UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE Working paper
CONFERENCE OF EUROPEAN STATISTICIANS 27 July 2016

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

The Register Utilisation Tool: A Practical implementation of GSIM as support in register-based research

Magnus Eriksson, The Swedish Research Council, magnus.eriksson@vr.se

Abstract:
As a step in fulfilling a government commission to support register based research the Swedish Research Council decided to create a GSIM (Conceptual group) implementation. The choice to use GSIM was preceded by requirements work and an evaluation of several metadata frameworks/standards.

The requirements work was conducted with researchers from different fields and different levels of experience in register based research. Because of the legal and ethical constraints that apply in the business domain the separation of metadata and data were a core business requirement coming into the project. The main effect goal was shortening the researchers "time-to-data".

The core requirements gathered was that the application should: 1) Enable variable search by meaning/concepts, 2) Provide the metadata needed to support the researcher during evaluation of a variable in relation to the research question and 3) during harmonization efforts, 4) Create the preconditions for expressing and communicating the design/selection of variables in an unambiguous way and 5) provide support during analysis of a variables quality & sources, collection methods etc.

The first four core requirements were included in the development cycles that resulted in the release of a test version to be evaluated by the researchers. The requirements regarding referential metadata are to be included in a later stage.

The selection of GSIM was based on the frameworks: 1) Separation of meaning and representation, 2) Strong support for handling codelists and classifications, 3) Domain independence/generic qualities, 4) The frameworks strong support in the international community.

The implementation resulted in an application named "The Register Utilisation Tool (RUT)". RUT provides an infrastructure to support the researcher during conceptual search, evaluation of variables in relation to the research question, variable harmonization and communication with the register holder.

After implementing the application we concluded that the use of GSIM as a common framework both enables researchers during search, evaluation and design and enhances metadata maintainability. The perspectives and level of granularity that the researcher is able to provide in order to communicate the selected variable in an unambiguous way are also more versatile supported by GSIM.

1 Introduction

In 2014 the Swedish Research Council received a government commission to create a business function and an infrastructure to support register based research. The commission was divided into three sub-projects where provision of information, advisory and education functions for register-based research was one.

This sub-project intends to support the register based research community by providing an information portal and a metadata search and analysis tool. The portal contains general information about rules, regulations and activities related to register based research in order to support the researcher throughout the research project.

The portal also acts as entry point to the Register Utilisation Tool (RUT) which is a variable search, register discovery and selection design tool. RUT provides an infrastructure to support the researcher with conceptual search, evaluation of variables in relation to the research question, variable harmonization and unambiguous communication with the register holder.

Before the development project was set up an evaluation of several different metadata frameworks and standards took place and requirements were gathered from a reference group with researchers from different fields of register based research. The evaluation together with the core business requirements lead to the decision to use a selection of GSIM, mainly the conceptual part, as the core information model for the solution.

2 Business requirements

Because of the legal and ethical constraints that apply in the business domain the separation of metadata and data were a core business requirement coming into the project. The next prerequisite was one of the projects main effect goals, to shorten the researchers "time to data", that is the time from when the researchers start the work identifying and selecting data to support their research question until they have access to the relevant data.

These prerequisites set the starting point for the requirements work that took place in collaboration with a reference group made up of researchers active in the field of register based research. The researchers in the group came from different research fields within the social sciences and medical sciences and have different levels of experience in register based research.

The core requirements set in collaboration with the reference group was that the project should:

1. Provide functions for variable search that do not require knowledge of register owners, registers or variable names and a way to search by meaning/concepts

2. Provide metadata to support the researcher during evaluation of a variable in relation to the research question.

3. Not put resources into harmonizing variables in general since the variables harmonization potential are study specific and can only be decided by the researchers. Instead the emphasis should be on providing the metadata needed to support the researchers during harmonization analysis. That is, each register holders definitions should be presented for evaluation instead of making the register holders adapt to a common vocabulary (see figure 1 below).

4. Give the researcher easy access to metadata on variable meaning, representation and populations in order to provide support for communication of design in a clear and distinct way. Including changes in meaning & representation over time.

