Moving towards metadata driven production: the case of National Accounts input data

Anette Morgils Hertz[1], Søren Kristensen[2]

[1]*Statistics Denmark, National Accounts, Denmark*

[2]*Statistics Denmark, Methodology and analysis, Denmark*

Abstract

Statistics Denmark has as one of its goals to ensure that the production of official statistics is metadata driven. To support this we have developed a documentation portal (the Metadata bank), which aims to gather all relevant metadata in one application. The Metadata bank is based on GSIM and contains the information objects that we have found are necessary to give a full metadata account. This include concepts, variables, quality assessments, classifications and correspondence tables. The Metadata bank will make it much easier for users to find comprehensive and consistent metadata. Equally important it will significantly improve the use of metadata in the production of statistics, thereby enhancing the quality of official statistics.

In this paper, we will present and discuss our efforts to integrate the Metadata bank into the Danish National Accounts' new source data system. By integrating the two, we have managed to build a metadata-driven source data system, that can transform the heterogeneous data sources used to compile the Danish National Accounts into the standardized classifications used in the compilation of the National Accounts. Using specific cases, we will focus on the challenges we met when fitting GSIM to data and describe how the system uses metadata from the Metadata bank to transform the source data.

Keywords: GSIM, Metadata driven, correspondence tables, National Accounts

1. Introduction

The users of the national accounts are very focused on the revisions of the national accounts, from the first estimate of GDP is calculated in the quarterly system until the final estimate is calculated in the supply-and-use-table based annual accounts system. This has made us look for ways to reduce the revisions and generally increase the efficiency of our compilation processes. One initiative to increase efficiency and potentially reduce revisions was to build a metadata driven source data system for the national accounts.

At the same time, there has been a lot of focus in Statistics Denmark on consolidating metadata in a new system - the Metadata bank. This paper focus on how the national accounts source data system has made use of this new development to meet the goal of making the process metadata driven. Simultaneously, the work on the source data system has proven to be a very good test of the Metadata bank. It has both underlined some clear benefits of the system and highlighted some issues we should be aware of when using the Metadata bank as the core metadata storage in a metadata driven production system.

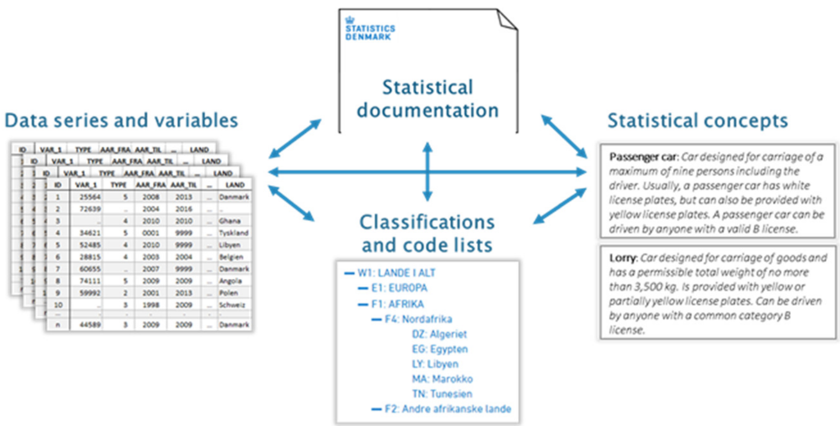
In this paper we will introduce the Metadata bank and the new Danish National Accounts source data system, describe how the systems have been used to create metadata driven production, and discuss some of the issues we have encountered in the process.

1.1 The Metadata bank

The Metadata bank is a result of an ambitious goal to integrate all relevant metadata objects in one application. It is based on GSIM in the sense that we have selected a number of key objects from the model for implementation in the Metadata bank. These range from the strategical level, e.g. the statistical program, to objects describing individual data points. The Metadata bank does not include all GSIM objects – far from it – but we are planning to gradually add more objects to the Metadata bank as we learn more from its use in practice. The aim of the Metadata bank is to ensure consistency and coherency across metadata objects, to improve their documentation, and to create a tool for metadata driven production of statistics.

