Towards Common Metadata Using GSIM and DDI 3.2

by Mogens Grosen Nielsen and Flemming Dannevang[1]

Abstract

Since the 1990's, Statistics Denmark has provided users with rich metadata, including classifications, quality declarations, and variables and concept systems. In 2011, we initiated projects aimed at a common integrated metadata system in order to improve quality and facilitate dissemination of statistics. At the beginning of 2015, Statistics Denmark launched a DDI-based system handling concepts and quality information for 237 statistics. At the ongoing project Statistics Denmark are including variables, concepts and classifications in one common metadata system. GSIM and DDI 3.2 are used as standards for building models at the conceptual, logical and implementation levels; Colectica is the software of choice. We demonstrate in this paper that the road towards common metadata in a statistical context requires a) improvement and precision in the terminology we use when we talk about metadata, b) better understanding of the role of metadata in relation to our users, and c) greater recognition of the role of metadata in production of statistics. The paper presents various perspectives on metadata in our context, including metadata in a helicopter perspective, metadata in business processes, and metadata implemented in Colectica using GSIM and DDI 3.2.

Keywords

Frame of reference, metadata, GSIM, DDI, Colectica

Introduction

Since 2011 Statistics Denmark has been working on building a common metadata system based on DDI and other standards among which quality standards from Eurostat play a key role. In January 2015, the standardised description of quality was in place for 237 statistics. The ongoing work focuses on variables, concepts and classifications. As in all other metadata systems the important part is to work towards reuse of common metadata. By common metadata we mean metadata about concepts, classification, and variables that are stored only in one place and reused across various domains and in all relevant business processes, including dissemination processes. Civil status, firm, and income concepts are examples of common metadata in our system. Common metadata are used together with local metadata that are specific to a statistical domain. Metadata introduces a high degree of complexity and requires a great deal of effort from various disciplines and from the whole organisation. The claim in this paper is that the

Common metadata are used together with local metadata that are specific to a statistical domain.

path towards common metadata in a statistical context requires a) improvement and precision in the terminology that is used when we talk about metadata b) better understanding of the role of metadata in relation to our users, and c) greater recognition of the role of metadata in the production of statistics.

In section 2, we first give a short history of the development of metadata at Statistics Denmark and then provide a description of the challenges with regard to opinions on metadata. We also touch upon the difficulties related to the introduction of the Generic Statistical Information Model (GSIM) and the Generic Statistical Business Process Model (GSBPM). In section 3, we offer a helicopter perspective on metadata in order to give an idea of how to approach metadata, focusing on metadata as frames of reference. Section 4 introduces metadata and business processes. We first introduce a general model. Hereafter we present a model on how to integrate metadata in business processes. The last part of the section focuses on processes related to user needs. Section 5 gives a longer introduction to models and terminology used in GSIM. It can be seen as our approach to integrate not only GSIM, but also DDI 3.2 in a Colectica [4] implementation.

Metadata at Statistics Denmark: Present situation and challenges

For several years, Statistics Denmark operated with separate classifications, quality declarations, and variables and concept systems. Some of these were founded on international standards, but they lived their own lives without integration. In 2010, we decided to improve the integration of these systems, including going from separate metadata to creation and definition of common, reusable metadata. For example, concepts are partly defined in both variable and concept sub-systems. They must be common in order to get consistent definitions to be reused in various areas. On the dissemination side, we wanted to use common metadata together with other metadata to give users easier access to our products (metadata used to direct users to the right products) as well as better information about what we offer.

In 2010, we discussed whether to keep the existing systems and build an integration component on top of those or build a new integrated metadata system. At a METIS meeting in Geneva in 2011, it became clear to us that DDI could help us with the integration, reuse and dissemination of metadata. Note that the integration was the main driver in relation to dissemination. We wanted to give the users easy ways to navigate from, for example, a variable in a dataset to common metadata on classifications and code lists relevant for this variable. We launched a pilot study using Colectica, which is based on DDI. The results from the study were promising. In 2012, Statistics Denmark got an EU grant focusing on improved horizontal and vertical integration of metadata. We used this opportunity to introduce both DDI as a metadata standard and Colectica, which is a DDI-based metadata tool. In addition to DDI, we introduced the SDMX reference metadata on quality. This was basically a list of quality concepts and corresponding values. This implementation was and is compliant with the Single Integrated Metadata Structure (SIMS). The SIMS standard is developed and used by Eurostat and member states. It covers metadata for quality reporting and as such it is only a small brick in the bigger SDMX framework used by Eurostat. The DDI 3.1 standard had been improved and it was possible using Colectica to "plug in" SIMS into our metadata system. In January 2015, the standardised description of quality was in place for 237 statistics at Statistics Denmark. They are now an integrated part of our dissemination system.

In spring 2015, a new strategy on quality and metadata was approved by the top management at Statistics Denmark. The vision focuses on fulfilment of user needs, implementation of quality, and efficiency. Regarding metadata, the strategy stresses the use of standards, end-to-end production, reuse and active metadata.

We have since introduced new projects as part of the implementation. In the project on quality mentioned above and the present ongoing projects on concepts, variables and classifications, we are experiencing difficulties in people's perception of metadata and especially the role of metadata in the production processes. It has been difficult to allocate resources and to change processes (including some stovepipe production) towards using common standards.

In addition, there is a widespread opinion that documentation is something you attach to your statistical products after they have been published. There is little attention paid to how the statistics and documentation about the statistics are produced and should be used. User needs are often introduced too late in the production process or without the efforts necessary [1], [2]. Our observations also show that there is little awareness among our subject matter experts of the benefits of using common models, both in terms of precise definitions of information objects like population, statistical unit, etc., but also in how to apply common work procedures introduced in GSBPM. We have tried to introduce GSIM, although the general opinion is that GSIM seems to be too complicated. In addition, there is a widespread opinion that GSIM seems mainly to fit into organisations having a lot of resources for automation.

In order to create common metadata we must handle these challenges. In section 4, we focus on how to integrate use and production of metadata into production processes (4.1) and how to improve the dissemination of metadata (4.2). Section 5 focuses on the related challenges achieving common metadata terminology. All three aspects of metadata need to be addressed in order to get metadata to support not only automation but also improvements in the information about our statistics that we disseminate to our users. .

The helicopter perspective: Metadata as compatible frames of reference

The term "statistical system" is being used in many different ways. From a helicopter perspective, it is helpful to use the following distinction: a statistical system as a system of statistics or a statistical system as a system for producing statistics [5]. In short, a system of statistics is about the statistics we produce and a system for producing statistics is how we produce the statistics.

Working towards a situation with common metadata, we need some basics on reality, data, and information. The diagram below shows the interplay [5]::

In short, the mental process can be expressed by the following equation introduced by Børje Langefors: *"I = i(D, S, t)")" where I is the information (or knowledge) produced from the data D and the pre-knowledge S, by the interpretation process i, during the time t.[...] In the general case, S in the equation is the result of the total life experience of the individual. It is obvious from this that not every individual will receive the intended information from even simple data." [6]*

But what about common and reusable metadata which we are aiming at?