5. Provide support during analysis of a variables quality & sources, collection methods etc.

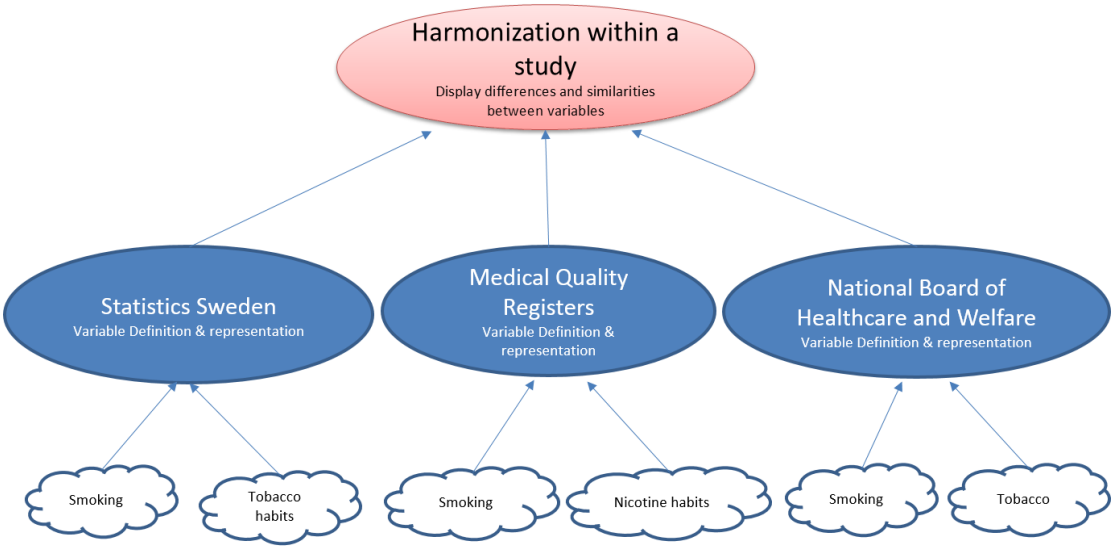

Fig 1 Presenting variable metadata for evaluation and harmonization within study instead of harmonizing variables between registers.

2.1 Perspectives

During the requirements work there where 4 perspectives on a variable that needed metadata in order to provide the researcher with the needed support during evaluation of a variable in relation to the research question and during the harmonization process. The project decided to include three of the perspectives in the first delivery. Detailed metadata on these three perspectives are also essential for communicating the design to the register holder in a clear and distinct way.

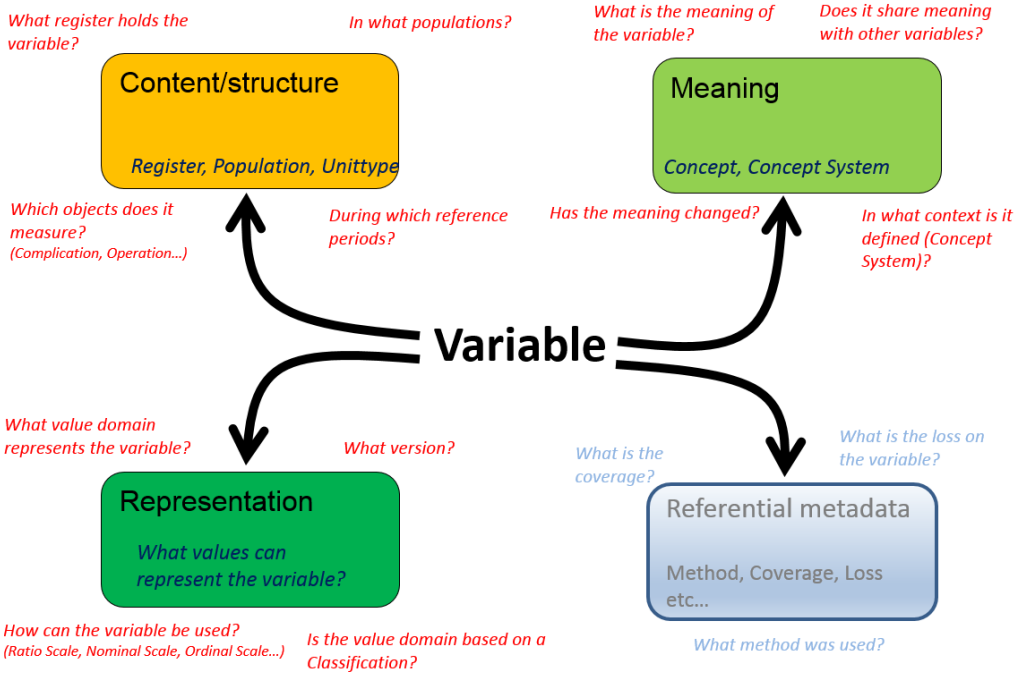

Fig 2 Illustration of perspectives. The fourth perspective regards referential metadata broken down to the variable level such as method, loss, coverage etc. and will be included in a later phase.

