**Concepts**

*- making the complex simple*

Sanne Nielsen, Statistics Denmark, svn@dst.dk

**Abstract**

*For some time, Statistics Denmark has been working on a unified metadata container for the following metadata elements: classifications, documentation, variables and concepts. In 2018, a working group established a plan for the implementation of concepts which will be implemented in 2019. The Metadata container is based on the Colectica software supporting relevant international standards (GSIM and DDI).*

*With this project Statistics Denmark will be one step closer to delivering standardized, high quality documentation on concepts in statistics, with the possibility to disseminate documentation on all our products ; a feature the users demands.*

*A concept is defined in GSIM as a unit of thought differentiated by characteristics, and the difference between a concept and a variable has been one of the subjects addressed by the working group.*

*The focus is to have a one-place-only and once-only approach to concepts, and to create an organization for Quality Assurance that is parallel to the QA process on our other metadata products.*

*The main challenges are among others scarce resources, as well as very different approaches to concepts within the organization (some what due to lack of instructions and procedures), and a very academic tradition in the discussion of metadata.*

*As a result of these challenges the strategy has been to create a broad process across the organization, with focus on what we can do, that will give us the largest effect. This has given us a very pragmatic and energetic process that will give us results with a small investment.*

*The project runs through 2019, and when it is completed, Statistics Denmark will have an application for concepts, where all defined concepts are contained in a single container, and an organizational structure with the necessary skills. Information from the container can be integrated into our publications channels via an API providing users access to all relevant information about our data.*

**Keywords:** *Concepts, metadata,**documentation, user needs*

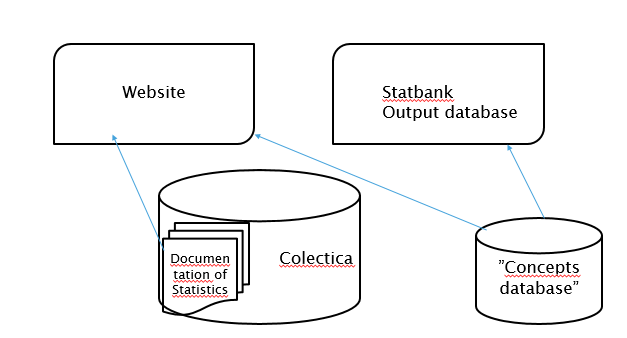
# How we got to where we are

At Statistics Denmark we have been working systematically with metadata for many years. The first work on concepts grew from user needs. Every user survey we made, told us that there was a need for definitions on what we measure in our statistics. We had the will and the ideas, and we did make progress, but we also experienced adversity.

The first Statistics Denmark work on concepts was a dissemination assignment entailing the gathering of concepts already described, and disseminating them on our aggregated output database, StatBank, and our website. Therefore, the storage of concepts data was connected to these two platforms. We knew, there was a theoretical discussion on what constitutes as a concept, that we needed to take into account, but at that time, we did not have the competences to address this. Therefore, our work was centered on the dissemination part in StatBank and website, and on describing the content of these two platforms. We did not focus on separating concepts from variables, or connecting concepts to other types of metadata.

A few years later, my colleagues in Methodology implemented Statistics Denmark Documentation of Statistics in the Single Integrated Metadata Structure (SIMS). SIMS contains concepts, but there was no real connection between the SIMS concepts and the concepts gathered on StatBank.

The result of this architecture is illustrated in figure 1. Two separate storages of concepts, which feeds information to the dissemination platforms.



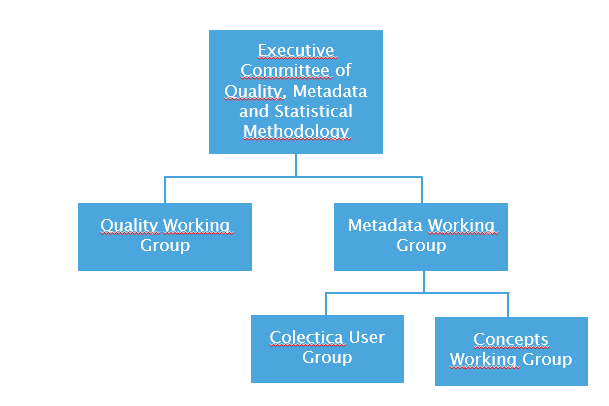
*Illustration 1: Storage of concepts as-is*

The consequences of this architecture was:

1. Redundant data between the two databases. This gave us more than one definition pr. concept. The definition varied from the same content with slightly different use of words, to cases were there was actually more than one concept.
2. Uneven quality of the concepts. Since we had no quality description on concepts, we had a large variety in quality, which gave an uneven impression on our dissemination of concepts.
3. Uneven coverage between subject areas. Since we had no definition on what constitutes as a concept, we had many interpretations across the subject areas.

# How we leap to the next level, and what we hope to achieve

So with redundant data and uneven coverage and quality, it was painfully clear that we needed a plan, so in 2018 a dedicated working group set out to create just that. The working group referred to our Metadata Working Group, which again refers to the Executive Committee on Quality, Metadata and Statistical Methodology



*Illustration 2: Organization of the metadata work at Statistics Denmark*

The purpose of the working group was to organize a plan for improved handling of concepts in Statistics Denmark, with the aim on one-place-only and once-only data-entry and quality measures for concepts.

