NSM 2022
NORDIC STATISTICAL MEETING

Implementing coherent metadata for production and dissemination

Charlotte Nielsen, Senior Adviser, Methodology and Analysis, Statistics Denmark, cln@dst.dk

Rasmus Anker Stavad, Senior Adviser and Metadata Coordinator, Methodology and Analysis, Statistics Denmark, rkr@dst.dk

Abstract

In recent years, Statistics Denmark has worked on implementing consistent and coherent metadata to support metadata driven production. The implementation follows international standards and focuses on increasing efficiency, by using standardized metadata actively and systematically in the production of official statistics and for dissemination to users. The content and organization of structural metadata follow the recommendations and standard outlined in UNECE's Generic Statistical Information Model (GSIM). The GSIM model excels by supporting the information objects that go into and out of the individual process steps of the Generic Statistical Business Process Model (GSBPM). This paper will present Statistics Denmark's new and improved documentation portal, which showcase relationships between Quality reports (SIMS), classifications and code lists, as well as the documentation of registers and variables. Examples of coherent metadata and how they are presented for the users will be given.

Keywords: Coherent metadata, documentation portal, GSIM, GSBPM, standardisation, metadata driven production

1. Introduction

For the better part of the last decade, Statistics Denmark has been on a mission to implement coherent, quality assured statistical metadata for production and dissemination of official statistics. With that said, statistical metadata has always been an integral part of official statistics, think classifications for compartmentalising figures in smaller groups or text in footnotes to tables and graphs, explaining the finer details.

Nonetheless, this metadata mission is historically unparalleled, in terms of consolidating statistical metadata into one common centralised repository for shared use in production and dissemination of official statistics. Additionally, there has been a significant maturation and recognition within Statistics Denmark of the value and efficiency gains of coherent, quality assured statistical metadata. Historically, Statistics Denmark's statistical metadata were produced and stored in different silo systems, with no explicit alignment to international standards. This has now changed. Statistical documentation, classification and code lists, statistical concepts and data series and variables are now consolidated and stored in one system, in accordance with international standards.

This paper sums up the process and achievements of the last decades' work with consolidating Statistics Denmark's statistical metadata into one system. The system of choice being *Colectica*. In simplified terms, Colectica is a user interface/editor for developers and metadata producers of standards-based metadata documentation.[1] Colectica uses the *Data Documentation Initiative* (DDI)[2] as its native storage format and makes use of a relational metadata repository stored in *PostgreSQL*. With statistical metadata, being as interrelated as it is, the architecture of Colectica, DDI and PostgreSQL aligns well with each other for documenting official statistics. This is where the fun begins.

DDI is more or less congruent with the Generic Statistical Information Model (GSIM) and speaks the same language as other metadata standards such as the Single Integrated Metadata Standard (SIMS) for quality reports and the Neuchâtel terminology model for classification object types and their attributes.