3 Approach

3.1 Evaluation

After an evaluation of different frameworks and standards in relation to the above mentioned prerequisites and requirements the choice fell on using GSIM as a metadata framework on the conceptual level and to create a logical and physical model based on GSIM with some minor additions to support the project requirements.

The choice of GSIM was based on the frameworks:

1. **Principle of separating meaning and representation**. This provides a foundation for implementing search by variable meaning and, even more important, to provide the conceptual support that the researchers need to evaluate the variable meaning in relation to the research question and a defined study variable during harmonization efforts.
2. **Strong support for handling codelists and classifications** with information objects covering the different aspects of representation and its historic changes. This part of the framework are very important during the evaluation of the variables harmonization potential.
3. **Domain independence**. From the projects point of view the choice of metadata framework/standard also needed to be influenced by how generic the framework/standard was since the register holders that provide the metadata come from a wide variety of business domains such as Statistics Sweden, The National board of health and welfare, Biobanks, Medical Quality registers and Cohorts etc. The framework also needed to provide a common language regarding metadata from these domains and be relatively easy to communicate to the register holders.
4. **Strong support in the international community.** The management/governance of GSIM by the UNECE High-Level Group for the Modernization of Official Statistics give the framework a strong support in the international community.

3.2 Selection

The researcher emphasis on meaning and representation resulted in a selection composed of mainly information objects from the GSIM Concepts Group with some addition from the Business Group.

3.2.1 Business Group Selection – objects regarding Register, Variants and Change.

The researchers in the reference group expressed a need for a high level understanding of the structured list of objects that constitutes the register within which the variable is stored and collected. They also showed a significant interest in the reasons for changes within a register.

In order to meet these needs we decided to implement information objects regarding the purpose, goal and design of the register that holds the variable. When approaching the registers we soon realized that the register most often served as an umbrella term for one or more variants that actually hold the population data.

We then selected “Statistical Program” as the object to hold purpose and goal for both the register and the variant and added the “Statistical Program Design” for information regarding the method and design for collection of the register data.

Finally we needed an information object to meet the researchers need to understand the reasons behind changes that entails additions of variables in a register/variant. We opted to include the

information objects describing “Statistical Need”, “Business Case” and “Change Definition” for this purpose.

In order to illustrate the register-variant hierarchy we added a recursive composition relationship on “Statistical Program”.

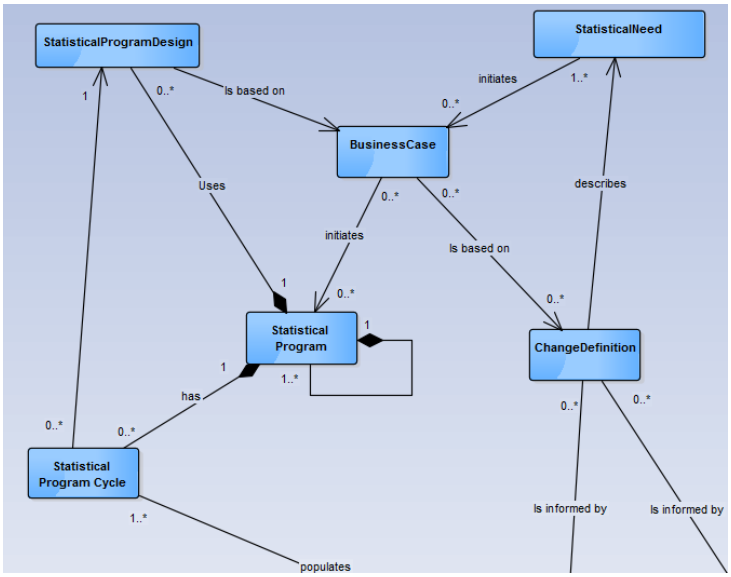

Fig 3 Selection - objects regarding register, variants and change

We also included an association between “Statistical Program Cycle” and “Population” to complement the relation between “Change Definition” and “Population” in order to express our usage of the Information Objects more clearly.