The working group had 11 active members; four from IT, three from Dissemination, two from Methodology and Analysis and two from statistical production. It was an important point, that the working group was open to everyone with an interest in concepts. The working group held nine meetings in 2018 and submitted the plan in October 2018. The plan was accepted by the Executive Committee of Quality, Metadata and Statistical Methodology, although we were only given very few resources to implement it.

So what was different in this plan, from what we have done earlier?

1. The working group had defined what a concept is!
2. The plan only targeted concepts, not every other documentation need users might have.
3. On the system level the plan included already implemented software
4. On the organizational level the plan used the experience og the organization, that Statistics Denmark had already used in the work on implementing SIMS.
5. Also the plan included the question of needed competences

This paper will have a short description on each of these points.

With this plan, we aim to achieve standardized, high quality concepts to help our users read and understand our products, one-place-only and once-only storage, which allows us to reuse concepts across all subject areas, new dissemination possibilities and minimizing maintenance (both metadata and system).

# What is a concept?

It turns out this simple question can trigger a very long discussion. The answer is very well established in scientific practice, but how will we establish it at Statistics Denmark?

The international standards are more or less agreeing on what a concept is, and how it is defined. Generic Statistical Information Model (GSIM) defines it as **a unit of thought differentiated by characteristics[[1]](#footnote-1)** and the international standard (ISO) has is as **unit of knowledge created by a unique combination of characteristics**.[[2]](#footnote-2)

Our experience was we needed to have descriptions of concepts more specific and closer to the everyday at Statistics Denmark. Also there is a need to make an operational separation of concepts from variables. Therefore, the working group formed an operational definition described as:[[3]](#footnote-3)

1. A concept describes the subject or the reality that we want to describe through our statistics in abstract terms. The concept is used to understand the reality.
2. A variable is a product of statistics, and is used to understand a specific statistics.

|  |  |
| --- | --- |
| **Concept** | **Variable** |
| Philosophical  Describes reality  Idea, ideally  Orthography:  Often written in lowercase | Technical  Describes elements in statistics  Specific  Orthography:  Often written in uppercase,  shortened |
| Example from DST:  Civilstand (Marital status) | Example from DST:  CIVST |

*Illustration 3: Concepts and variables*

This illustration on separation of variables and concepts will be supplemented by relevant examples of concepts central to a statistics, and the variables from the same statistics, making the distinction specific.

# Concepts, and only concepts

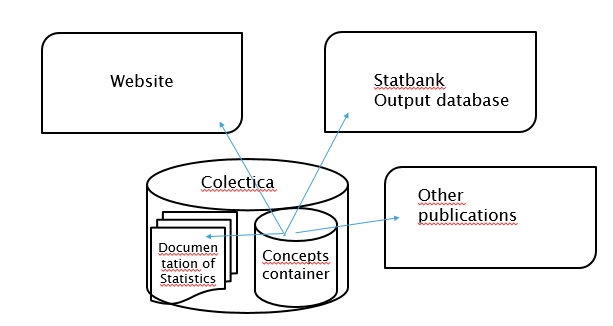
With a clear definition at hand, the working group could now start analysing how many concepts we would expect to include. To make the job manageable the focus was firstly to handle the concepts DST already had described. The concepts already existing are stored in two different databases, presented through three separate user interfaces (shown in illustration 1). The “concepts database” which feeds data to Statbank, contains 2640 concepts. But that number had to be revised with our new separation of variable and concepts. When we implemented the “concepts database”, the idea was to describe all the structural metadata a Statbank table can contain. But with our “new” definition of concepts as little as 10% of the content in “concepts database” is indeed concepts. The rest is most likely variables. From Documentation of Statistics we have 1900 concepts, but the existing data model allows unique concepts within each of 400 documentations. This means that we can expect some overlap, so that the most popular concepts exists more than once.

By scoping the plan to include only concepts, not variables, it is no longer a question of describing more than 5000 concepts, but more like 500-1000. This project still recognizes the need for high quality descriptions of variables, but that would be another project. Focusing the resources on concepts this year will not solve all user needs, but it will solve some. It is eating an elephant one bite at a time, sizing the bite to fit the resources at hand.

# One system – one-place-only

One of the big advantages the working group had in this work, is that the Colectica system for administration of metadata already was in operation. Colectica was implemented at Statistics Denmark when the SIMS standard was implemented. Later classifications was implemented in Colectica. So we have an exciting system at hand that “just” needs to be expanded with new needs.

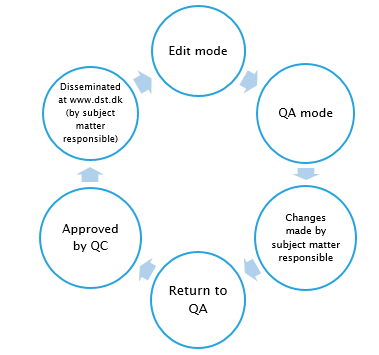
The vision of metadata in Statistics Denmark is to have four types of metadata elements (variable, concepts, documentation and classifications) all in one system, all linked to each other. A single unified system gives us a number of advantages: reuse of metadata, minimizing maintenance (on both metadata and system) and the possibility to link metadata elements together.