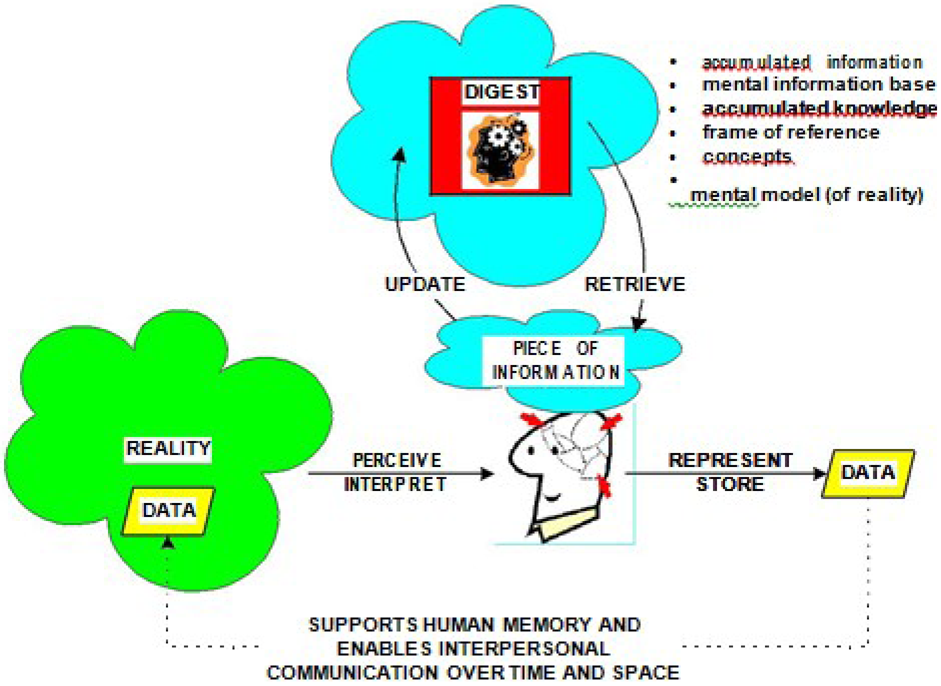


Figure 3.1 Reality, information and data.

*"Sharing of data (over time and space) is a proxy process for sharing of information. Sharing of information is fundamentally impossible. We can only do our best to improve the chances that different persons sharing the same data will interpret them in the same or at least similar ways. How can we do that?" We use "*compatible frames of reference: *A person's interpretation of data depends on the person's frame of reference, which consists of concepts and information in the person's mind. If two persons have the same or at least compatible, frames of reference, it seems likely that they will interpret the same data in similar ways"; [5]*

The challenge is to ensure compatible frames of reference. This is done by creating metadata: data about data. Metadata can be shared and communicated between users. *"Communication of metadata is subject to the same fundamental difficulties as communication of the basic data that they describe, but even so, adequate metadata will reduce the range of possible interpretations of the data that they describe, and thus improve the chances of different persons making similar interpretations of the same data." [5]*

Metadata has an essential role in facilitating compatible frames of reference. GSIM and other models can help in creating these frames of reference by introducing common terminology. Sundgren notes that communication about metadata shares the same difficulties as communication in general. These difficulties can be viewed at all levels: from global to simple communication between two people. Much research in this field introduces terminology on social systems and organisational learning [7], [8]. It is beyond the scope of this paper, but may be the most important aspect on the road towards a successful understanding and implementation of metadata.

Based on the considerations above, we can distinguish between two levels of terminology: general terminology on metadata, and domain specific terminology on metadata. The first level is terminology with regard to metadata constructs themselves (e.g., what is a "Variable" compared with a "Represented Variable" or "Concept"). The second level (closer to most end users) is instances of metadata constructs (e.g., the specific variable "Income of Business" and its definition, etc.). Both are associated with considerations of clarity and consistency. For example, "Income of Business" may vary as depending on what is included and excluded in the calculation of "Income". This is a key consideration. Often there exist multiple "standard" definitions e.g., definitions associated with accounting standards vs. definitions associated with the international System of National Accounts (SNA). Sometimes the challenge may be to explain how one definition relates to another, and/or to explain data quality considerations when data collected using one frame of reference (e.g. accounting / business reporting) are used to produce estimates for data using a different frame of reference (SNA).

The table below shows relations between general metadata terminology and domain specific metadata with respect to frames of reference of producers and users of statistics.

In section 5, we will focus on terminology related to metadata. We remedy this by introducing GSIM at several levels in order to have metadata terminology as a frame of reference. Note that the introduction in section 5 is directed to metadata experts inside NSIs. Whether we succeed depends on the ways in which we communicate. By going from a very abstract level to the logical and physical levels we believe that it will be easier to establish the needed terminology.

Metadata and business processes

How should we handle the role of metadata in the production of statistics and in relation to end users? We must establish processes that provide the right knowledge for the user. But more precisely, how should we build processes and interact with users? In order to move

	Frames of reference of producers	Frames of reference of various users
General terminology for statistical metadata	1. Complex metadata terminology for producers inside NSI's (examples: instance variable vs represented variable; classifications vs code-lists)	2. Simplified metadata terminology used both by internal and external users. (examples: classification, variable, concept, population used for dissemination e.g. search-tools)
Domain specific statistical metadata	3. Domain specific metadata (examples: detailed description of the definition of income of person)	4. Domain specific metadata differentiated and communicated with respect to frames of reference of various users (examples: short and detailed description of income directed to various user segments)

Table 3.1 Different kinds of frames of reference related to general metadata terminology and domain specific statistical metadata.

