UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE Working paper
CONFERENCE OF EUROPEAN STATISTICIANS 25 August 2016

Workshop on Implementing Standards for Statistical Modernisation,
21 – 23 September 2016

Fulfilling user-needs, improving quality and efficiency using GSIM and other standards.

Mogens Grosen Nielsen*
mgn@dst.dk
*Chief Adviser, Statistics Denmark

Abstract: 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. The modelling work aims at building models compliant with GSIM. As in all other metadata systems, the important part is to work towards reuse of common metadata. By common metadata, we understand metadata about concepts, classifications, variables that are stored only in one place, linked together, and reused across various domains and in all relevant business processes, including dissemination processes. The implementation of this idea introduces a lot of complexity and requires effort from various disciplines and from the whole organisation. The claim in this paper is that the path towards common metadata, fulfilment of user-needs, improved quality and efficiency require 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) better understanding of the role of metadata in the production processes and d) the use of common overall models for business process change and Enterprise Architecture to ensure successful implementation.

1. 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 about 300 statistics. The ongoing work focuses on variables, concepts and classifications. The modelling work aims at building models compliant with GSIM.

As in all other metadata systems, the important part is to work towards reuse of common metadata. By common metadata, we understand metadata about concepts, classifications, variables that are stored only in one place, linked together, and reused across various domains and in all relevant business processes, including dissemination processes. Descriptions and definitions of concepts such as marital status, firm and income are examples on common metadata. In our solution, we combine common metadata and local metadata that are specific for a statistical domain. Both common and local metadata are stored in the metadata-system. An example of this is the linking of the common concept marital status to marital status as variables in concrete dataset across waves and domains.

Metadata introduces a lot of complexity and requires effort from various disciplines and from the whole organisation. The claim in this paper is that the path towards common metadata in a statistical context requires a) improvement and precision in the terminology used when we talk about metadata, b) better understanding of the role of metadata in production of statistics, and lastly c) better understanding of the role of metadata in relation to our users. d) use of common overall models for business process change and Enterprise Architecture to ensure successful implementation.

In section 2, we first give a short history on the development of metadata at Statistics Denmark. The section then gives a description of the challenges with regard to opinions on metadata. We also touch upon the work

on the introduction of the Generic Statistical Information Model (GSIM) and the Generic Statistical Business Process Model (GSBPM).

In section 3, we give a helicopter perspective on metadata in order to give an idea on how to approach metadata, focusing on metadata as frames of reference.

In section 4, we introduce metadata and business processes. The chapter starts with a general introduction about modernisation, value chains, including how we can move towards process centric organisations. Hereafter, we introduce two widespread and tested models on methodology. The first model is about business process management. We are introducing a model from the book Business Process Change by Paul Harmon [10]. The second model is an Enterprise Architecture model. We are introducing a tailored version of The Open Group Architecture Framework TOGAF [19]. We close the section with ideas and models on how to integrate metadata in business processes and how to use metadata related to user needs.

Section 5 presents models and terminology used in GSIM. The section shows our approach to introduce not only GSIM, but also moving towards DDI 3.2 and its implementation in the software package Colectica [4].

Section 6 of the paper aims at putting the pieces together showing how we use the models standards and insight described in this paper in an on-going project.

The conclusion closed the paper stressing the importance of standards and the benefit from the Nordic cooperation.

2. Metadata at Statistics Denmark: present situation and challenges

For several years, Statistics Denmark operated separate systems for classifications, quality declarations, variables and concepts. Some of the systems 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 creation and definition of common metadata. On the dissemination side, we wanted to use metadata to give users easier access to our products. We wanted to direct users to the right products, but also to give users better information about our products, once they found them.

At a UNECE Metis meeting in Geneva in 2011[1], it became clear to us that DDI could help us with the integration, reuse and dissemination of metadata. From 2012 to 2015, we implemented a DDI based model using Colectica as our standard-software. Besides the DDI-standard, we introduced the SDMX reference metadata on quality complying with Single Integrated Metadata Structure (SIMS), which Eurostat was developing at the same time. In January 2015, a quality management system and a standardised description of quality were in place for about 300 statistics. Today, we publish all as an integrated part of our dissemination system.

In spring 2015, the top management at Statistics Denmark approved a new strategy on quality and metadata. 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.

Since then, we have introduced new projects as part of the implementation. We have used parts of GSIM on concepts, variables and classifications. Statistics Denmark is now introducing terminology and implementing GSIM using DDI and Colectica. We work closely together with our colleagues in the other Nordic countries on clarifying various aspects in a common overall model.

In the presently 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 a challenge to allocate resources and to change processes towards using common standards, including getting rid of some stovepipe production.

There has been a widespread opinion among statisticians that documentation is something you attach to your statistical products after the statistician has published the statistics. Statisticians often introduce user-needs

[1] Statistics Denmark was represented by Lars Thygesen and Mogens Grosen Nielsen. Link: http://www.unece.org/stats/documents/2011.10.metis.html#/

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 on using common models, both in terms of precise definitions of information objects like population, statistical unit etc., but also on how to apply common work procedures introduced in GSBPM.

3. The helicopter perspective: metadata as compatible frames of reference

The starting point for this section draws on articles by Bo Sundgren who analysed processes of understanding of data and information and the creation of knowledge.

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 contents of 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]:

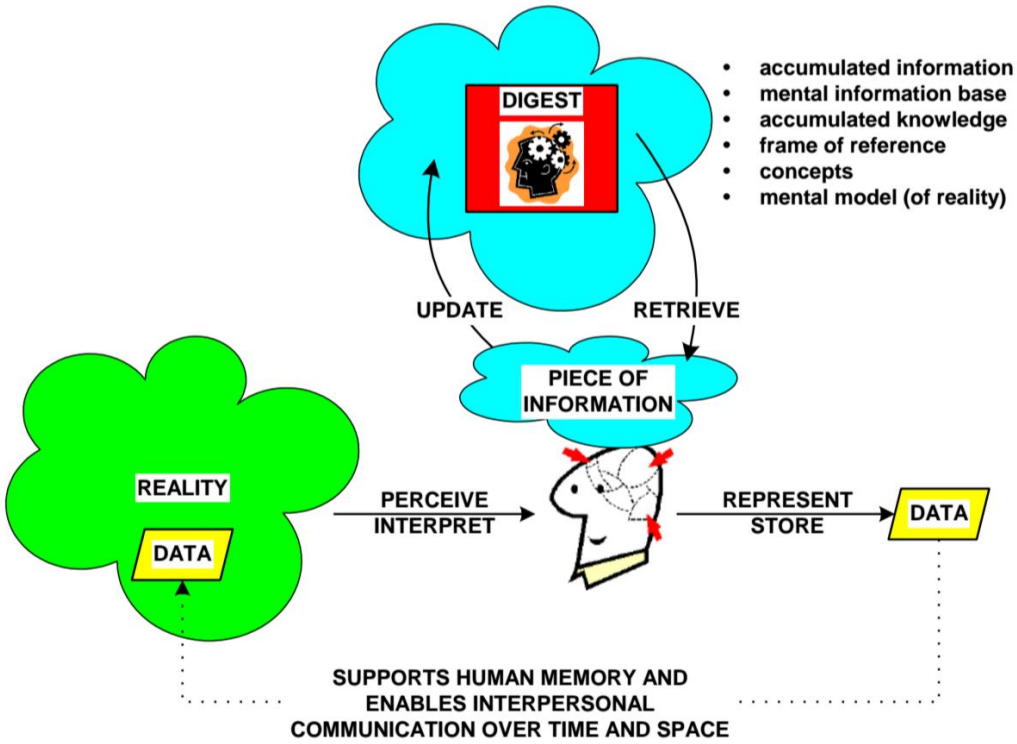


Figure 3.1 Reality, information and data (from [5]).

In short, the mental process can be expressed by the following equation introduced by Börje 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 is the role of common and reusable metadata, which we are aiming at?

Bo Sundgren states *"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 then to ensure compatible frames of reference. It is here that metadata or data about data enters the scene. Metadata can help user and producers on to road towards common interpretation of data. The NSI's should construct the needed foundation for metadata and communicate metadata between users. Metadata thus have 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. Bo Sundgren notes that communication about metadata shares the same difficulties as communication in general.

"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]

Difficulties in communication occur at all levels from global communication at the Internet 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 the organisational aspect is maybe 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.

	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.

In section 5, we will introduce complex terminology on metadata. We do this by introducing GSIM and related levels towards a concrete implementation.

4. 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 inside the production and dissemination of statistics in such a way, that users will understand the result and be able to use and combine statistics across different domains? In order to move in this direction we must make sure that statistics are not produced in silos, and this has a bearing on the organisational diagram.

The sections below first introduce statistical modernisation. Hereafter follows various models focusing and moving towards process-centric organisation. The final section focuses on methodology.

4.1 Modernisation of the statistical production

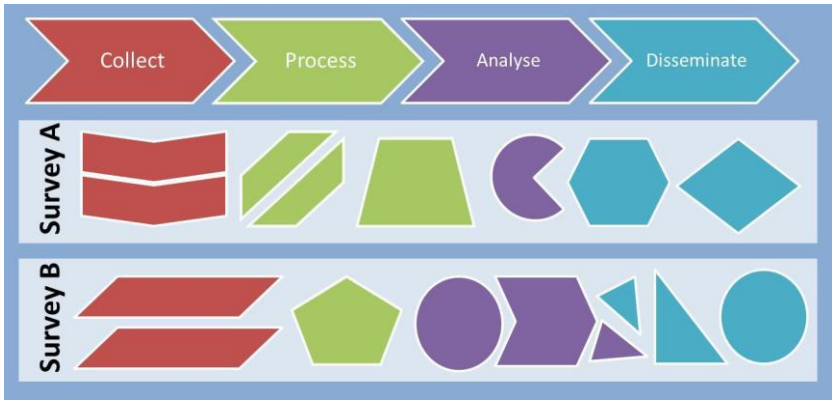
All statistical organisations around the world have since their creation been working on applying the best methods, the best technology, and best way to organise. The work in the field of processes, quality and metadata has in the last 20-30 years focused on the use of common standards. These activities have speeded up in the last 5-10 years by the creation of an international High Level Group and many other initiatives. This is due to the observation that all statistical organisations are facing the same challenges:

- New demands for statistical information including more statistical indicators, more sectorial and territorial details, improved timeliness and better quality.
- Governments need help on formulation of good policy, not only at national level
- Reduction of financial resources
- The ICT development and the explosion of the amount of information available via Internet.

The UNECE Conference of European Statisticians (more than 60 countries worldwide) established The High Level Group in 2010 with the following objectives:

- *"To promote common standards, models, tools and methods to support the modernisation of official statistics;*
- *To drive new developments in the production, organisation and products of official statistics, ensuring effective coordination and information sharing within official statistics, and with relevant external bodies;*
- *To advise the Bureau of the CES on the direction of strategic developments in the modernisation of official statistics, and ensure that there is a maximum of convergence and coordination within the statistical "industry".*" [2]

The main idea behind the modernisation efforts is often illustrated with the following two diagrams.



[2] Link: http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics

Figure 4.1. as-is situations with specialised business processes, methods and IT systems for each survey / output.

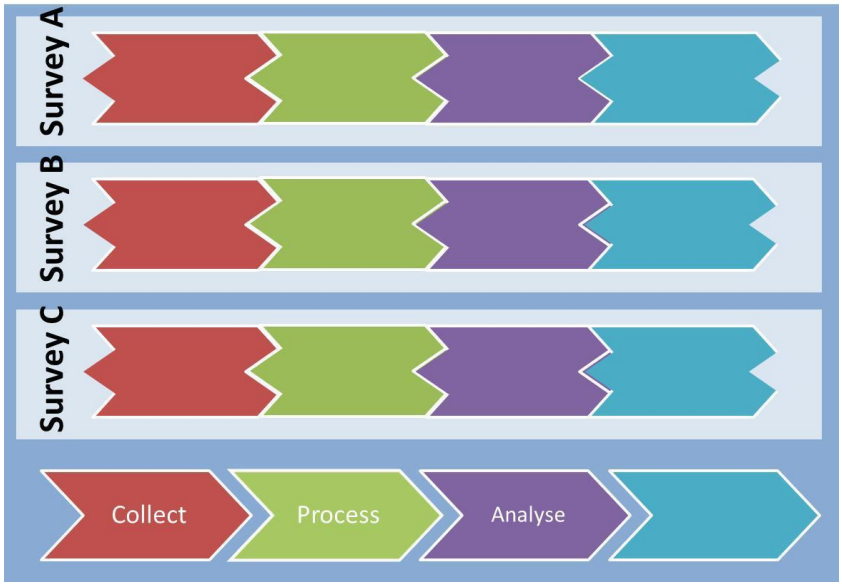


Figure 4.2. to-be situation with the result of the standardisation within an organisation

The first diagram shows the "as-is" situations with specialised business processes, methods and IT systems for each survey / output. The second diagram shows the "to-be" situation with the result of the standardisation within an organisation. In order to reach this within and across organisations, a lot of work on standardisation is needed.