Metadata for data series, variables, classifications, concepts, and statistical documentations, which all existed in their own separate systems, are now not only stored in the same system, they are also linked to each other (see figure 1 below). Prior to this, metadata existed in separate systems and too often only in individual statistical production systems. It was not uncommon, and still is not, that variable descriptions and classifications are stored and updated in individual excel sheets.

Figure 1: The vision: Coherent Metadata



Since its launch in January 2023, our work has focused on populating the system with all the relevant metadata. This is no small task and it is still in focus for the next years to come. However, as the Metadata bank has matured we have gradually started focussing more on metadata driven production.

1.2 The source data system

The source data system for the Danish national accounts is a database system that can receive data delivered for the regular production of the national accounts in a standardised format. Data for the national accounts primarily consists of different primary statistics such as the balance of payment, PRODCOM, the household budget survey etc. Historically these primary statistics have been delivered twice, once for the annual and once for the quarterly accounts. The transformation to national accounts concepts has been carried out in both of the two compilation systems, sometimes resulting in small differences that could end up causing revisions.

In the new source data system the transformation from source data to national accounts concepts is done once, using metadata from the Metadata bank, where after the two different compilation processes start.

The goal of the new source data system is to ensure that only data with the correct format and valid codes can be delivered into the national accounts source data system. The initial data validation is performed and data is transformed into national accounts concepts using statistical classifications and correspondence tables from the Metadata bank. Another goal for the source data system is that it should be self-documenting and enable reproducibility of the compiled national accounts by logging the exact version of metadata used for transforming the delivered data. To put it in GSBPM terms, the source data system will cover the Collect Phase, with all four sub-processes, but it will also cover sub-processes 5.1 Integrate data and 5.2 Classify and Code

Entering the statistical classifications and the correspondence tables into the Metadata bank is done in close cooperation with Statistics Denmark's quality unit. This unit is the owner of the Metadata bank and is in charge of operating and developing it. Updating and storing the classifications and correspondence tables in the Metadata bank is key to making the statistical production metadata driven.

Since the source data system is one of the first examples of metadata driven production using the Metadata bank as the central storage facility a number of issues has been identified. These issues will need our attention when moving towards even more metadata-driven production. Some of the more critical issues we identified when working on integrating the Metadata bank and the source data system are:

1. Downtime in the Metadata bank directly effects the source data system, as it needs access to statistical classifications and correspondence tables when validating data deliveries and transforming them to national accounts concepts
2. Statistical classifications need to be uploaded to the Metadata bank in a timely manner. This means before their "valid from" date, such that they can be accessed as soon as the primary statistics start delivering data where the statistical classification should be used for validation
3. Likewise correspondence tables should be uploaded in time to be used for transformation of the delivered data
4. The number of statistical classifications and correspondence tables stored in the Metadata bank will increase drastically. This may create a bottleneck if the quality unit alone is able to upload new classifications and correspondence tables to the Metadata bank
5. The "valid to" and "valid from" dates should follow the same date format for all statistical classifications and correspondence tables
6. The correspondence tables need to depict valid correspondences, hence there should be built in checks between the validity of the statistical classifications and the validity of each of the correspondences in the tables

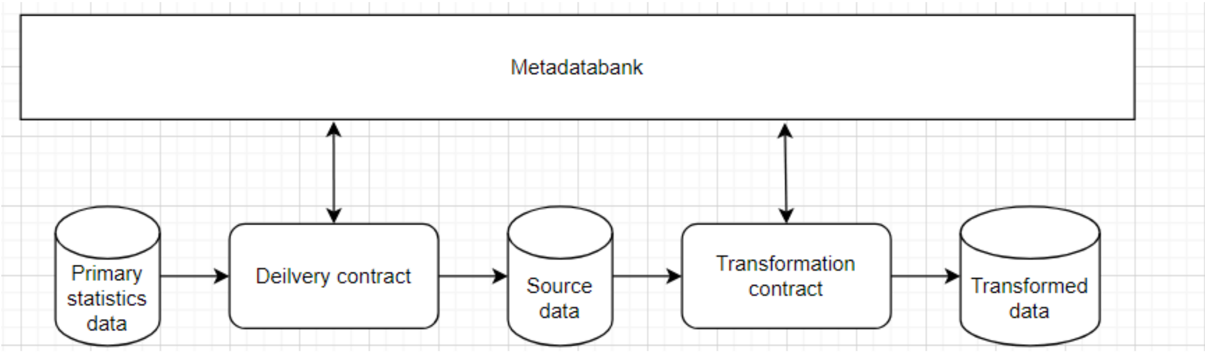
In this paper we will focus on the critical issues described in 2, 3 and 5 to ensure as smooth a metadata-driven production as possible.