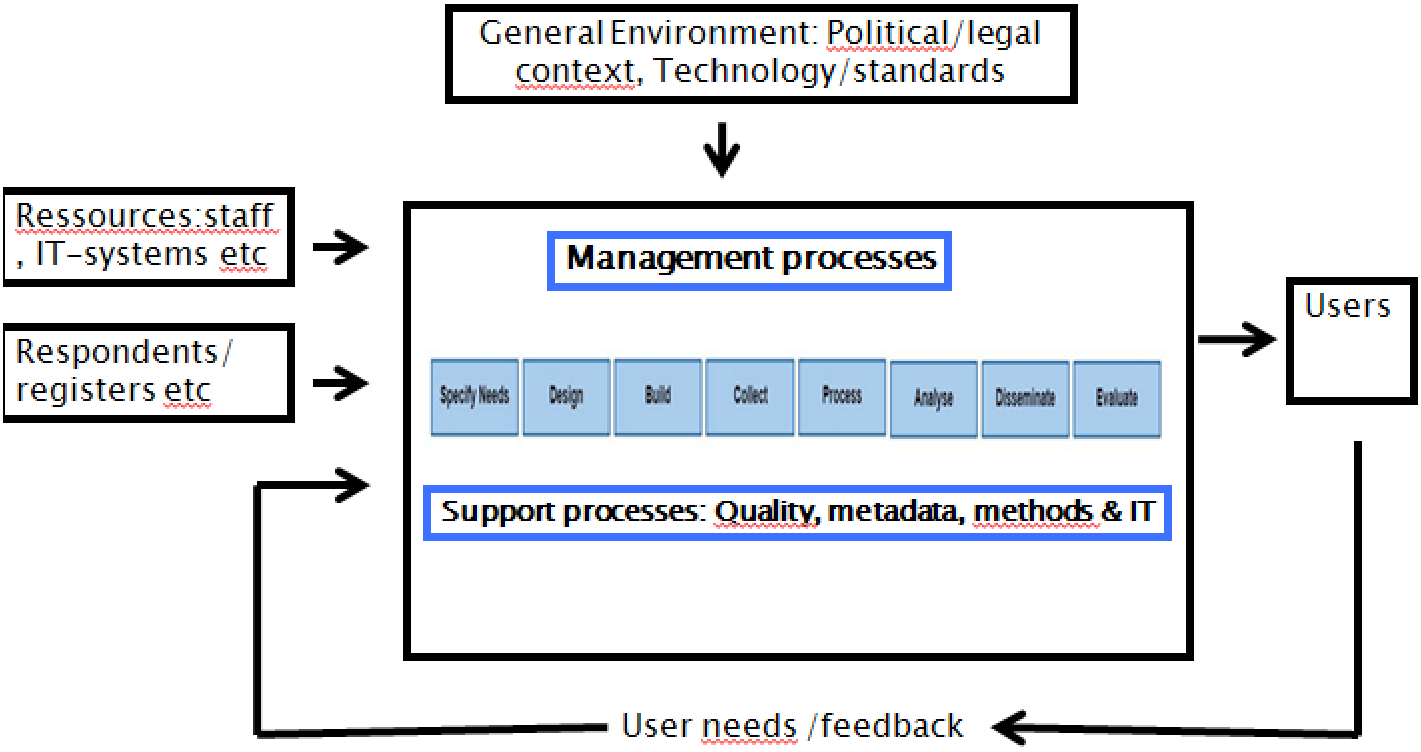

General Environment: Political/legal context, Technology/standards
Ressources:staff , IT-systems etc
Respondents/ registers etc
Management processes
Specify Needs
Design
Build
Collect
Process
Analyse
Disseminate
Evaluate
Support processes: Quality, metadata, methods & IT
Users
User needs /feedback

Figure 4.1 Business process perspective with environment elements.

in this direction we must have a general picture of statistical organisations. If we do not want silos depicted in a traditional organisational diagram, what should we have instead? The high level model of the process-centric organisation described below is a starting point.

TThe construction of the model is inspired by the idea of value chains introduced by Porter in the 1980's [9] and ideas on business process management [10] but also more recent ideas and practices related to GSBPM and GSIM. This way of thinking implies that you organise all processes from start to finish in such a way that each process adds value to users. This is reflected in distinctions between core processes, high level management processes and support processes. Quality, metadata, IT, etc., are seen as supporting processes. The focus must be on core processes delivering value to users. Management and support processes must be designed to assist the core processes. The production processes, including processes on metadata, must be designed to fulfil goals for the organisation – including goals on cost effectiveness.

The model gives an overall framework for the production of statistics. When combining the model with sources of inspiration above, it will be possible to build a complete framework for how to manage metadata, how to work with surveys, how to use standards on processes, metadata, etc., and how to build IT solutions.

The ideas on end-to-end production have existed for many years in the international statistical community, but very few NSIs have managed to walk successfully down that road. Examples of more large-scale implementations can be found in Canada, Australia and Sweden. They have all put a great deal of resources into building advanced cross-department solutions. The challenge for smaller NSIs with fewer resources is how to benefit from these ideas using standards and standard solutions taking small manageable steps in the right direction.

Integration of metadata in business processes

We must make sure that metadata is integrated in our processes in order to fulfil user needs, vision and goals outlined for Statistics Denmark. The diagram below shows an overall diagram of the business process architecture including how we expect to integrate metadata into GSBPM. The GSBPM phases are marked with a red rectangle. The boxes in the bottom show the production and use of metadata in the GSBPM phases.

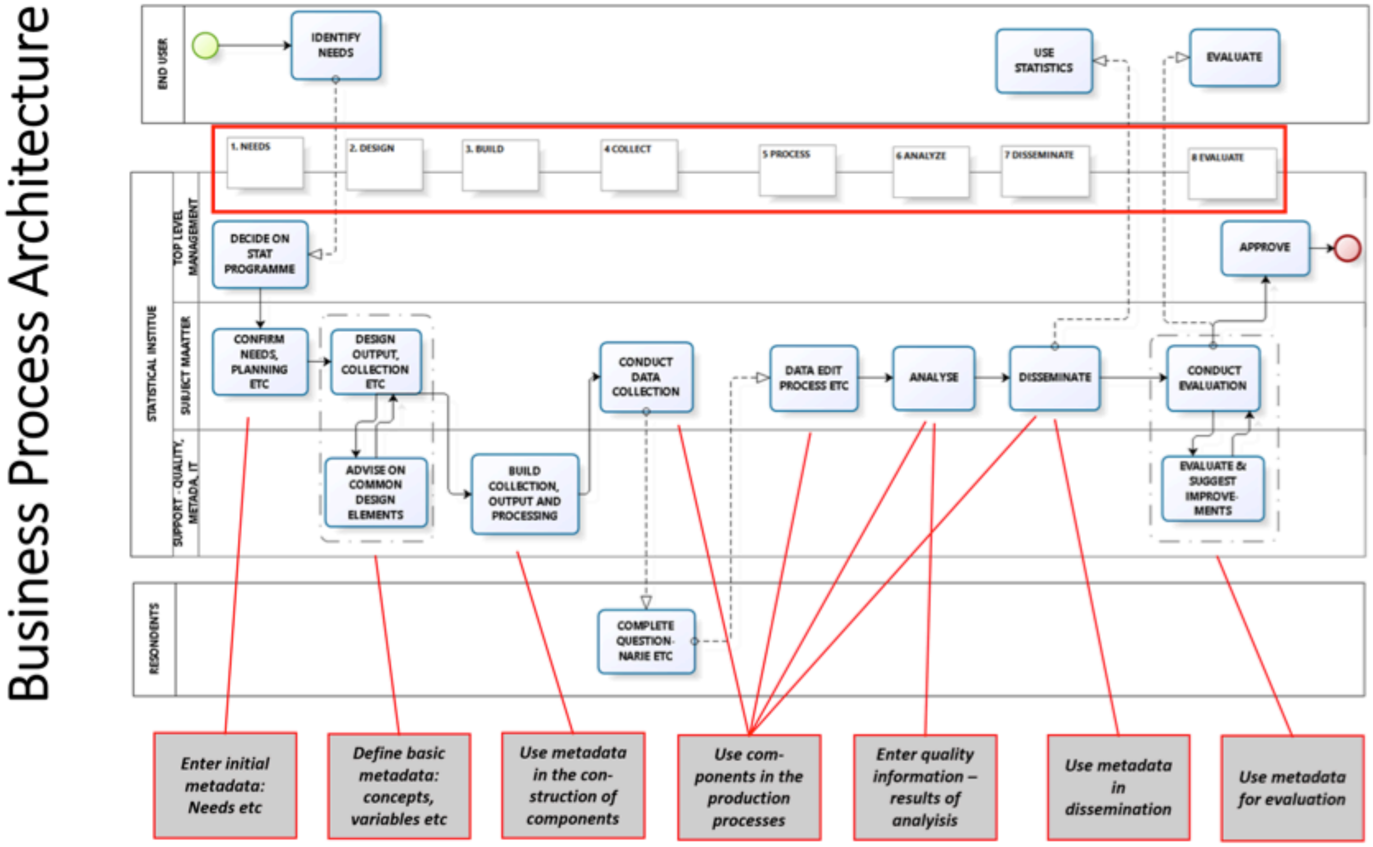

Business Process Architecture
END USER
IDENTIFY NEEDS
USE STATISTICS
EVALUATE
1. NEEDS
2. DESIGN
3. BUILD
4 COLLECT
5 PROCESS
6 ANALYZE
7 DISSEMINATE
8 EVALUATE
STATISTICAL INSTITUTE
TOP LEVEL MANAGEMENT
DECIDE ON STAT PROGRAMME
APPROVE
SUBJECT MAATTER
CONFIRM NEEDS, PLANNING ETC
DESIGN OUTPUT, COLLECTION ETC
CONDUCT DATA COLLECTION
DATA EDIT PROCESS ETC
ANALYSE
DISSEMINATE
CONDUCT EVALUATION
SUPPORT - QUALITY, METADA, IT
ADVISE ON COMMON DESIGN ELEMENTS
BUILD COLLECTION, OUTPUT AND PROCESSING
EVALUATE & SUGGEST IMPROVE-MENTS
RESPONDENTS
COMPLETE QUESTION-NARIE ETC
Enter initial metadata: Needs etc
Define basic metadata: concepts, variables etc
Use metadata in the con-struction of components
Use com-ponents in the production processes
Enter quality information – results of analisis
Use metadata in dissemination
Use metadata for evaluation