3.2.2 Conceptual Group - objects regarding Concepts, Concept Systems and meaning.

In order to give the researcher access to the meaning of the variables in an efficient way we included “Concept” and “Concept System” into the solution since they provide the infrastructure to express meaning separate from the concept specializations and the representation.

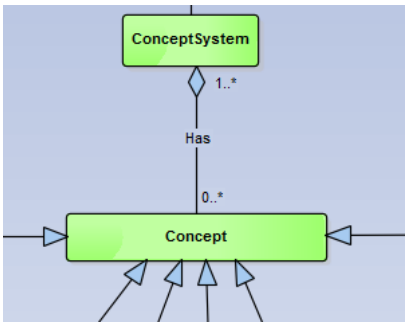

Fig 4 Providing the researcher with access to the meaning of the variables

Since we set the definition in a **Concept** separate from the Variable, Unit Type, Population and Category we get a way of handling the fact that the term names for the Variables and Unit Types can differ over time and between registers and variants although the meaning, that is the Concept definition, is the same. The other way around we also manage the cases when the term names are the same but the meaning differ.

Concept System – Since our requirements state that register holders definitions should be presented to the researchers for evaluation instead of making the register holders adapt to a common vocabulary we give each register holder the responsibility for their own definitions. Of course we give the advice that, when possible, common vocabularies should be referenced instead of creating new definitions but the choice of definition lies on the register holder.

The use of Concept Systems as a way of grouping Concepts and concept relations gives the register holder an opportunity to present different perspectives of its Concepts, in our case often depending on what register variant the researcher is interested in, and is thus very useful.

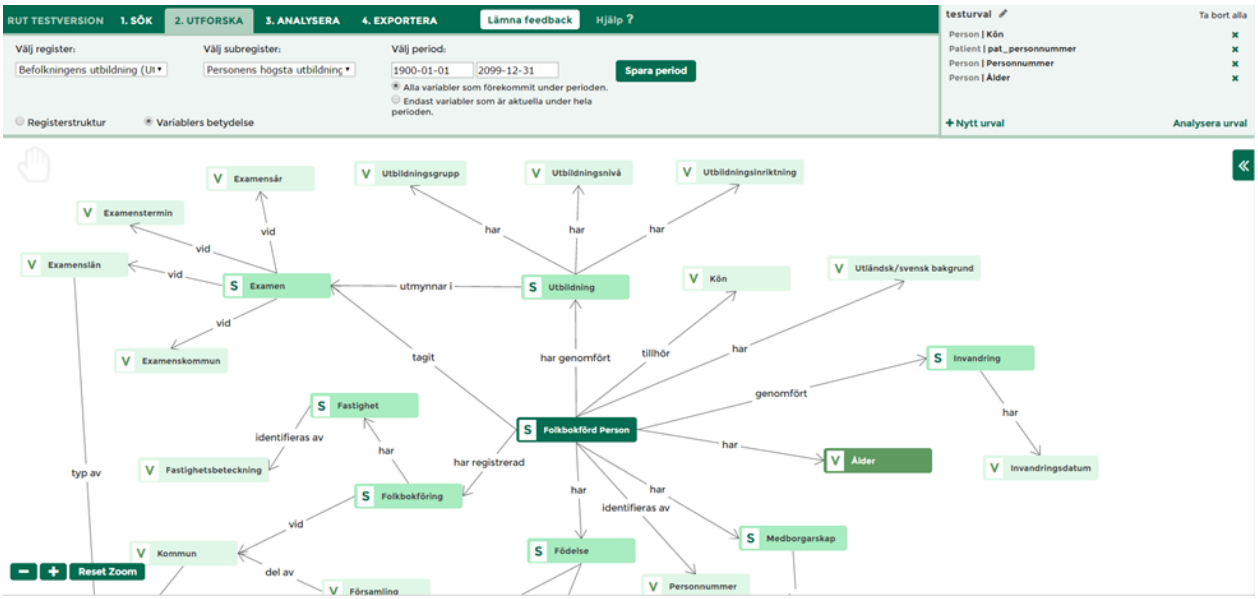

Fig 5 Presenting a Concept System in the application

3.2.3 Concept specializations - Population, Unit Type, Variable

Population. The population constrained by time and geography is of course of great importance for the researchers and is included in our GSIM selection as an specification of the UnitType(-s).

Variable in detail. In order to give easier access to, and a better overview of, the Variables in a register we wanted to provide more detail to the GSIM variable to be able to visualize the variable as two parts, the "Variable UnitType" and the "Variable Concept", e.g. Father [Variable UnitType] + Income [Variable Concept].