[1] Colectica

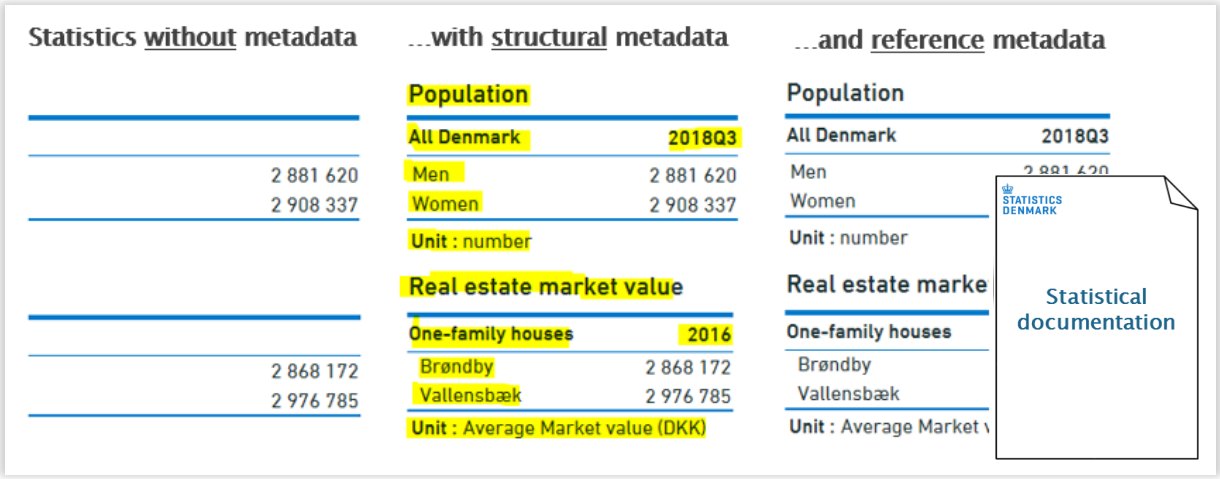
[2] DDI Alliance

2. About the study

2.1 Structural versus reference metadata

Statistics Denmark operate with two types of statistical metadata: *structural metadata* and *reference metadata*. **Structural metadata** identify statistical data, e.g. titles, code lists, time dimensions, unit of measure and variable names etc. and must go together with statistical data. **Reference metadata** describe statistical concepts and methodologies used for the production of statistics and provide information on quality to help users with the interpretation of the data. Contrary to structural metadata, reference metadata can be detached from the figures, meaning that they *can* be produced separately from the statistics to which they refer.[3]

Figure 1. Statistics with and without structural and reference metadata



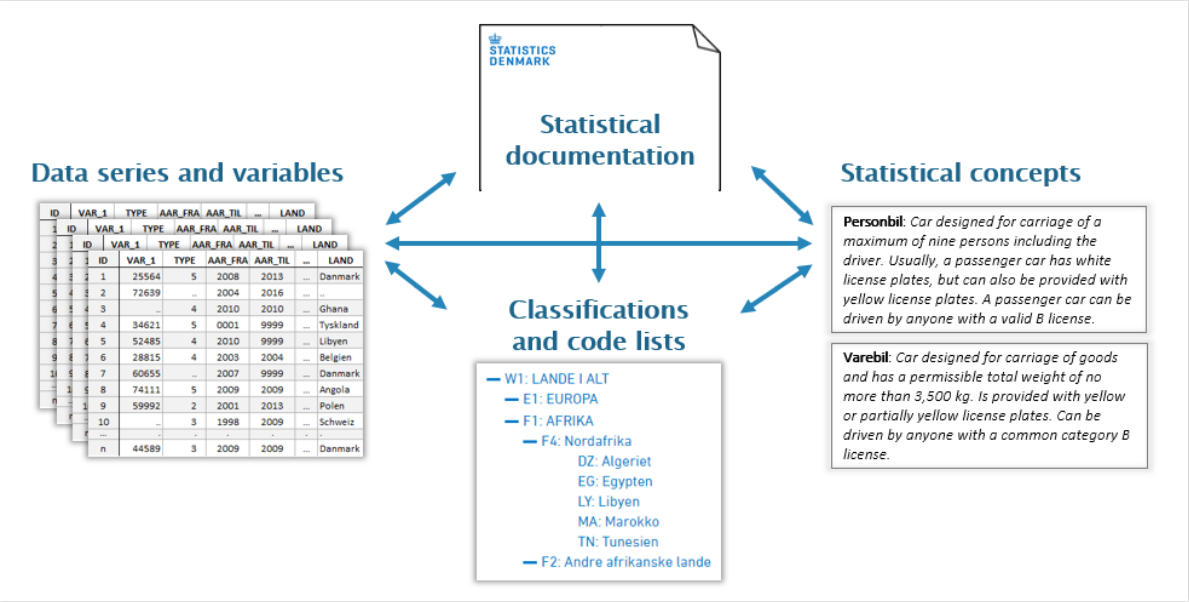
2.2 Four pools of statistical metadata

For documenting official statistics, Statistics Denmark has pooled statistical metadata into four groups: 1) statistical documentation, 2) classification and code lists, 3) statistical concepts and 4) data series and variables.