Figure 4.2 Diagram showing GSBPM business processes, workflow and the use of metadata

The overall idea is to start with users in the needs phase outlined in the GSBPM. The next phase starts with the design of the outputs to be disseminated in the dissemination phase. These inputs will drive what will need to be collected, derived, etc. during the statistical production. In this way what users need in the end (“documentation”) is driving the definition of metadata to be used during statistical collection, processing and analysis. With the structural design for “documentation” having been established early in the process, the assembling should become much more straightforward.

Another important point behind the diagram is the idea of metadata driven production [13] [14]. The Australian Bureau of Statistics define metadata driven production as ‘configurable, rule-based and modular ways of producing statistics’ [13]. The sub processes for is phase 4-8 are typically the starting point when designing and implementing components for the metadata driven production. Subject matter staff and support staff define metadata on variables, concepts business rules etc. in phase 2. These metadata are used in phase 3 in the construction of components. These component are ”plugged into” production process systems in phase 4 to 8..

Metadata and user needs

According to the model presented in section 3 it is important to handle frames of reference. How do we do that? In order to establish frames of reference we must have dialogues with users in workshops where we invite users, trying to achieve a double purpose. The first purpose is to learn as much as possible about how users wish to use statistics and the problems they seem to encounter. The second purpose is to make them understand terminology and ways to use metadata to improve their search and use of information, and allow them to better understand the way in which we handle metadata.

An example of the improved dissemination of metadata is the implementation of summary and detailed levels in the dissemination of quality information at *www.dst.dk*. In this way, we try to target both the general public and users who need detailed information (e.g. researchers). In general, we must establish processes that consider various users' input. This will primarily happen in GSBPM phase one (Needs) and in GSBPM phase seven (Disseminate). This aspect is discussed in several papers [1] [2] [3] including a model on how to find out about user needs for metadata.

Metadata and the implementation of a GSIM-compliant DDI model in Colectica

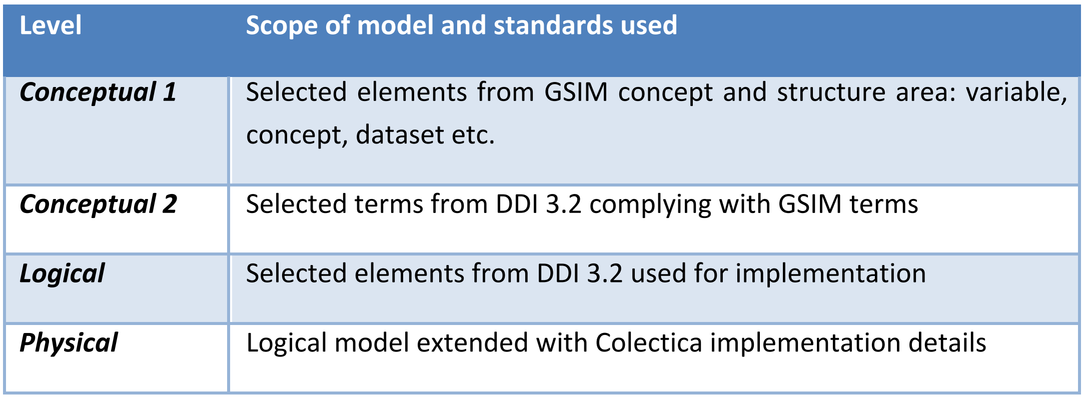
Introduction to mapping

The claim in the paper is that we need a) improvement and precision in the terminology we use when we talk about metadata, b) better understanding of the role of metadata in relation to our users, and c) better understanding of the role of metadata in production of statistics. The last two parts have been discussed in the previous sections. Besides these, there is a need for improvement and precision of terminology we use when creating and using metadata about statistics. This is where GSIM comes in. The GSIM model is complementary to GSPBM and together they aim to improve processes, communication and automation as depicted in figure 4.2. Another important aspect of GSIM is, that it works as a reference model of information objects that can aligned with the DDI and SDMX standards. This gives us the benefit of the work on these standards as well as access to standard software.

The aim of this section is to introduce GSIM and related standards in order to move towards a common terminology. The introduction is mainly complex metadata terminology for staff with knowledge of and experience in modelling statistical metadata.

As discussed in section 3, this must be supplemented by a more intuitive and narrow terminology that must be used in relation to end users. End user terminology must only include intuitively known terms respecting the end user frame of reference like statistical unit, property of a unit, population, unit-type, variables at the dataset, etc. As written above, the content and form of metadata depend on the common understanding of metadata terminology reached in discussions with end users, e.g., presentations in metadata portals.
It must also be noted that we are mainly focusing on a part of GSIM telling us about systems of statistics. This means that we mainly use elements like concepts, variables, classifications, populations, and unit type to name the most important terms. We are aware that GSIM has the ambition of having metadata about systems for producing statistics. In a broader scope this must be taken into account, e.g., how definitions of variables for output are derived from definitions of input variables.

The information objects in GSIM are described by definitions, attributes and relationships; however, the model is not fully developed in terms of attributes. Areas of the model such as classifications have enhanced sets of attributes whereas the model in other areas just includes the attribute's name and description. When aligning DDI with GSIM, we aim to use only DDI elements and associations which are present in GSIM. On the attribute level, this is not possible due to the simplicity of GSIM. We expect that the two models will become closer during the next couples of years, and by applying this strategy we will more easily be able to migrate towards newer versions of the standards, both at a logical and at a physical level.



Level	Scope of model and standards used
Conceptual 1	Selected elements from GSIM concept and structure area: variable, concept, dataset etc.
Conceptual 2	Selected terms from DDI 3.2 complying with GSIM terms
Logical	Selected elements from DDI 3.2 used for implementation
Physical	Logical model extended with Colectica implementation details

Figure 5.1 Modelling at various levels

The development of information-systems typically involves three levels of abstraction. These contain requirements of modelling, Use Case modelling, Class definition, etc. Modelling traditionally includes a three-tiered approach [11]:

- Conceptual Level – the basic entities of a proposed system and relationships between them
- Logical Level – specifies entities with a full set of attributes and their relationships without implementation details
- Physical Level – defines the physical structure for a technology specific format

The models at each of the three levels of abstraction correspond to Model Driven Architecture (MDA) concepts. In our implementation we have an additional top level, the GSIM conceptual model, which is mapped into a GSIM-compliant DDI conceptual model.