2. How the metadata driven source data system works

It is important to stress that successful metadata driven production is not merely a question of solving technical issues and designing a system to carry metadata. It also involves close and continuous cooperation between owners and users of data. Prior to each primary statistics' integration into the national accounts source data system a "delivery contract" is agreed upon. Here the format of the delivery from the primary statistics domain to the source data system is designed in terms of the names of the variables, and their datatypes. If a variable is to be verified against a statistical classification from the Metadata bank, then the exact name and version of the classification is assigned to the variable. Based on this contract an empty "source data" table is created, ready to store any future deliveries.

Another “transformation contract” is created focusing on the variable names and datatypes of the transformed data. For each variable the contract contains how the variable should be assigned data. It could be a 1:1 from the source data table, or a new variable could be created based on a correspondence table from the Metadata bank. In this case the exact name of the correspondence table and its version is assigned in the transformation contract. The process is depicted in figure 2.

Figure 2: From primary statistics to transformed data fitting the national accounts concepts



When both the delivery and the transformation contract is in place the system is built such that when a delivery is initiated the process of validating whether the data matches the criteria for the source data table starts automatically. If the delivery fails the office of the primary statistics is notified that something is wrong and that they have to fix that before their next attempt of delivery. If the delivery of the source data is successful, on the other hand, then the transformation process starts automatically. If the transformation then fails, the source data team is contacted, and will have to figure out why the transformation failed. If the transformation is successful the national accountants is notified that their data is ready.

2.1 Metadata-driven validation and transformation

Let’s look at an example. The producer of primary statistics X delivers a simple dataset as shown in table 1.

Table 1: An example of a simple source dataset

Periode	SITC	Value
01-01-2022	00111	1000
01-01-2022	00150	2000
01-01-2023	11101	3000

The variable SITC should be validated against metadata from the Metadata bank figure 3. Hence it is critical that the statistical classification is available in the Metadata bank (critical issue 2). The validation consists of comparing the validity of the given SITC code and the reported period with the codes validity in the statistical classification. Since all three codes are valid, and the data fits the format specified in the delivery contract data ends up in the source data table.

Figure 3: A screenshot from the Metadata bank of the beginning of the statistical classification of SITC.

KLASSIFIKATION

VIS HELE KLASSIFIKATIONEN

Niveau	Kode	Tekst	Lang tekst	Beskrivelse	Inkulderer	Inkluderer også	Ekskluderer	Retsgrundlag	Måleenhed	Gyldig fra	Gyldig til
1	00111	Hornkvæg til avlsbrug								01-01-2000	
1	00119	Levende hornkvæg, undt. til avlsbrug								01-01-2000	
1	00121	Levende får								01-01-2000	
1	00122	Levende geder								01-01-2000	
1	00131	Svin til avlsbrug								01-01-2000	
1	00139	Levende svin, undt. til avlsbrug								01-01-2000	
1	00141	Levende fjerkræ af vægt max 185 gram								01-01-2000	

The transformation process is then initiated. In the transformation contract it is specified that a new variable, called "subcategory", be created using a correspondence table between SITC and a subcategory-code designed by the national accountants. Again it is critical that the correspondence table is available in the Metadata bank and that the "valid from" and "valid to" dates are correct for each correspondence (critical issue 3). Figure 4 shows the correspondence table.