Historically, the four pools of statistical metadata were more or less a given and the logical outcome of underlying standards. Prior to the acquisition of Colectica and the adaptation of DDI, statistical documentation and data series and variables were stored in one system, classifications and code lists in another and statistical concepts in a third, all of them with close to no overall alignment with international standards.

[3] Eurostat - ESS Reference Metadata Reporting Standards

Figure 2. Interrelated statistical metadata



The vision and purpose at mission start, was to implement coherent metadata to support metadata-driven production and to:

- Meet the needs of internal and external users of metadata on statistical outputs in relation to their desired use of statistics (fit-for-purpose).
- Ensure that the quality of statistical products and processes are described and comply with quality requirements.
- Conduct systematic quality checks of statistical products and processes, and implement quality improvements.
- Increase efficiency so that also employees of Statistics Denmark can use shared metadata to complete their tasks, by maintaining metadata in one metadata system and making it easily accessible to all.

In the most simplified of simplified terms, firstly, the task was to create placeholder repositories for statistical metadata and align them with international standards. This meant mapping out the different standards within the four corners of metadata. Secondly, to transfer existing statistical metadata objects from old systems to the new repository and adapt to these international standards. Thirdly, quality assure statistical metadata while expanding the quantity of metadata and finally, provide users and producers with coherent, quality assured statistical metadata.

2.2.1 Statistical documentation

For statistical documentation, the Single Integrated Metadata Structure (SIMS)[4] was applied as the standard and the new statistical documentations were deployed in 2014. We went from no metadata standard for reference metadata to being fully compliant with the SIMS. The primary efficiency gain, being able to upload local documentation directly to Eurostat in the ESS Metadata Handler. Documentation of statistics for all statistical products are available on our webpage in both Danish and English.

2.2.2 Classifications and code lists

For classifications and code lists, the Neuchâtel terminology model for classification object types and their attributes[5] was adopted. Documentation of classifications and code lists were deployed in 2018. Classifications are available on the website of Statistics Denmark, e.g. Regions, provinces and municipalities (Danish adaptation of NUTS). We went from a low to a high degree of standardisation and centralisation of statistical classifications.

2.2.3 Statistical concepts

For statistical concepts the ISO 704 *Terminology work – principles and methods*[6] was adopted. Shared quality assured one-place-only and once-only descriptions of statistical concepts, imbedded in statistical documentations, was deployed by 2020. We have yet to roll out statistical concepts on a larger scale and completely harmonize across statistical documentation, the website and the StatBank.

2.2.4 Data series and variables

Documentation on data series and variables are being deployed at present. We are in the process of moving existing documentation from our old system to Colectica. The tricky part has been aligning with GSIM[7]. The key modernisation feature going forward is the introduction of the *variable cascade*, which includes splitting documentation of variables into conceptual variables, represented variables and instance variables.

---

[4] European Statistical System (ESS) handbook for quality and metadata reports

[5] Neuchâtel terminology model for classification object types and their attributes

[6] ISO 704 Terminology work – principles and methods

[7] Generic Statistical Information Model (GSIM)

2.3 Quality assurance

As stated in the mission, not only should metadata be produced and made readily available to users and producers, it shall also be quality assured.

Processes have been set up to systematically quality assure **statistical documentation** connected to every single publication of statistics. Every publication of a statistical product must be covered by an updated statistical documentation (reference metadata) wherein the content, processing and five quality dimensions (relevance, accuracy and reliability, timeliness and punctuality, comparability and accessibility and reliability) is elaborated in detail. Changes to the statistical documentation is read by a peer who specialise in compliance with quality requirements in SIMS prior to publication.