4.2 From silos to process centric organisations

An important aspect of the discussion of modernisation of statistics focuses on the problem of organisation and thinking in silos. This problem is not limited to statistics. It is present in many other types of organisations. At the moment, there is a discussion at the Danish hospitals on how to improve the quality of treatment with the patient in focus. The top management introduced a rule, that all patients must have only one doctor as contact. This did not work, due to budget constraint in various departments. Now the management is discussing a new solution on how to ensure cross-cutting mechanisms to improve the patient's treatment, including the patient's perception of the treatment.

This discussion is present in the literature on business process management as well. *"... most companies had focused on dividing processes into specific activities that were assigned to specific departments. Each department developed its own standards and procedures to manage the activities delegated to it. Along the way, in many cases, departments became focused on doing their own activities in their own way, without much regard for the overall process. This is often referred to as silo thinking, an image that suggests that each department on the organisation chart is its own isolated silo*" [10].

As a reaction to silo thinking, we see a movement toward to dynamic "social-system thinking". Regarding processes, the change has gone from focusing on optimizing stable processes to focusing on what users need, and how the organisations can adjust their processes accordingly. Assembly lines at the Ford T factory are an example of the former way of organizing. Dynamic production lines in the present-day Toyota factories are an example of the latter type of business processes. The focus in the former is on how to split the business processes into separate vertical functional units. The focus in the latter is how to build dynamic business processes as value-chains based on user needs and feedback from users.

The figure below stresses the idea of value-chain and system thinking. A central aspect in system thinking is that organisations continuously must make self-corrections based on feedback from users.

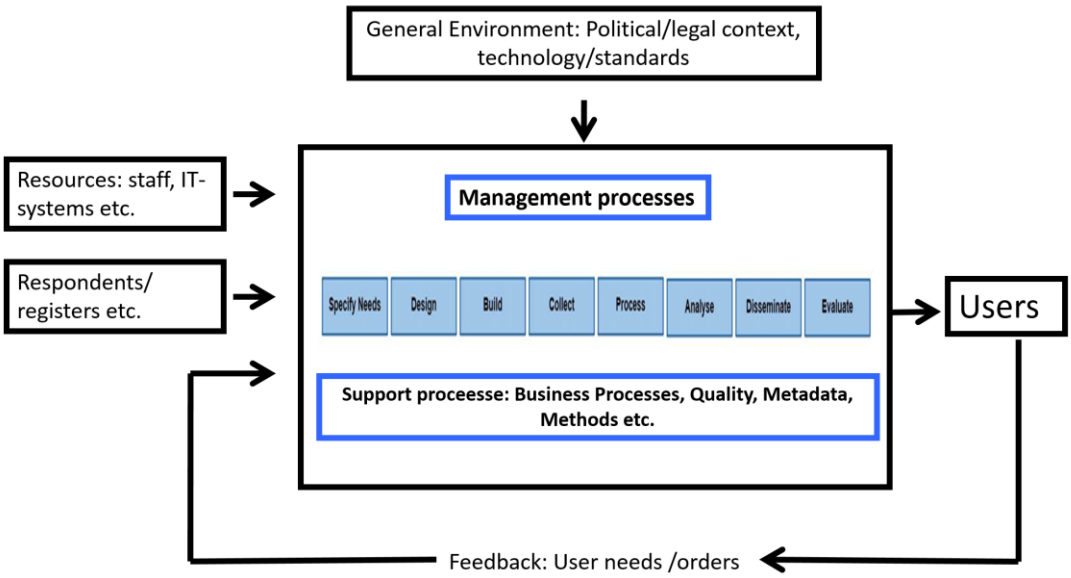


Figure 4.3 Business process perspective with environment elements.

The construction of the model is inspired by the idea of value-chains introduced by Porter in the 1980's [9] and ideas on system-thinking, including focus on feedback-mechanisms, in business process management [10]. In essence, the systems perspective emphasises that everything is connected to everything else and that it is often worthwhile to model businesses and processes in terms of flows and feedback loops. Systems thinking emphasises that any given employee or unit or activity is part of a larger entity and that ultimately those entities, working together, have to be justified by the results they produce for the user.

The value-chain thinking implies that you organise all processes from start to finish in such a way that each process adds value based on input and feedback from users. The ideas of using value-chain and system thinking 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 implementation can be found in Canada, Australia and Sweden. They have all put many resources into building advanced cross-department solutions. The challenge for smaller NSIs with less or very few resources is how to benefit from these ideas using standards and standard solutions, taking small manageable steps in the right direction.

4.3 Integration of metadata in production processes

Having the value chain thinking in mind, we must integrate metadata in production processes in order to obtain processes that are more efficient and improve fulfilment of user needs. The diagram below shows 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.

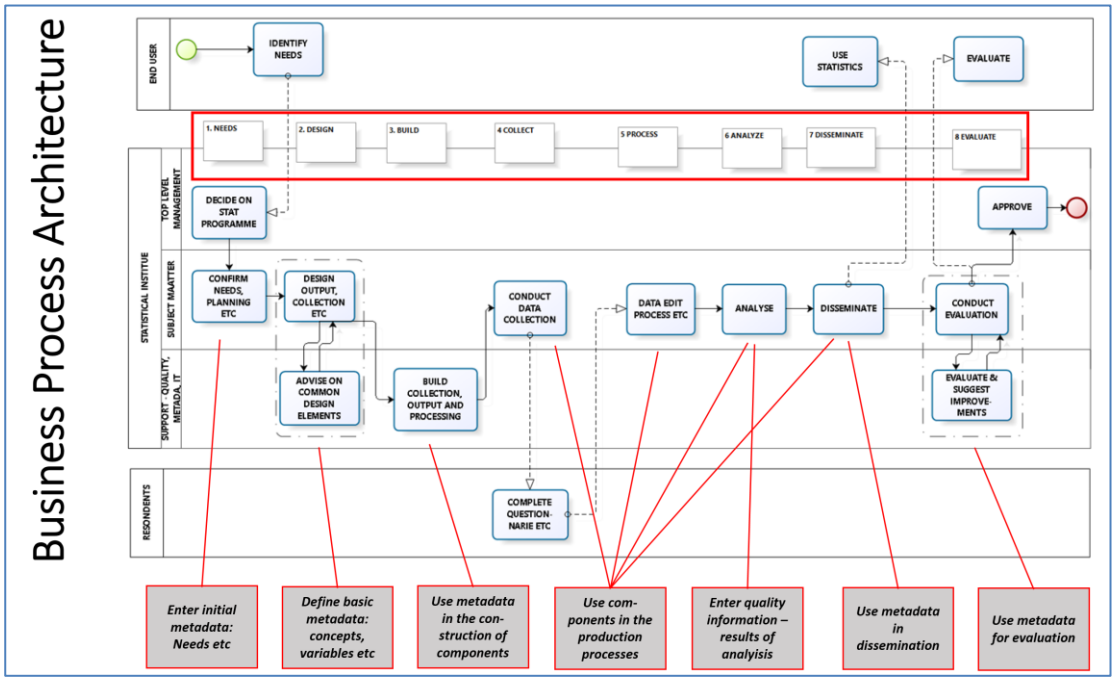


Figure 4.4. Diagram showing GSBPM business processes, workflow and the use of metadata

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 is driving the definition of metadata to be used during statistical collection, processing, analysis and dissemination. 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 [20] [21]. The Australian Bureau of Statistics defines metadata driven production as 'configurable, rule-based and modular ways of producing statistics' [20]. The sub-processes for phases 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. The components are "plugged into" production process systems in phases 4 to 8.