Figure 4: A screenshot from the Metadata bank of the beginning of the correspondence table between SITC and a subcategory-code

Korrespondancetabel

EKSPORTER

Filtrer tabel:

Fra Kode	Fra Tekst	Gyldig fra	Til kode	Til Tekst	Gyldig til
00111	Hornkvæg til avlsbrug	01-01-2000	00	nan	31-12-9999
00119	Levende hornkvæg, undt. til avlsbrug	01-01-2000	00	nan	31-12-9999
00121	Levende får	01-01-2000	00	nan	31-12-9999
00122	Levende geder	01-01-2000	00	nan	31-12-9999
00131	Svin til avlsbrug	01-01-2000	00	nan	31-12-9999

In Table 2 the resulting transformed data table is shown. It contains the newly developed variable "Subcategory", which gets its input from the correspondence table in figure 4. When the transformation is successful, the national accountants are notified that data is available for the compilation process to start. And the metadata driven source data system has done its job.

Table 2: An example of a simple transformation dataset

Periode	SITC	Value	Subcategory
01-01-2022	00111	1000	00
01-01-2022	00150	2000	00
01-01-2023	11101	3000	10

2.2 Date formats: an example of an issue encountered in using the system

Setting up a new production (the source data system) and linking it to another new system (the Metadata bank) in order to support metadata driven production, we have encountered a number of seemingly minor issues, which needed to be solved for the production process to work. An example of this is the date formats used in the classification tables (critical issue 5 described above). The source data system communicates with the Metadata bank via an API. There is a Python package that given certain input calls the Metadata bank and extract what is requested. This Python package was developed during the development of the source data system.

During the development of the Python package and the subsequent testing, it was discovered that there were examples of different date formats for the "valid to" and "valid from" fields for the statistical classifications in the Metadata bank. It had been assumed that the “valid to” and “valid from” fields would always return a date of the format dd-mm-yyyy, however some statistical classifications only use yyyy in the “valid to” and “valid from” in the Metadata bank. Let us consider an example: The statistical classification for Country codes following the ISO 3166 standard. Here we had that “Western Germany” had the country code DE from 1970 to 1989, and that “Germany” had the country code DE from 1990 and counting. Since a metadata-driven system needs a specific date, it was decided that it is the administrators of the Metadata bank that should translate a “valid from” date of 1970 to 01-01-1970 and a “valid to” date from 1989 to 31-12-1989, such that it is done similar for all users of the API.

3. Conclusions

There is little doubt that using the Metadata bank and creating a metadata driven production system has many benefits. When integrated into a production system the Metadata bank will ensure a transparent production process as well as secure that the production process is reproducible.

The Metadata bank is the central place for storing and updating metadata. The Metadata bank is available to all internal users, which helps to ensure that the correct metadata is used in the statistical production process. The Metadata bank has the ability to keep track of vintage versions of statistical classifications and correspondence tables, which helps ensuring that metadata driven systems can reproduce historic results, by accessing prior versions of metadata. Before the launch of the Metadata bank, it was not uncommon to update metadata in local production systems, which was a great source of errors in the statistical production, as there were no system to keep track of vintage metadata versions, and the “same” metadata therefore would be in several production systems.

As the example of integrating the Metadata bank and the new Danish national accounts source data system shows, the storing of metadata in a central Metadata bank does not come without challenges. In the metadata driven statistical production system, the Metadata bank becomes critical production infrastructure. There is no doubt that it is the way forward, but there are a number of issues which have to be acknowledged and addressed in order to successfully scale and further develop metadata driven systems.

Close attention has to be given to the timeliness of the metadata to be used in the production cycle. It has to be ready and available at the time the production cycle starts.

Internal consistency is important as illustrated by the date format issue. Inconsistencies may not be identified until metadata from several sources are integrated, so laying down some general rules, for instance regarding date formats, could minimize the problem. Adequate IT and man power resources has to be allocated to the Metadata bank so as to ensure its reliability, as it becomes a key part in the production systems. Finally, introducing metadata driven production also raises a series of questions regarding roles and responsibilities between the quality section and the producers of statistics.