*Illustration 4: Storage to-be*

# Reusing organization

One of the large advantages we have in this project is that we have a Quality Assurance (QA) organization we can reuse. The QA organization was created, when the SIMS standard was implemented. Methodology developed a QA cycle and put it to use on the SIMS standard.



*Illustration 5: Quality Assurance cycle for Documentation of Statistics*

This cycle includes roles for a Quality Coordinator and a Subject Matter Responsible, and contains the flow that a Documentation of Statistics undertakes when changes are made (or a new Documentation is created) through the QA process, to when it is approved and disseminated.

The flow is handled by a workflow in our metadata system Coletica. With the roles and the flow already defined, is was easy to reuse this on the concepts. At the present time the final organization on concepts is not finished, but we do know that the QA flow will be somewhat similar to the one of Documentation of Statistics for two reasons. First, we want to keep the organization simple, and nothing is simpler than what you already know and use. Second, the QA process on Documentation of Statistics has shown to our organization how important QA is to the quality of a product.

# Competences

In November 2016 Statistics Denmark held a workshop on terminology with the participation of leading competences on the subject[[4]](#footnote-4). The topics were (among others) constructing concept hierarchies and how to develop concept definitions. The workshop was really an eye-opener for us. There was so much information presented to us which we could easily use operationally, including many thoughts on what constitutes as a concept, and how to describe them.

The participants on the workshop broadly represented Statistics Denmark, and many of them later participated in the Concepts Working Group or the Metadata Working Group. So the competences and knowledge gained through the workshop is working within the project. Also much of the information gained at the workshop is now part of the introduction material the project is developing for the future operational organization. This contains very specific guidelines on how to describe concepts as on illustration 5*.* Each of the rules on this illustration are elaborated in the guidelines.

1. Concepts definitions must be short and simple
2. Mind the target group
3. There must be no unknown concepts
4. Concepts must be mutually reconciled
5. Concept definitions must adequate; not to broad, not to narrow.
6. Concept definitions must not be circular.
7. Concept definitions describes what the concept is, not what it is not.

*Illustration 5: Example from the guidelines developed from the workshop*

It might seem novel, but for us it was a step of great importance to invite knowledge from terminologists into the organization. It made quite a difference to have specialized input on concepts in a statistical context, from someone who was not in statistics.

The work now lies in disseminating this knowledge out in the organization. The plan is to build our guidelines as specific and to the point as possible, while using specific examples from our everyday work.

# Conclusions

We believe this project will leap us closer to our vision of coherent metadata. First, the scope of the projects. This contains concepts, and only concepts. Variables or other documentation needs is not included in this project. We have scoped one thing, and we have been working on fitting the assignment to the resources at hand. This has made quite a difference in the project, that the aim no longer is to describe “everything our users could possible want”, but only to describe central concepts for each statistics.

Second, we have added to our competences on terminology and concepts. It has been of great value to have terminology experts help us interpret GSIM and ISO.

Third, in this project we have had the advantage of already having a metadata handling system in operation. The system already contained documentation (SIMS) and classifications, and we have been locked on the options and limitations build into Colectica. In terms of creating progress in the project it has been a large advantage. The decision on aiming for coherent metadata in one system was already taken. So this project was not a project of choosing or developing a metadata system, which earlier projects had been to some extent. It was a project, where the developing was adapting an already existing system to specific needs.

Fourth and lastly, reusing the QA organization from SIMS, has made the QA process easy to accept. The idea of QA is not new and the process will be similar to the one on Documentation of Statistics. Concepts are using already implemented processes as stepping stones instead of inventing a separate QA.

In short my advice to anyone who are planning a similar project will be to scope your projects according to your resources, get terminology experts to help you straighten out the basics, and reuse as much as possible on both organization and system.

# **References**

International Organization for Standardization. (2009). Terminology work – Vocabulary; Part 1: Theory and application (ISO Standard No. 1087-1:2000(E/F))

International Organization for Standardization. (2000). Terminology work – principles and methods (ISO Standard No. 704:2009(E)).

Nistrup Madsen, Bodil. (1999), Terminologi – principper og metoder, Gads forlag.

UNECE (2015): Generic Statistical Information Model (GSIM): Statistical Classifications Model. Meeting of the Expert Group on International Statistical Classifications New York, 19-22 May 2015 (42 pp).

1. GSIM version 1.1 and version 1.5 [↑](#footnote-ref-1)
2. ISO 1087-1:2000 [↑](#footnote-ref-2)
3. With the inspiration from professor Bo Sundgren, who has done acknowledged work in information science and statistical methodology, [↑](#footnote-ref-3)
4. Professor Bodil Nistrup Madsen and associate professor Hanne Erdman Thomsen. Copenhagen Business School [↑](#footnote-ref-4)