4.4 Metadata and user needs

According to the model presented in figure 3.1, 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 Statistics Denmark's web site 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.

4.5 Methodology: Business Process Management and Enterprise architecture

Moving towards a process centric organisation is a difficult task. It requires a coordinated effort from many disciplines. It is crucial to apply good methodologies that have proved their value in practice. In the on-going

project, we introduced two widespread and tested models on methodology. The first model is a tailored version on Business Process Management. Paul Harmon describes the full model in the book "Business Process Change – a guide for Business Managers and BPM and Six Sigma Professionals" [10]. The second model is an Enterprise Architecture model. We are using a tailored version of The Open Group Architecture Framework TOGAF [19].

The figure below shows a model of how we use the BPM methodology.

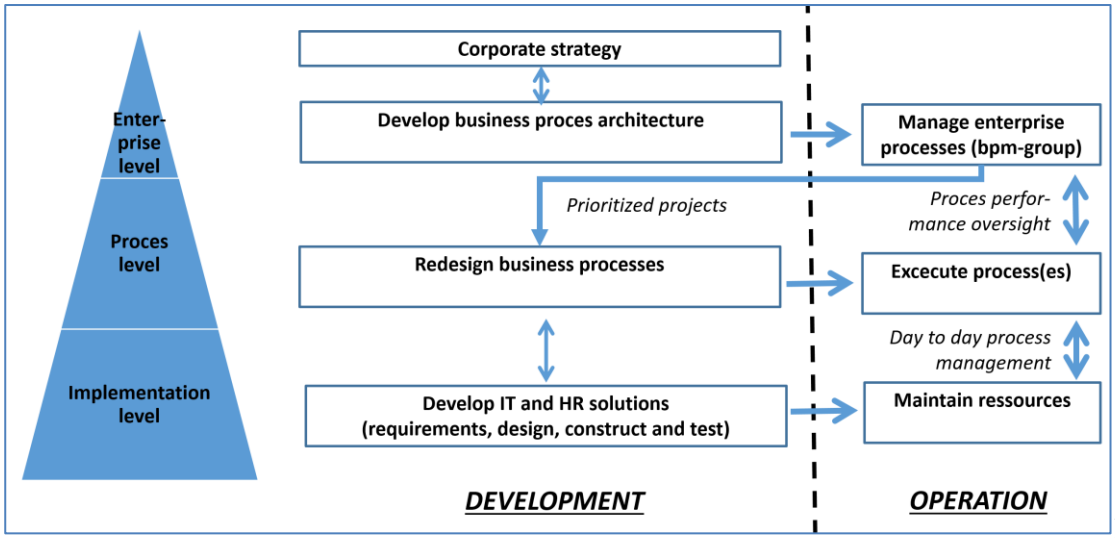


Figure 4.5 Model showing main activities in the BPM methodology

The model is introduced as part of the on-going project on metadata. Statistics Denmark will in the coming months discuss this model at workshops on the use of standards at the metadata project.

The model distinguishes between activities related to development and operation. It operates at three levels:

Enterprise level: The main activity at this level is to define business processes including performance measures. This activity covers all business processes. At this level, executives shift from focusing on specific processes to focus on the entire organisation as a system of interacting processes. They are working to maximise the effectiveness of the whole system. This includes giving responsibilities for managing and measuring the processes in the organisation. This entire set of models and measures and resources assigned to them is referred to as the *business process architecture.*

Business process level: The activities at this level include identifying, documenting, modeling, analyzing, redesigning and innovating specific value streams or individual processes to improve process performance. The main activity at this level is to suggest and redesign specific business processes. This activity covers a single or a group of specific business processes. You start by investigating and describing the existing processes based on the goals you want to achieve. The goals may improve documentation and sharing of knowledge, introduction of a quality framework, improving the use and management of metadata, better use of IT including standardisation and automation, improved organisation, better reactions to user needs, improved methods, etc.

Implementation level: At this level HR and IT teams design and implement human systems and software systems to implement business processes. Many available methodologies are appropriate at the Implementation level

When redesigning processes it is important to handle and coordinate changes in business and IT. In the on-going project we use an Enterprise Architecture Framework in order to handle the complexity that is being introduced when e.g. building software-components that can be reused across many domains. We use The Open Group Architecture Framework. (TOGAF). TOGAF is a widely used global architecture model. It offers a systematic methodology to handle changes in business and IT. In addition, knowledge on TOGAF can help us when discussing the use of the European Enterprise Reference Framework (EARF) [14] in cooperation with Eurostat.

We have studied this framework and we expect no difficulties in aligning with this framework, since Eurostat has aligned EARF with TOGAF and several other models and standards that we are introducing as well.

"The work at hand draws on and borrows as much as possible concepts and definitions of the widely recognised Open Group Enterprise Architecture Framework TOGAF, ESS EARF also builds upon the various results from initiatives in the official statistics industry: GSBPM (Version 5.0), GSIM (Version 1.1) and GAMSO (Version 1.0) have been used as a reference throughout the ESS EARF. The ESS Business Capability model is inspired by GSIM (Version 1.1) and GAMSO (Version 1.0), which defines the activities that take place within a typical statistical organization generalizing GSBPM with an extension in the areas of "Strategy", "Capability" and "Corporate Support"." [14]

The aim at Statistics Denmark is to use a simple model that is easy to communicate and easy to use. We are therefore implementing a tailored version of TOGAF that matches the specific situation in Statistics Denmark. It is also important to stress that TOGAF can be used *either* as a general architecture framework for all domains and activities in an organisation *or* it can be used for a subdomain. In the on-going project, we are using TOGAF for the metadata sub-domain.

As part of our tailored version of TOGAF we prepare a business architecture document with the following elements:

1. Vision, business principles and business goals
2. Scope
3. Stakeholders
4. Solution concept
5. Existing and future architecture
6. Gap analysis - and candidate roadmaps
7. Roadmap (projects on developing and implementation of solution)

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

The three claims in the paper are that we need a) improvement and precision in the terminology used when we talk about metadata, and lastly c) better understanding of the role of metadata in relation to our users.