Within **classifications and code lists**, quality assurance is far more ad hoc, but equally systematic in its practice. Whenever a classification or code list needs to be created or updated, a peer who specialises in compliance with Neuchâtel, assists the responsible statistician with understanding, which classification object types are necessary, to load a classification into Colectica. Overall, classifications are compliant only if categories are mutually exclusive, exhaustive, normative, stabile and valid for a given time period. In practice however, things are not black and white. Furthermore, there are requirements for logic and discipline in the codes used, definition of levels, uniformity in hierarchy depth etc. On top of that, there are variants, correspondence tables and often issues with the national implementation of international nomenclature.

Quality assurance of **statistical concepts** sits somewhere in between the recurring nature of statistical documentation and the ad hoc nature of classifications and code lists. Statistical concepts are meant to be shared between statistical products. Therefore, they must rely on agreed upon definitions and be described in an appropriate manner. For this, we rely on eight “rules” for writing good statistical concept descriptions. 1) Write short and simple, one sentence if possible; 2) consider the target audience; 3) clarity, do not use specialist language, 4) coordinate mutually with other concept descriptions, 5) write adequate, not to narrow or broad, 6) no circularity, 7) no negative definitions and 8) write in singular terms.

Within documentation of **data series and variables**, the establishment of quality assurance procedures is a work-in-progress, to be decided later this year.