We then get the opportunity to provide a visualization of the logical grouping of Variable Concepts within Variable UnitTypes. Through the Variable we have the relation to the Unit Type e.g. Person[UnitType]->”Father Income”.

This gives the researcher a better overview of the register variables then what we would be able to offer if we were displaying them in a long list under the Unit Type. In addition to this it also provides a way to handle reuse. By using these more detailed information objects as separate specializations of concept we are also able to handle differences in naming of variables that have different names but the same meaning in an even better way.

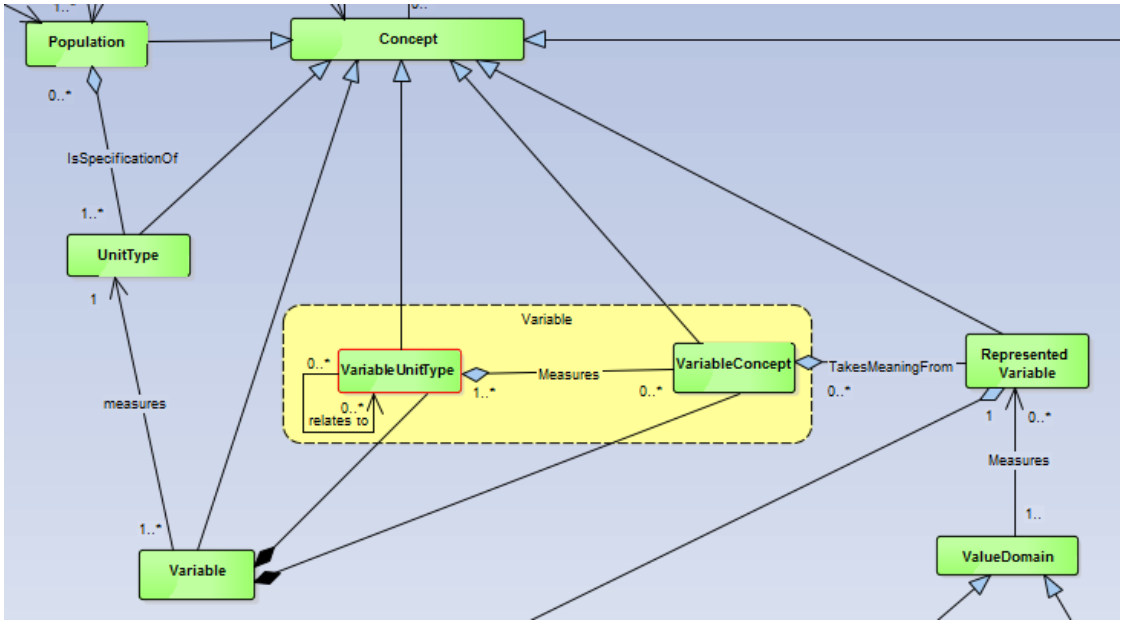

Fig 6 Variable in detail

See example below from a population register – Statistical Program for family:

Population	UnitType	Variable	Related to Concepts
Registred individuals in Sweden in ages between 18-64 years 2010	Person	Persons Country of birth	Registered Person, Country of birth
	Person	Father Country of birth	Biological Father & Country of birth
	Person	Mother Country of birth	Biological Mother & Country of birth

Fig 7 Before

Population	UnitType	Variable	Domain UnitType	Variable Concept	Related to Concepts
Registred individuals in Sweden in ages between 18-64 years 2010	Person	Persons Country of birth.	Person	Country of birth	Registered Person, Country of birth
	Person	Father Country of birth.	Father	Country of birth	Registered Person, Biological Father, Country of birth
	Person	Mother Country of birth.	Mother	Country of birth	Registered Person, Biological Mother, Country of birth