With regard to improvement and precision of the metadata terminology, we have used GSIM as the starting point. The GSIM model is complementary to GSPBM and together they aim to improve processes, communication and automation. Another important aspect of GSIM is that it works as umbrella covering the implementation of metadata using the DDI and the SDMX standards such as Eurostat's reporting formats ESMS and ESQRS. This gives us the benefit of the work being carried out worldwide on these standards, as well as access to standard software.

5.1 Introduction to standards for information objects and mapping[3]

The Generic Statistical information model (GSIM) is the first internationally endorsed reference framework for statistical information. This overarching conceptual framework will play an important part in modernizing, streamlining and aligning the standards and production associated with official statistics at both national and international levels.

The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioural, and economic sciences. Do we now have one additional standard to manage? The answer is no, the DDI community has developed (and will develop) the DDI-standard so that it aligns with GSIM.

[3] Parts of this work can be found in a paper by Mogens Grosen Nielsen and Flemming Dannevang presented at the European DDI conference in December 2015 [12]

GSIM is complex and covers a lot, including how to model production and management processes. In this paper, we mainly use elements like concepts, variables, classifications, populations, unit type to name the most important.

When aligning DDI with GSIM, we aim to use DDI elements and associations, which we can map from GSIM to DDI. 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.

The development of information models typically involves modelling at three levels [11].

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

Table 5.1 shows the levels we are using, including the type of standards we use at each 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

Table 5.1 Modelling at various levels

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

1) Mapping
    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.
    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.

5.2 Mapping from conceptual to logical and physical level

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

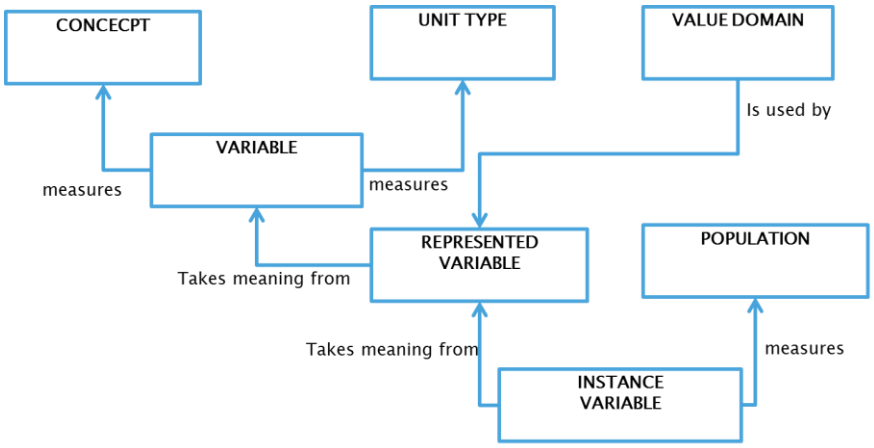


Figure 5.2 GSIM Variables

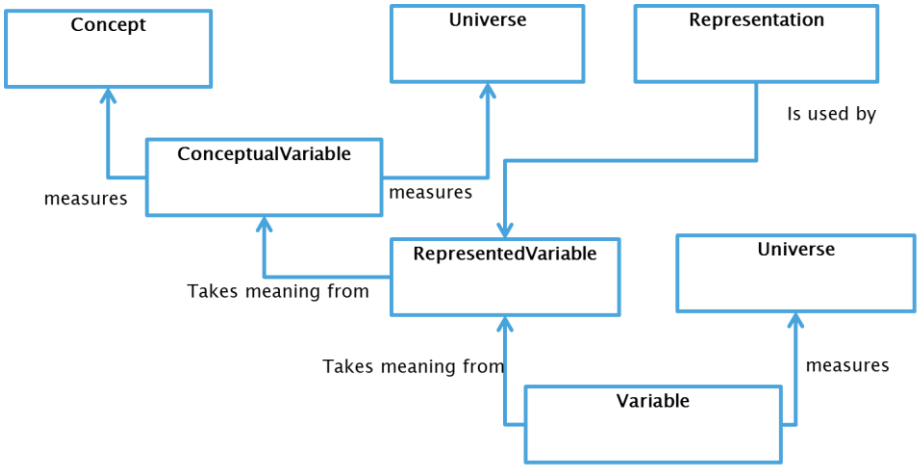


Figure 5.3 DD1 Variables aligned with GSIM

It should be noted 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.* 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 two simple cases will be used. 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.

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*

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.