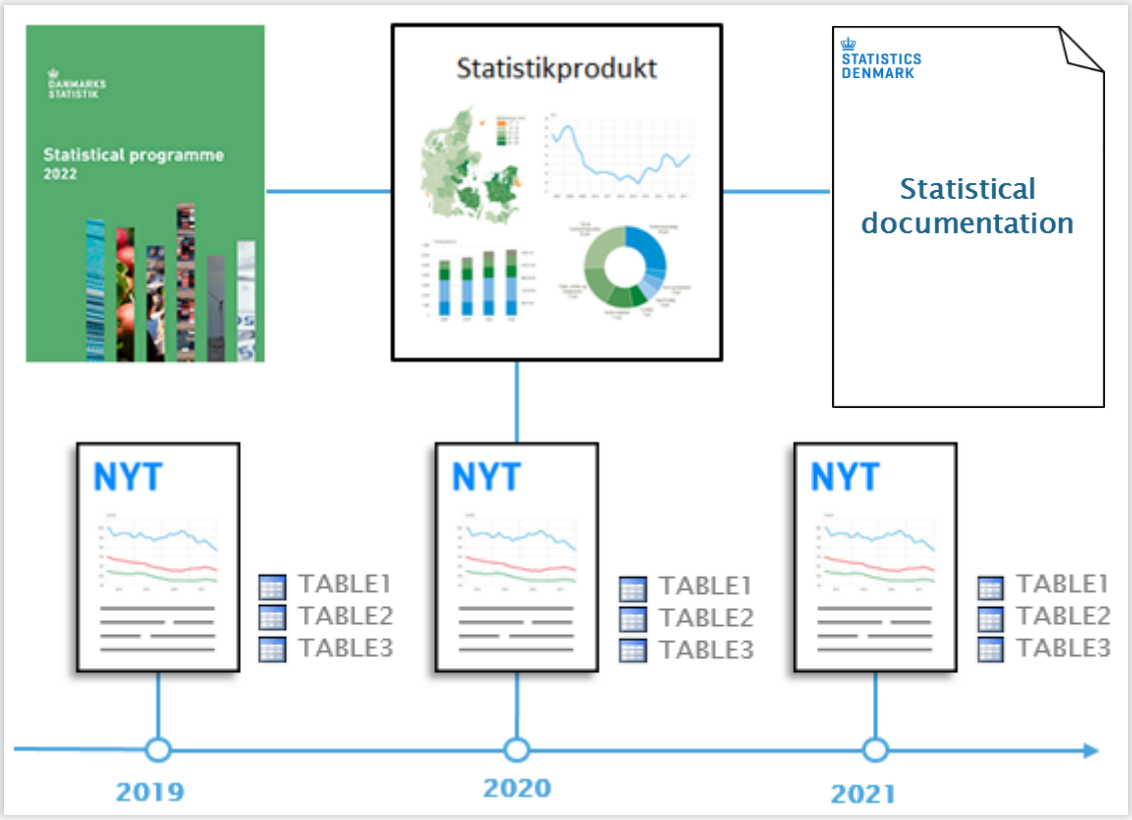
2.4 Statistical products

So far, we have mentioned in passing *statistical products* without really touching upon what that means in relation to statistical metadata. According to an OECD definition,

"**Statistical Products** are, generally, information dissemination products that are published or otherwise made available for public use that describe, estimate, forecast, or analyse the characteristics of groups, customarily without identifying the persons, organisations, or individual data observations that comprise such groups".[8]

In a sort of side mission to implementing coherent metadata, we have taken the above definition, thought about what that means to us and institutionalised it accordingly.

Figure 3. Statistical products in Statistics Denmark



At the heart of our statistical products are the published figures. These are made available through scheduled releases in tables in the StatBank and in a press release (Danish only) on the same day. From the tables and press release there is a link to reference metadata in a statistical documentation.

[8] OECD Glossary of Statistical Terms - Statistical Products

Each year, Statistics Denmark releases a Statistical programme, containing an overview of all official statistical products produced by Statistics Denmark. Each of the statistical products are described briefly with information about the purpose and content of the statistics.

Statistical products are *branded* so that the title of the product, statistical documentation, press release and subject headline match across dissemination and communication platforms. Well known international examples of named statistical products are the *Consumer Price Index (CPI)* and *Labor Force Survey (LFS).*

A side benefit from this side mission has been a revision of our complete inventory of official statistical products. In terms of implementing coherent metadata aka the main mission, this work has been an important step when linking statistical metadata objects to specific statistical products.

2.5 Aligning everything with GSBPM

Whatever we do, we have the Generic Statistical Business Process Model (GSBPM)[9] in the back of our minds. Statistical metadata is an abundance of GSIM objects and as we know from the GSBPM, GSIM provides a set of standardised, consistently described information objects, which are the inputs and outputs for GSBPM sub-processes. Therefore, there really is no way around GSBPM when working with metadata.

On that note, we will be reading the recent UNECE report on linking GSBPM and GSIM[10] published in 2022 going forward as a frame for next steps. However, for now, we acknowledge that when we formulate statistical concepts it's GSBPM sub-process *1.4 Identify concepts*. When documenting variables it's GSBPM sub-process *2.2 Design variable descriptions*. When applying statistical classifications to data it's GSBPM sub-process *5.2 Classify and code*. The list goes on, but the point is that all statistical metadata objects has a place in the GSBPM.