The process of implementing GSIM-compliant DDI in Colectica involves the following steps:

1) Mapping

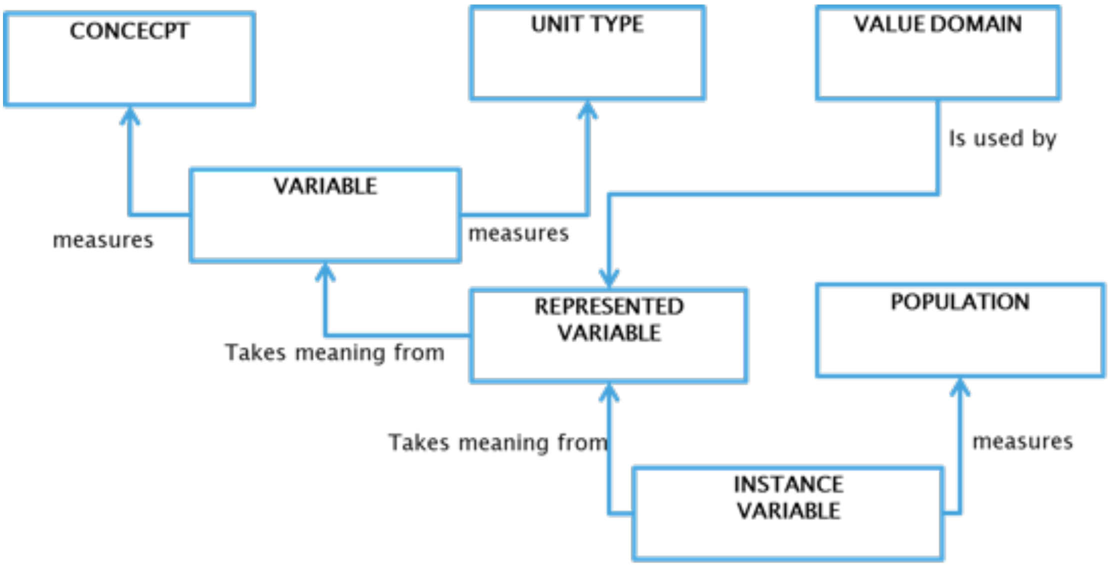


Figure 5.2 GSIM Variables

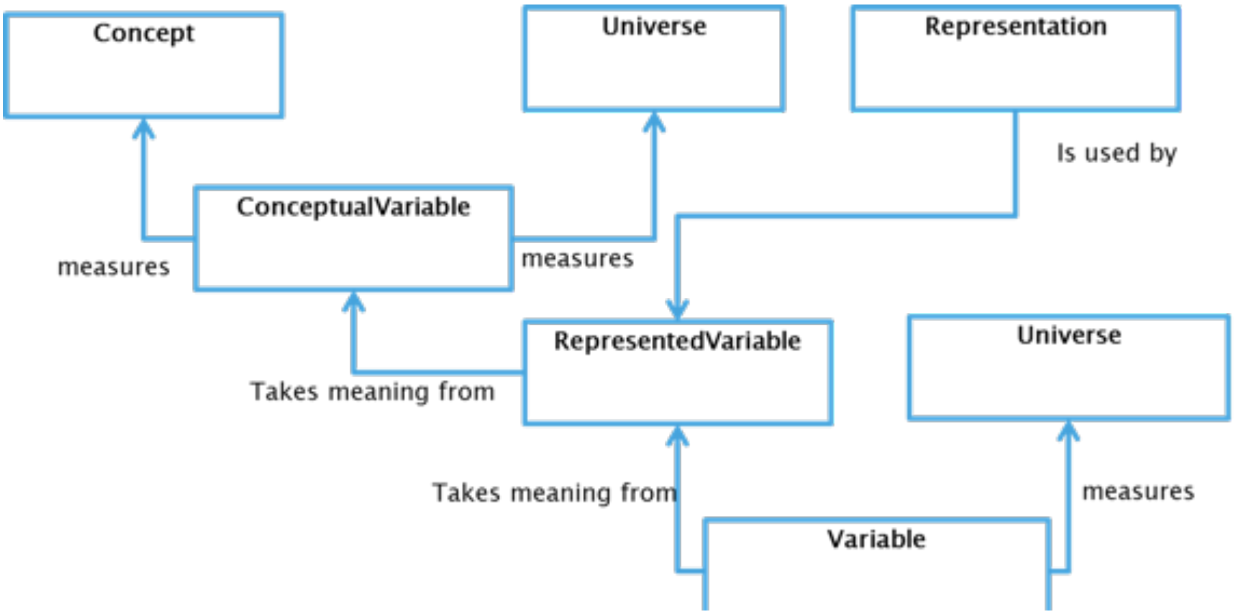


Figure 5.3 DD1 Variables aligned with GSIM

a. High-level mapping from GSIM to DDI: The purpose of the high-level mapping is to ensure compliance between GSIM and DDI in terms of association and cardinality. This mapping can be quite complex, as seen in classification. See paper about the Copenhagen mapping [12]. An easier example is mapping variables from GSIM. See figure 5.3 below.

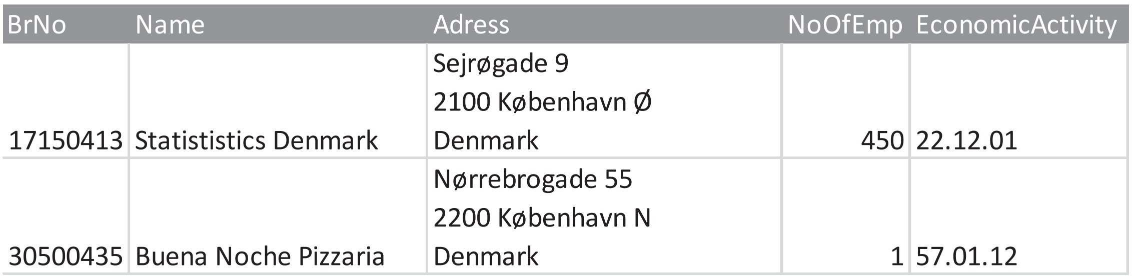
b. Low-level mapping, including adjustments from GSIM to DDI: Here the models are mapped on the attribute level. DDI provides some generic mechanisms for implementing user attributes.

2) Creating the Logical DDI 3.2 model. Once the mapping is in order, it is then possible to create the logical DDI model. Workflow-related logic is added.

3) Creating the Physical model: The GSIM-compliant DDI model implemented in Colectica is a common effort between Colectica and Statistics Denmark so that the revised DDI model in (2) is followed. The metadata also have to be organised physically in a way that facilitates reuse.

mapping

Fig 5.2 below shows selected GSIM objects and their relations. Figure 5.3 shows that the GSIM variables aligned with DDI 3.2 using the DDI terminology.



BrNo	Name	Adress	NoOfEmp	EconomicActivity
17150413	Statistics Denmark	Sejrøgade 9 2100 København Ø Denmark	450	22.12.01
30500435	Buena Noche Pizzaria	Nørrebrogade 55 2200 København N Denmark	1	57.01.12

Figure 5.4 BR_FROZEN_2014.XLSX

One should note that both Unit Type and Population are mapped into Universe. This is a fault in the DDI model as they are clearly not the same. In the GSIM model, the Instance variable is associated with the Represented variable which again is associated with the Variable.