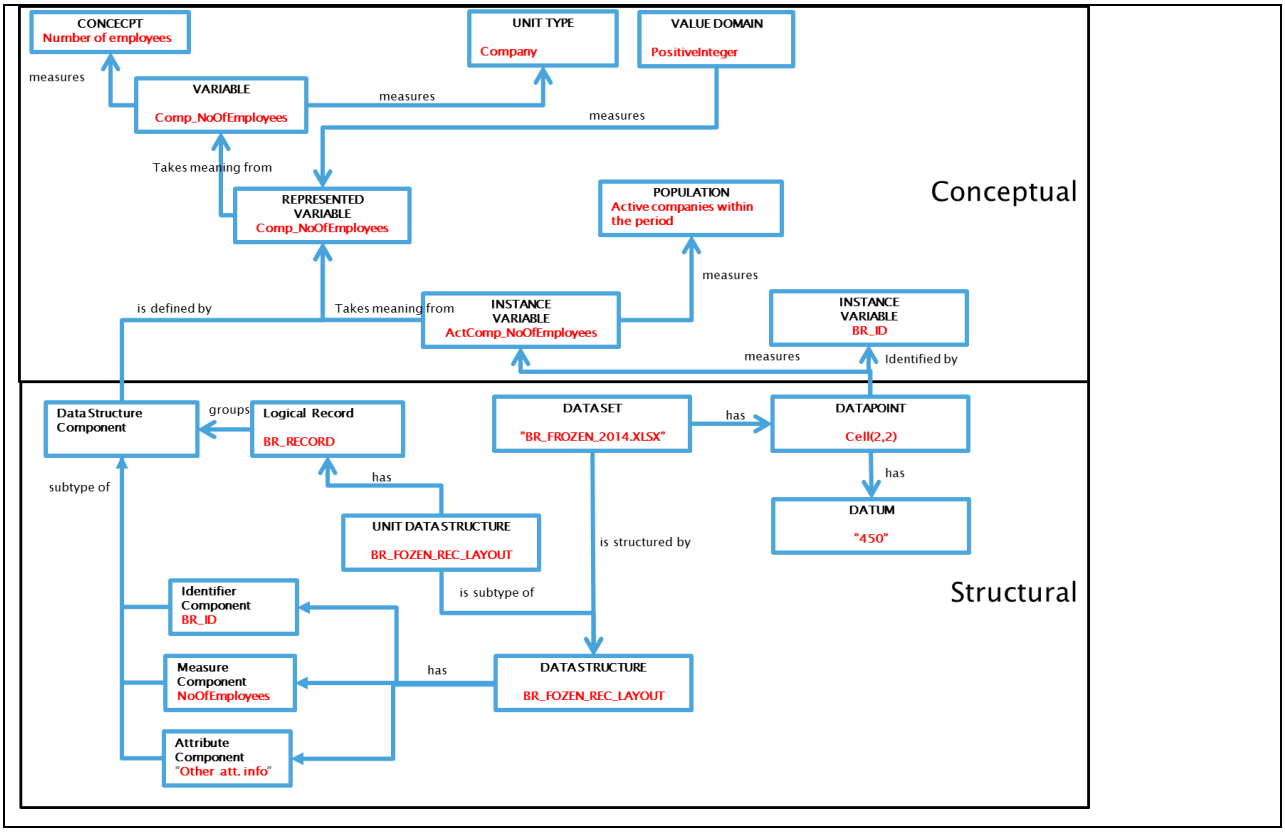


Figure 5.5 Conceptual model showing GISM concepts and example data from the Business Register

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

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

Fig 5.6 *SC_2014.XLSX*

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.

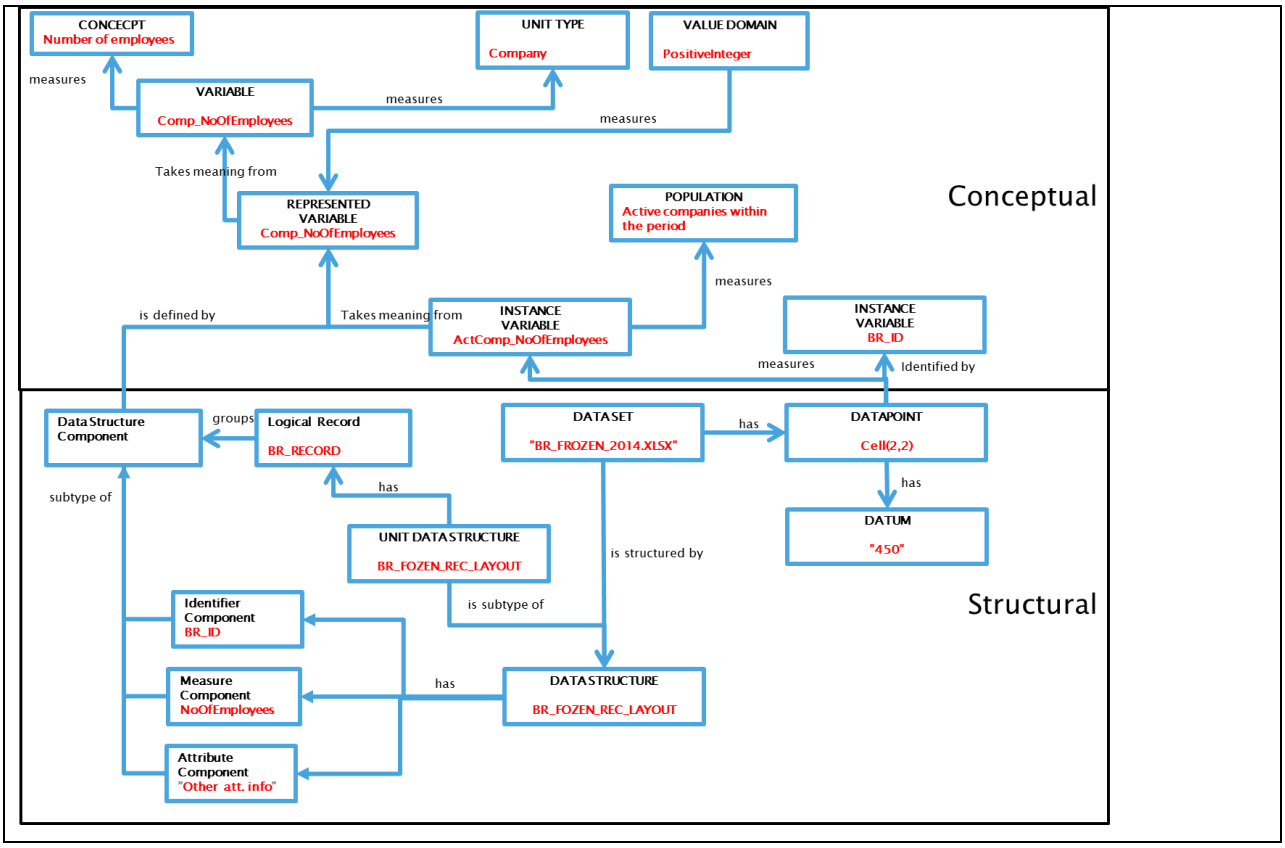


Figure 5.6 Conceptual model showing GISM concepts and example data about population

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 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.