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

[10] UNECE report on linking GSBPM and GSIM

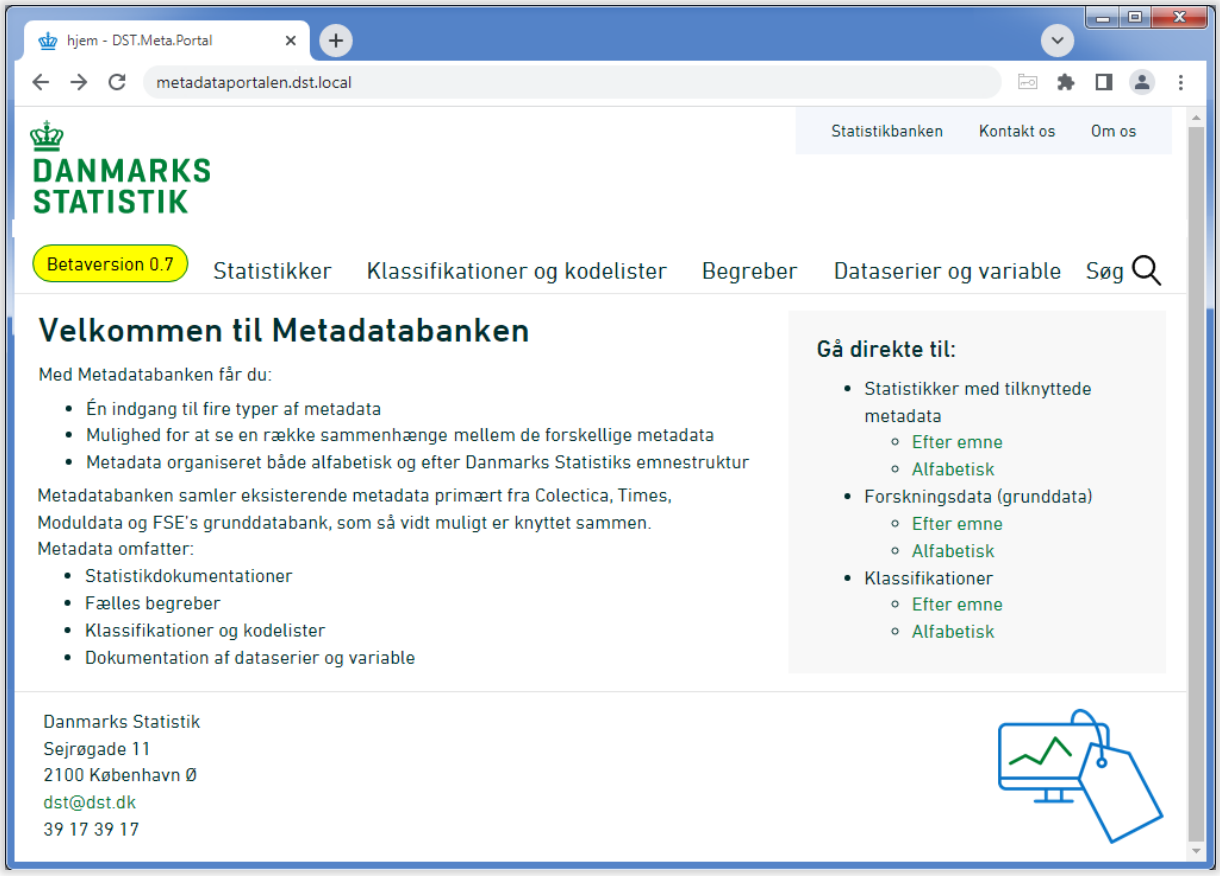
3. Results

As of 2022, Statistics Denmark is finalising the mission of implementing coherent metadata for production and dissemination of official statistics. The further the mission goes on, the more confident we become in our belief that the idea of coherent metadata is buoyant. During the better part of the past decade, the work done in this field has delivered results we can be proud of. It is, however, tricky to sum up the results from a decades work, so three achievements will be highlighted; the launch of the MetadataBank, Metadata linkage from input to output, and the use of centralised quality assured metadata in practice, in the production of official statistics.

3.1 Launch of the internal MetadataBank

Launching the internal MetadataBank in June of 2022 was a huge leap in terms of visualising all of the statistical metadata content stored in the centralised system.