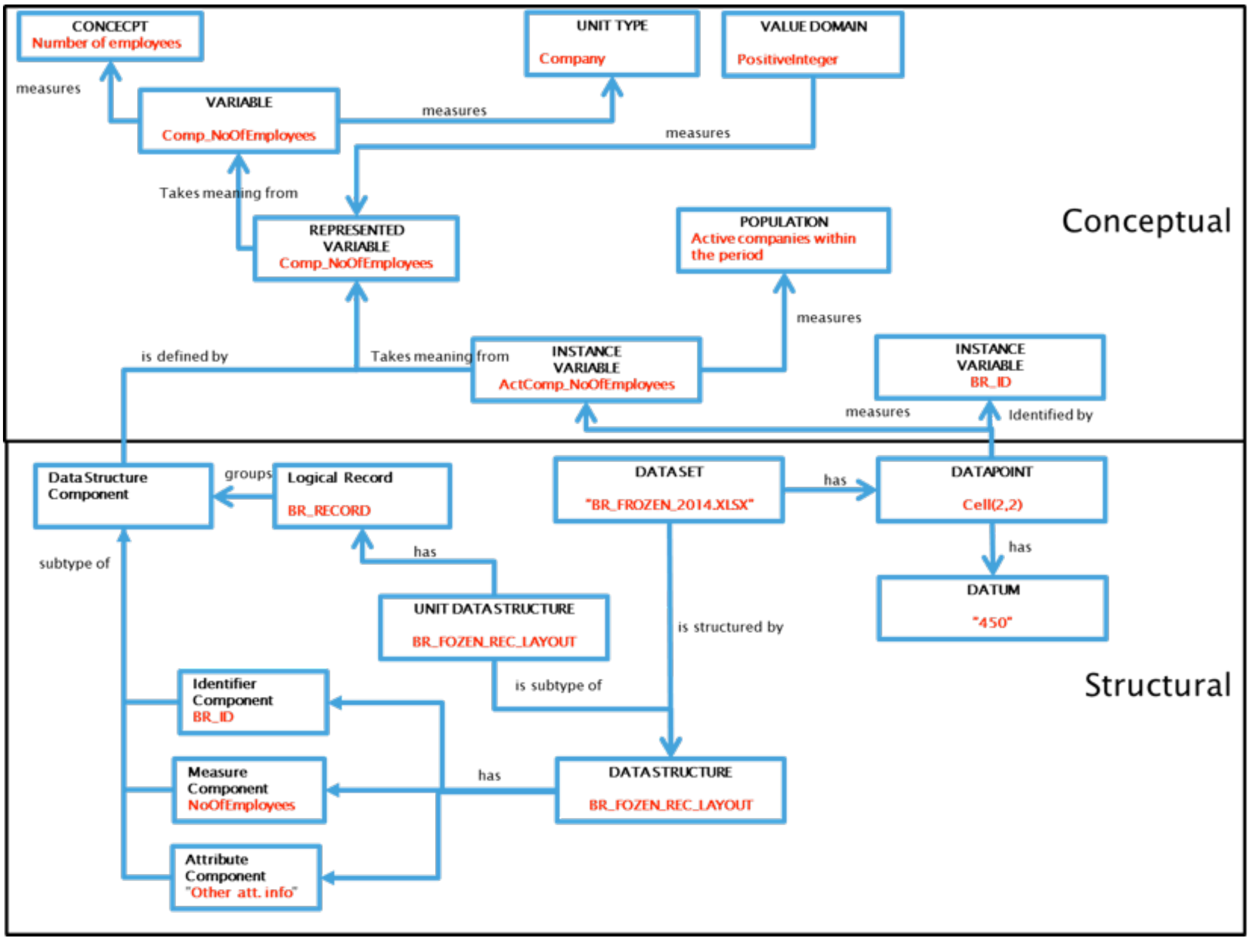
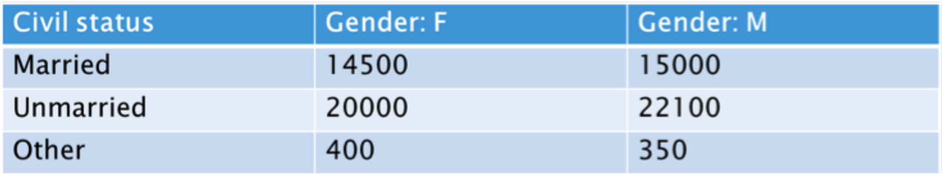


Figure 5.5 Conceptual model showing GISM concepts and example data from the Business Register

In DDI there is an additional association between Instance variable and Variable. Hence, we must prohibit the use of the additional association. In the following we will deal with the mapping from GSIM to DDI.

In order to become familiar with the terminology we will use two simple cases. Please note that for communication purposes the cases deal with physical representations rather than treating GSIM at a logical level. This masks how elements in data structures are reused but gives the reader a handle to a concrete example from the real world.

Case 1: Unit-dataset from the Business Register.
This is a case where we focus on microdata, which is called unit data in GSIM. A so-called "Frozen" version of the Business Register is disseminated in various forms once a year. One example would be an Excel spreadsheet for the year 2014, "sheet1" in the spreadsheet file; "BR_FROZEN_2014.XLSX". In Cell 2,4 a value of 450 is found, which is the number of employees at Statistics Denmark. The column header name is "NoOfEmp" . In addition, the dataset has columns with id, name, address and Economic activity.



Civil status	Gender: F	Gender: M
Married	14500	15000
Unmarried	20000	22100
Other	400	350

Fig 5.6 SC_2014.XLSX

Now we want metadata about the number 450. The diagram below shows selected information objects from GSIM. The GSIM concepts are shown with a black font and the object value is in red. For example, 450 is marked with red and put into the box called DATUM in the figure. Note also that the GSIM concepts are marked with bold in the text below.

So how are we to interpret the number 450? The GSIM model comes in handy here: In GSIM the value 450 is an attribute of the concept Datum. The Datum lives within a placeholder: the Data Point which lives in a Data Set structured by a Data Structure. For a Unit Data Set a Data Point will measure one unique Instance Variable, "ActComp_NoOfEmployees".

So now we have an intuitive understanding of what we are measuring, namely a number in a one-dimensional dataset that measures the instance variable "ActComp_NoOfEmployees". But we need more information and GSIM elegantly explains it all while making it all reusable.
To understand an Instance Variable it must be associated with a **Population** and a Representation, the **Represented Variable**. In this case the Population is "All active companies within the period" and the Represented Variable takes its meaning from "Comp_NoOfEmployees". To understand the representation of the Datum value is quite easy. This is not coded but simply a **Value Domain** of positive integers.

So are we there? Not yet. We need an explanation of the content of the variable "Comp_NoOfEmployees".

Every **Represented Variable** takes meaning from a conceptual **Variable**, which measures a **Concept** on a **Unit Type**. So now we are back to basics. When the statistics were designed years ago, someone expressed a wish to measure the number of employees on all active companies (the population) every year. They elaborated that a Company should be specified as a legal entity with a Business Register ID (The Unit Type). Furthermore, the **Concept** "Number of employees" should be specified as people working full-time all year.

We now have a conceptual interpretation for the Datum, but it would be nice to know which company we are dealing with (company identification). Let us therefore proceed to the structural part of GSIM.

A **Dataset** in GSIM is structured by a **Data Structure**, which contains **Data Structure Components**. In this case the Dataset is our Excel spreadsheet, which is structured by "BR_FROZEN_LAYOUT". The layout contains a Logical record, which is a reference to a data record independent of its physical location. The layout also points to three types of Data Structure Components: An **Identifier Component** is the unique identifier for the unit, here "BR_ID" with the value "17150413". The thing we are measuring is stored in the **Measure Component**, "Comp_NoOfEmployees".