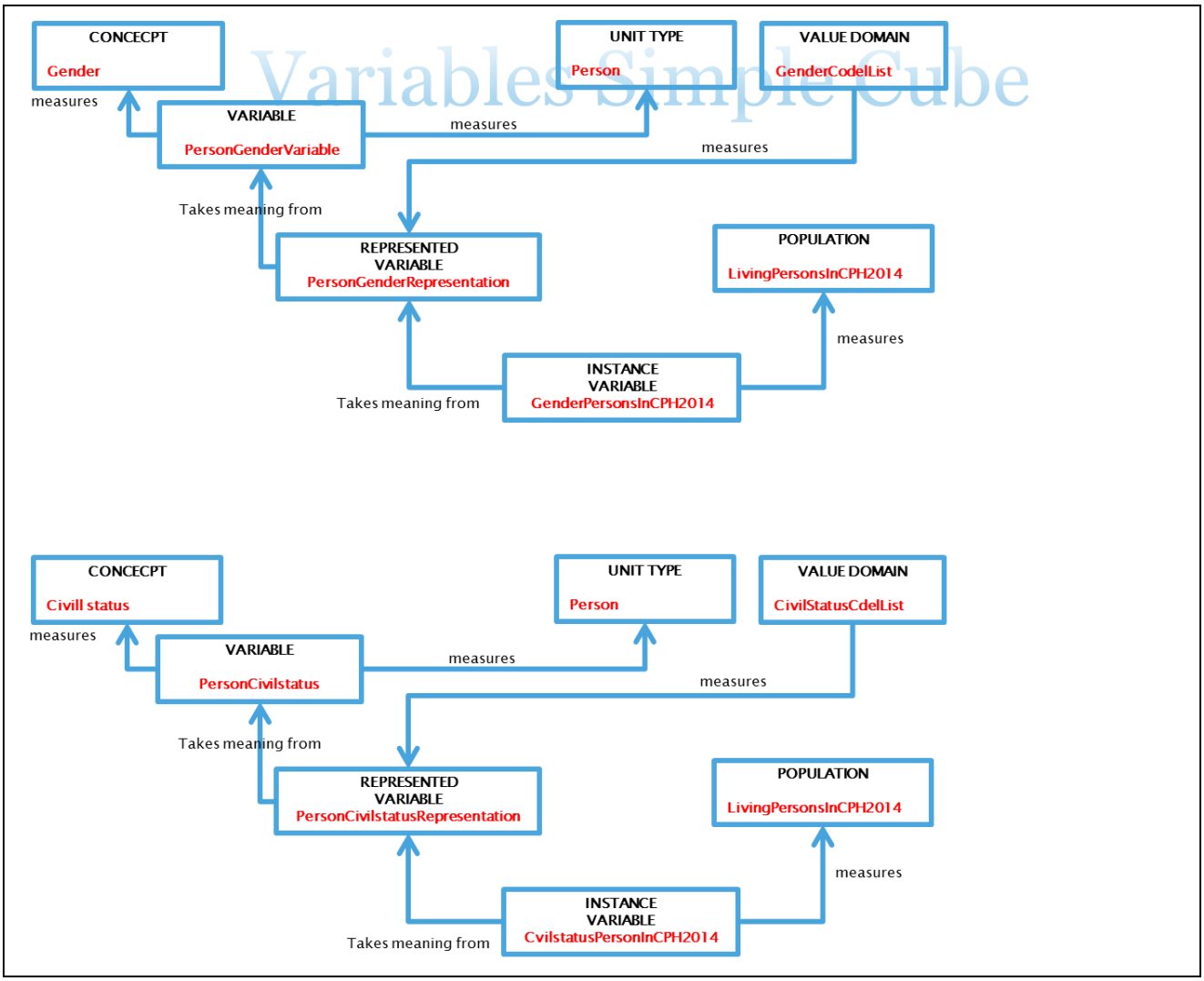


Figure 5.7 Gender and Civil status variables

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.

5.3 Implementation of DDI and Colectica

At the beginning 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 in the coming years.

The model below shows part of the physical DDI compliant model, we are working with now in Statistics Denmark.

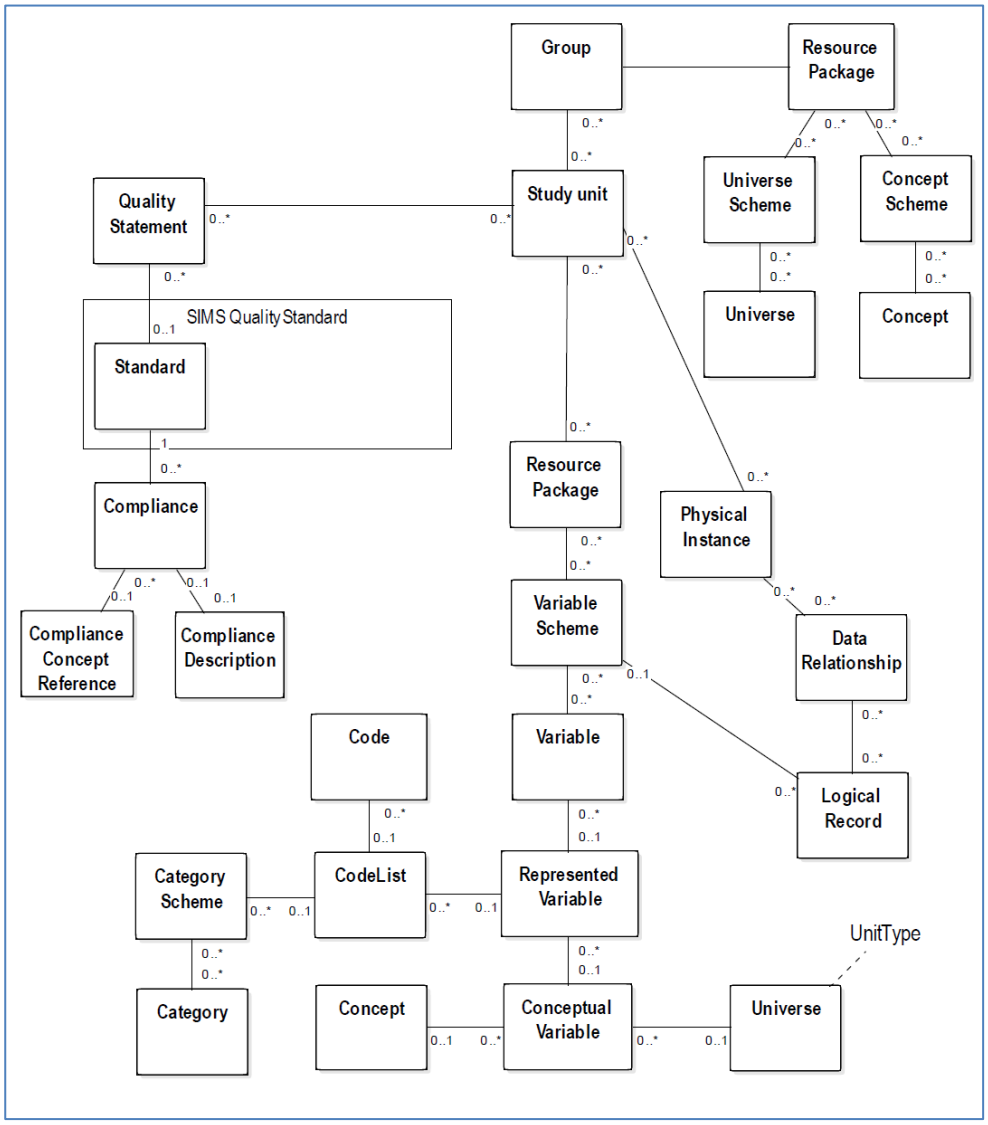


Figure 5.8 Logical model complying with DDI 3.2

At the top, we are using a group construction in DDI as a wrapper for all statistics and registers. We use the DDI term study unit for statistics and registers. To the top left, we have the model for quality information. To the right and at the bottom we have elements covering the description of variables, dataset, concepts and code-lists etc.

In addition to the work mentioned above, the Nordic countries have all been involved in the work on developing a common model for classifications. In 2013, the cooperation had representatives from Denmark (Statistics Denmark and the Danish Data Archive) and Statistics Sweden met in Copenhagen and performed a gap analysis to determine whether the DDI Lifecycle standard could be used to implement the GSIM Statistical Classifications Model. The result of work is a paper with the title Copenhagen mapping [13]. The paper describes how the GSIM Statistical Classifications Model maps to DDI 3.2, and offers a set of controlled vocabularies to be used by DDI 3.2 implementers who wish to describe classifications using the standard. These elements are now part of the DDI standard and Statistics Denmark uses these in the implementation of classifications in Colectica.