Figure 4. Front page of Statistics Denmark's internal MetadataBank

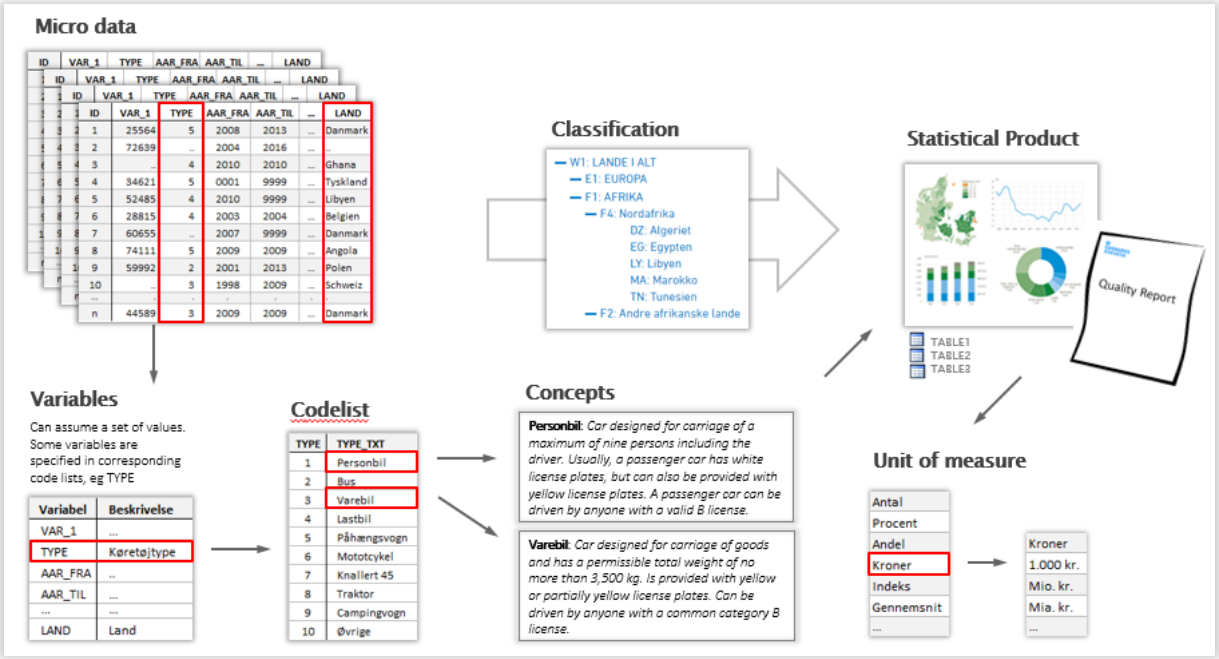


Internal metadata users and producers now have a tool for browsing statistical metadata, including relations between metadata objects.

3.2 Metadata linkage from data to output

During our mission, we had the sketch below in figure 5 as our reference point. Now we have most of the arrows covered. Thanks to GSIM and DDI, there are now links functioning links from data to variables, to classifications and code lists, to statistical products and their statistical documentation, including concepts. The linkage is actually, what makes it possible to navigate around in the MetadataBank.

Figure 5. Coherent metadata linked from data to output



3.3 Using coherent metadata in practice

Final highlight is that we have actually succeeded in putting the centralised system for statistical metadata in operation in the statistical production. Production systems are actively using Colectica's API to extract quality assured statistical metadata from the shared repository, i.e. using classifications for data editing and for data visualisation.

4. Looking ahead

We have come a long way, but there is still a bit of way to go, before we can confidently say we have completed our metadata mission on implementing coherent statistical metadata for production and dissemination. The non-exhaustive headlines of work for the next couple of years will be to:

- Introduce unit types into our metadata model and link to relevant objects
- Add a user interface layer to the MetadataBank in order to edit metadata
- Introduce quality assurance procedures for data series and variables
- Expand the pool of quality assured metadata in Colectica
- Use metadata from Colectica directly in questionnaires
- Further integrate metadata from Colectica and the StatBank
- Launch the MetadataBank for external users of metadata
- Enhance the use of Colectica API for external system users
- Link GSIM based metadata to GSBPM
- Fully adopt the FAIR principles for metadata

5. References

Colectica

DDI Alliance

Eurostat - ESS Reference Metadata Reporting Standards

European Statistical System (ESS) handbook for quality and metadata reports

Neuchâtel terminology model for classification object types and their attributes

ISO 704 Terminology work – principles and methods

Generic Statistical Information Model (GSIM)

OECD Glossary of Statistical Terms - Statistical Products

https://statswiki.unece.org/display/GSBPM

UNECE report on linking GSBPM and GSIM