With metadata about content, population, unit-type, etc., we now have information about value 450 in cell 2.4!

Case 2: A dimensional dataset from population
Data about the population in Denmark are disseminated in various forms. One version would be a dimensional dataset (a cube) for the year 2014, "sheet1" in the spreadsheet file "SC_2014.XLSX". In Cell 2.2 a value of 14500 is found. The column header name is "Gender", the first column is labelled Civil Status. The sheet name (not shown here) tells us the measure is LivingPersonsInCPH2014_NoOf

As in case 1 the selected objects are shown in the following object diagram where the GSIM concepts are with the colour black and the object value with the colour red. (see figure 5.7)

Now we want information about the number "14500". The understanding of a dimensional dataset is different from a unit dataset based on the data structure. As the Business Register example with unit data, the data point is interpreted by its instance, representation and

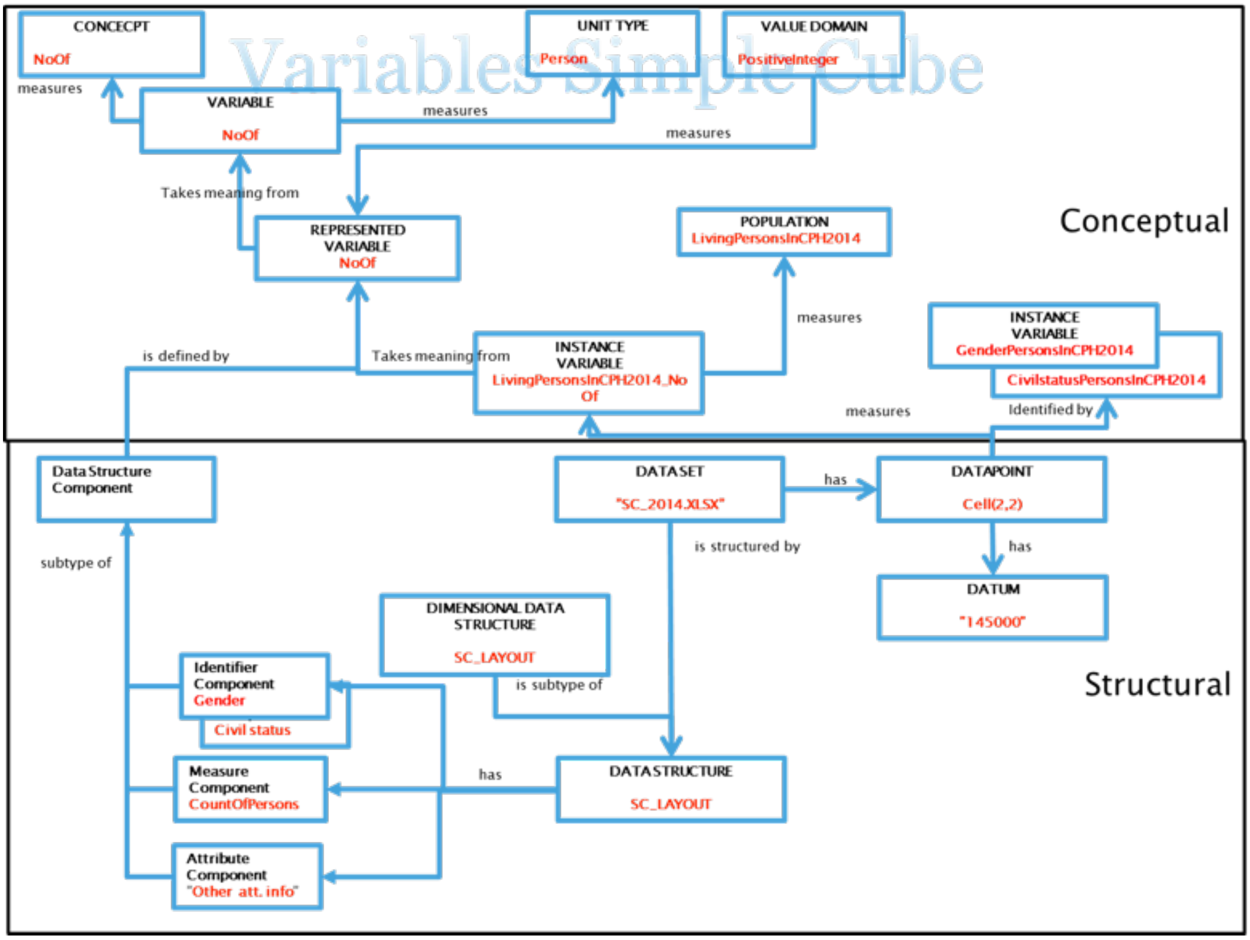


Figure 5.7 Conceptual model showing GISM concepts and example data about population

variable. For dimensional data, the measure component is connected to as many identifier components as the number of dimensions, each of which has its own representation and concept. This is shown in the figure below, which shows the conceptual part of GSIM for the two dimensions Civil status and Gender, (see figure 5.8).

So aside from understanding LivingPersonsInCph2014, we have to explain **Gender** *F* and **Civil status** *unmarried*.

The instance variable "*GenderPersonsInCPH2014*" takes its meaning from the represented variable "*PersonGenderRepresentation*" measured by a "*GenderCodelist*". The variable takes its meaning from the "*PersonGenderVariable*". Dealing with Civil status is equivalent.

Implementation of DDI and Colectica

In the start of 2015 Statistics Denmark completed the implementation of quality declarations in Colectica. In spring 2015, together with Colectica, we completed the modelling of classifications, which we are prototyping now. Feeling confident, we finally took the bold move to modelling and prototyping the conceptual part of GSIM variables as seen in our two cases for three statistics, and all of this is available in our coming portal. The structural part of linking it up with our statistics bank will be a new exciting project coming years.

Conclusion

In the beginning of the paper, we claimed that the road towards common metadata in a statistical context requires a) improvement and precision in the terminology that is used when we talk about metadata b) better understanding of the role of metadata in relation to our users, and c) greater recognition of the role of metadata in the production of statistics.

Based on this, we went through the history of metadata at Statistics Denmark and presented the challenges we have had with the existing situation of the metadata work. It was shown that widespread myths about metadata make it difficult for both producers and user of metadata to benefit from common metadata.
For example, a widespread opinion regarding metadata is that documentation is something you attach to your statistical products after they have been published. There is only a little reflection on how the statistics and documentation about the statistics are produced and should be used. User needs are often introduced too late in the production process or without the efforts necessary. In addition, there is little awareness about using common models, both in terms of precise definitions of information objects like population, statistical unit, etc., but also in how to apply common work procedures introduced in GSBPM.

In order to pave the road towards common metadata we gave a helicopter perspective on how to approach metadata focusing on the importance of metadata understood as a frame of reference.