5.4 Moving towards a common Nordic Metadata Model

Representatives from all Nordic countries are involved in detailed discussions of the models described in the previous section. This takes place as part of the Nordforsk project financed by the Nordic Council of Ministers. All countries will implement the model in different physical ways, but the big news is that all countries, including staff from the Swedish Research Council agree on a common understanding and

common models at conceptual and logical level. Because of this, we will be able to use common terms and common understanding when metadata specialists and non-metadata specialists talk about classifications, concepts, variables, code-lists etc.

This will be of great benefit both for the continued development of the model, but especially beneficial for the work on statistics internally at NSI's and in the cooperation between NSI.

6. Putting the pieces together in the on-going project

In this section, the aim is to put the pieces together and show how we proceed in practical life, using the insights explained in the preceding sections.

As part of a new EU-grant, we have a work package on investigating standards and aligning these with the new *ESS Vision 2020. Building the future of European statistics* [18]. We are testing the standards and models on a pilot basis as part of the work in the EU-grant.

The overall business goals for the project are to support the modernisation and integration of work at EU and national level using GSBPM, GSIM, SDMX and DDI. The specific goals are a) to improve and standardise work-processes using GSBPM, b) to improve the metadata system through the use of GSBPM , GSIM , DDI and SDMX and other related standards, and c) to improve exchange of statistical documentation with EU.

We use the Business Process Management model and TOGAF as described in chapter 4. Following the model at figure 4.5, we are working on a prioritised project on redesigning processes with regard to metadata. This implies that we start with business goals and business principles and hereafter we continue with descriptions of architectural views on business processes, data, applications and technology.

With regard to business principles, we use the European Statistics Code of Practice [16] with details from the Quality Assurance Framework of the European Statistical System [17]. The principles give us overall direction for the work. The following principles and indicators in the Code of Practice and Quality Assurance Framework are relevant for the on-going work:

Principle 7. Sound methodology, Indicator 7.2:
Procedures are in place to ensure that standard concepts, definitions and classifications are consistently applied throughout the statistical authority

Principle 10 Cost effectiveness, Indicator 10.4:
Statistical authorities promote and implement standardised solutions that increase effectiveness and efficiency.

Principle 15. Accessibility and Clarity, Indicator 15.5.
Metadata are documented according to standardised metadata systems

With business principles and business goals in place, we have set up a baseline and target architecture for business processes, data, applications and technology. Through a gap-analysis of baseline architecture and target architecture, we set up scenarios and suggest a way forward on the implementation of a solution that fulfils principles and goals. When we design the target architecture for business processes, data, applications and technology we also use the insight described in the chapters above. This includes how to get a) improvement and precision in the terminology used when we talk about metadata, b) better understanding of the role of metadata in production of statistics, and lastly c) better understanding of the role of metadata in relation to our users.

Overview of models and standards in the on-going project.

We have divided the project into the following sub-projects.

1. Preparation of documents on use of models and standards based on input from *ESS Vision 2020. Building the future of European statistics* and other sources. These documents will be discussed and finalised at workshops with internal and external users and experts.

2. Investigation of user needs and integration of metadata in dissemination
3. Migration and harmonisation of concepts, variables and classifications
4. Improving the transmission of quality reports to the EU
5. Integration of metadata as defined by GSIM in GSBPM workflows

In general, we are using the business process management and architecture methodology described in section 4. This results in overall documents with goals, principles and architecture elements. These documents provide the overall architectural models for business processes, applications, data and technology and they must serve as a guide for the subsequent work at the sub-projects.

In sub-project 1, we touch upon all models and standards.

In sub-project 2, we use the insight on role of metadata in relation to users as described earlier in the paper. We do this by using focus groups where we discuss various ways of the dissemination of metadata.

In sub-project 3 on concepts, variables and classifications, we use the insights described in section 5 on transforming GSIM into DDI.

In sub-project 4, we improve the integration of metadata-systems in Statistics Denmark using SIMS, ESMS and ESQRS.

In sub-project 5, we use the models and thinking presented in section 4.1 on how to integrate metadata into business processes. GSBPM, GSIM, DDI play important roles in this work.

7. Conclusions

In the beginning of the paper it was claimed that the road towards common metadata in a statistical context requires a) improvement and precision in the terminology used when we talk about metadata, b) better understanding of the role of metadata in production of statistics, c) better understanding of the role of metadata in relation to our users and lastly d) the use of common overall models for business process change and Enterprise Architecture to ensure successful implementation.

As shown in this paper, there is no simple road towards implementation of metadata in a statistical context. If we want to succeed, it is necessary to combine a lot of theoretical and practical knowledge and standards. But the most important fact is that benefits can only be accomplished by using standards and working together across organisational and national state boundaries.

This has been the case historically, especially shown by contributions from Bo Sundgren, whose approach to the general introduction to metadata is very fruitful.

As mentioned earlier in the paper representatives from all Nordic countries are involved in detailed discussions aiming at developing detailed models and a common understanding of the international model for statistical information objects (GSIM). All countries will implement metadata in various ways, but the big news is that all countries, including staff from the Swedish Research Council, agree on a common understanding and common models at conceptual and logical level. Because of this, we will be able to use common terms and common understanding when metadata specialists and non-metadata specialists talk about classifications, concepts, variables, code-lists, etc.

This work will be of benefit for the continued development of the models, but the main benefit comes from using the models to improve business-processes and to improve the information about the statistics for internal and external users.

References

[1] Thygesen, Lars and Nielsen, Mogens Grosen (2013). *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. (2004). *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 Organisational 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] Nielsen, Mogens Grosen Nielsen and Flemming Dannevang (2015). *Towards common metadata using GSIM and DDI 3.2*, Paper presented at the European DDI Conference December 2015.

[13] 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

[14] Eurostat, *ESS EA Reference Framework* (2015). Link: http://ec.europa.eu/eurostat/cros/system/files/ESS_Reference_architecture_v1.0_29.09.2015.pdf_en

[15] United Nations Economic Commission for Europe UNECE (2013). *Generic Statistical Information Model (GSIM): Statistical Classifications Model*, Geneva. Link: http://www1.unece.org/stat/platform/display/gsim/Statistical+Classification+Model

[16] Eurostat (2011), *European Statistics Code of Practice*. Link: http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

[17] Eurostat (2011), *Quality Assurance Framework of the European Statistical System.* Link: http://ec.europa.eu/eurostat/web/quality

[18] Eurostat (2014), *ESS Vision 2020. Building the future of European statistics*. Link: http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020

[19] The Open Group (2011), *The Open Group Architecture Framework TOGAF version 9.1*. Link: http://www.opengroup.org/subjectareas/enterprise/togaf/

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

[21] 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)