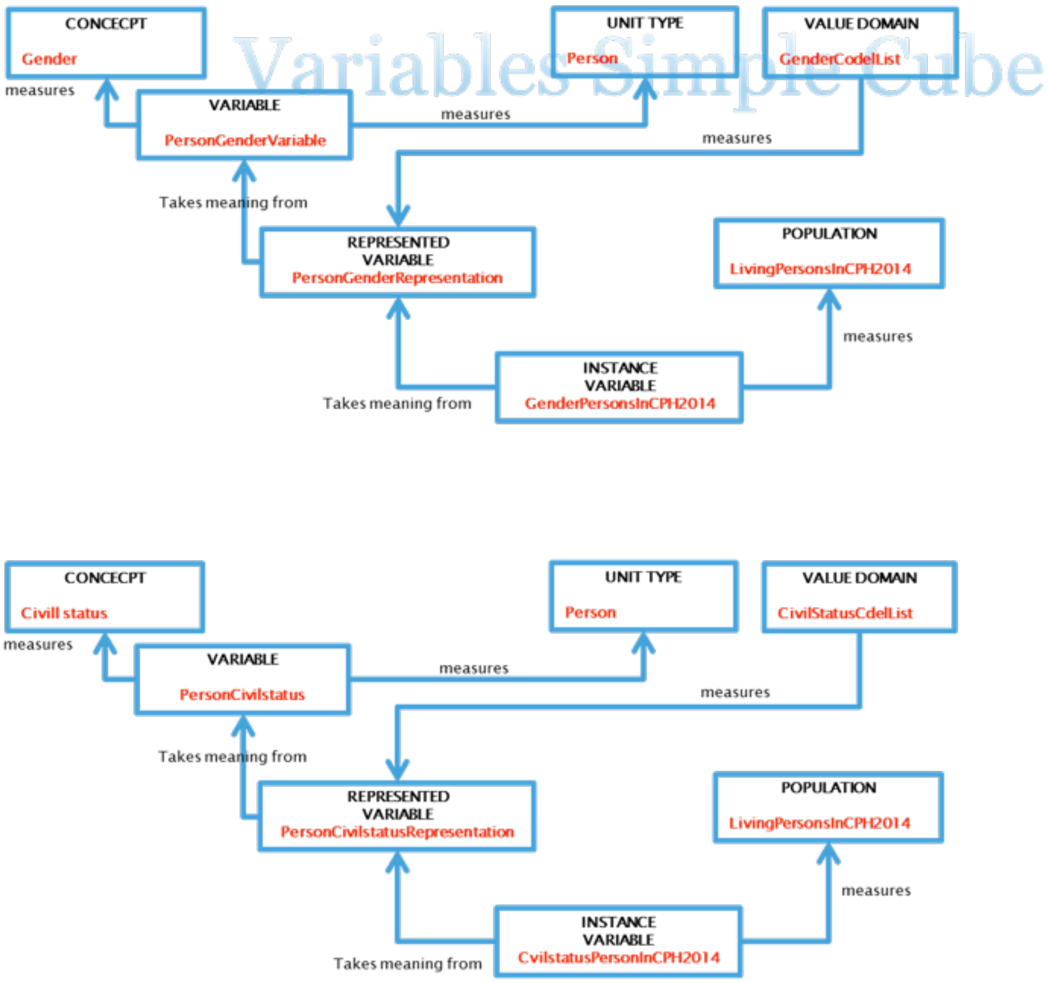


Figure 5.8 Gender and Civil status variables

Our claims regarding better understanding of the role of metadata in relation to our users and better understanding of the role of metadata in the production of statistics were dealt with in section 4. We stressed the importance of business process modelling and how metadata should be integrated into the business processes. This included a short note about metadata-driven production. Based on this, we talked about the importance of metadata in relation to users and how these aspects of metadata should be closely connected to what is going on in the business processes when talking about needs and dissemination.

Section 5 dealt with the last part of our claim, namely the need for improvement and precision in the terminology used when we talk about metadata. As with all areas of "modern life" the use of terminology plays an important role as well. We gave a longer introduction to models and terminology used in GSIM mainly for information model experts. We introduced the GSIM model as a complementary model to GSPBM. Together they aim to improve processes and communication, but also to support IT development and automation.

The introduction in section 5 in this paper contains information models for people with a frame of reference that requires knowledge of terminology and experience in modelling metadata. This is only one kind of use that can be distinguished from a more intuitive and narrow terminology that must be used in relation to end users. End-user metadata terminology must only include intuitively known terms respecting the end user frame of reference like statistical unit, property of a unit, population, unit-type, variables at the dataset, etc.

Building and implementing common metadata requires that many disciplines must be in play. In the paper, we argued in favor of a need for improved understanding of the role of metadata in relation to users, metadata in relation to production processes and improvement and precision and communication of common terminology. To succeed in creating common metadata requires that these three aspects are well elaborated and communicated by users and producers of metadata.

References

[1] Thygesen, Lars (2013) and Nielsen, Mogens Grosen; How to fulfil user needs – from industrial production of statistics to production of knowledge. Statistical Journal of the IAOS, Volume 29, Number 4 / 2013. IAOS Press

[2] Nielsen, Mogens Grosen and Thygesen, Lars (2011) How do end users of statistics want metadata? Paper at Metis workshop on Statistical Metadata, 5-7 October 2011

[3] Nielsen, Mogens Grosen and Thygesen, Lars (2014) Implementation of Eurostat Quality Declarations at Statistics Denmark with cost-effective use of standards. Paper presented at European Conference on Quality in Official Statistics, Vieanna 2-5 June 2014.

[4] Colectica - a tool for statistical metadata, developed by Colectica. Link: *www.colectica.com*

[5] Sundgren, B. (2004b). Statistical systems – some fundamentals. Statistics Sweden

[6] Langefors B. (1995). Essays on Infology - Summing up and Planning for the Future. Lund: Studentlitteratur

[7] Espejo, Raul (2000): Self-construction of desirable social systems in Kybernetes, Vol. 29 no. 7/8, MCB University Press

[8] Bednar, Peter M (2000) A Contextual Integration of Individual and Organizational Learning Perspectives as Part of IS Analysis, School of Computing and Management Sciences, Sheffield Hallam University Department of Informatics, Lund University, Vol 3, no. 3

[9] Porter, Michael (1985). Competitive Advantage: Creating and Sustaining Superior Performance, The Free Press

[10] Harmon, Paul (2007). Business Process Change – A Guide for Business Process Managers and BPM and Six Sigma Professionals. Massachusetts, USA.

[11] Sparx Systems (2011). From Conceptual Model to DBMS, Link to website: *http://community.sparxsystems.com/white-papers/669-data-modeling-from-conceptual-model-to-dbms*

[12] Iverson, Jeremy; Mogens Grosen Nielsen & Dan Smith (2014). The Copenhagen Mapping implementing the GSIM statistical classifications model with DDI lifecycle. Link: *http://cdn.colectica.com/TheCopenhagenMapping-Draft.pdf*

[13] Aurito Rivera, ABS; Simon Wall, ABS; Michael Glasson, ABS: "Metadata driven business process in the Australien Bureau of Statistics". Work Session on Statistical Metadata (Geneva, Switzerland 6-8 May 2013)

[14] Pedro Revilla, José Luis Maldonado, Francisco Hernández and José Manuel Bercebal National Statistical Institute, Spain. "Implementing a corporate-wide metadata driven production process at INE Spain". Work Session on Statistical Metadata (Geneva, Switzerland 6-8 May 2013)

Notes

[1] Mogens Grosen Nielsen, Chief Adviser, mgn@*dst..dk* and Flemming Dannevang, Senior Adviser, fda@*dst..dk* are employed at Statistics Denmark. Mogens Grosen Nielsen is head of metadata. Flemming Dannevang is working with the metadata task and other tasks.