Fig 8 After

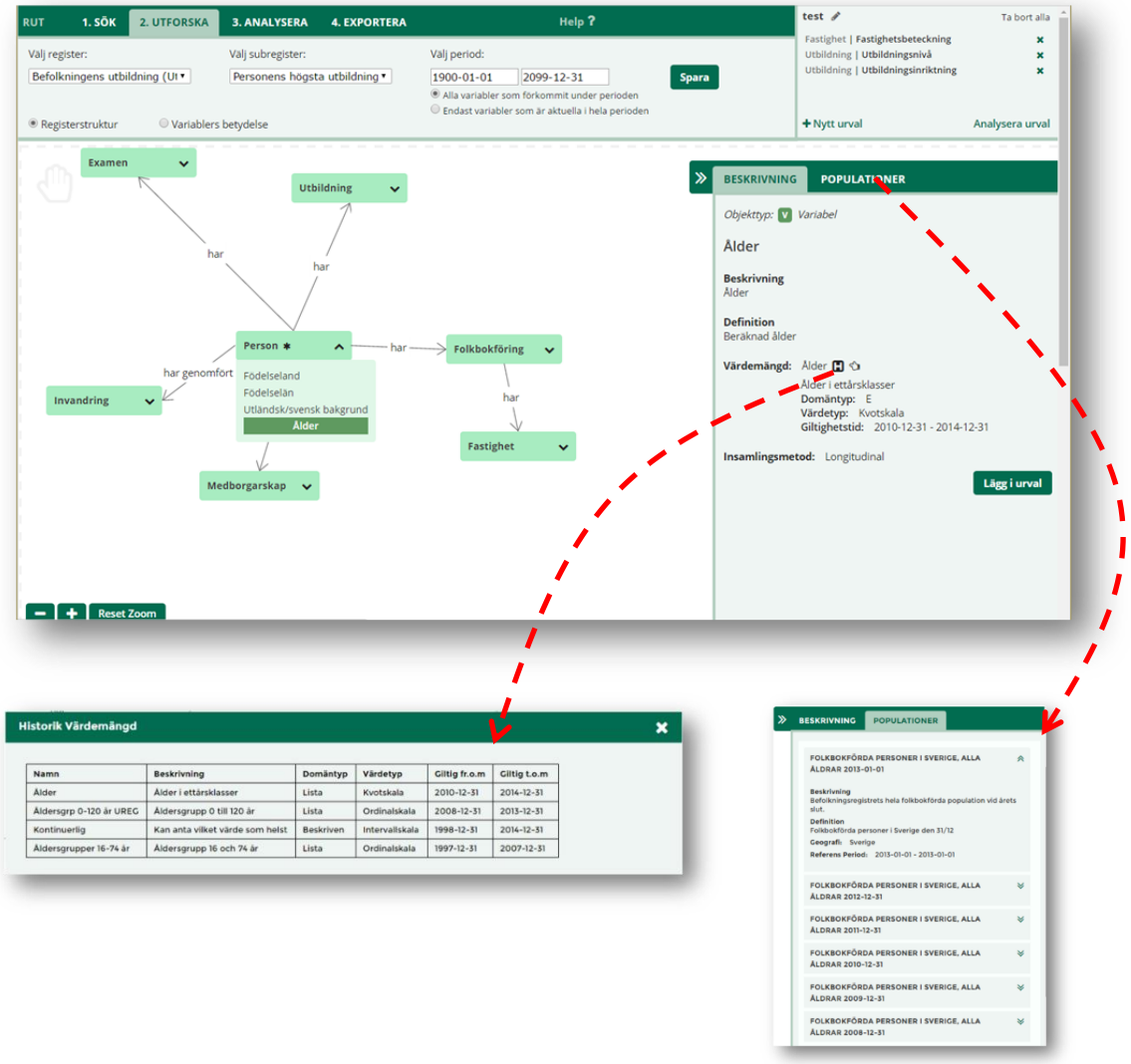

Fig 9 Screenshot - Variable grouping, Populations and presenting change history.

In the screenshot from the application above we can view the Variable Concepts grouped within Variable UnitTypes for a better overview over what the Variables are measuring. Since the representation is related to the used instance of the represented variable we present historic use of value domains when needed. The populations which the variable are measuring are also presented. Both of these are important for the researchers initial evaluation of the variable in relation to the research question.

One could argue that the use of the variable unit type would be the same as using concepts directly but by handling it as a specialization of a concept we get a better ability to handle differences in naming of variable unit types having the same meaning and vice versa.

Instance Variable, data or no data, that is the question? One of the main prerequisites coming into the project were the separation of metadata and data and because of legal constraints the solution where required not to hold any data. After some discussions and reviewing the GSIM examples we came to the conclusion that we needed a replacement for the instance variable that where not holding data but filling a quite similar role. We named this object the Contextual Variable and we use this object to hold information regarding when a represented variable has been

used within a population (and Unit Type). That is the reference period for the use combined with the source for the variables use during the reference period.

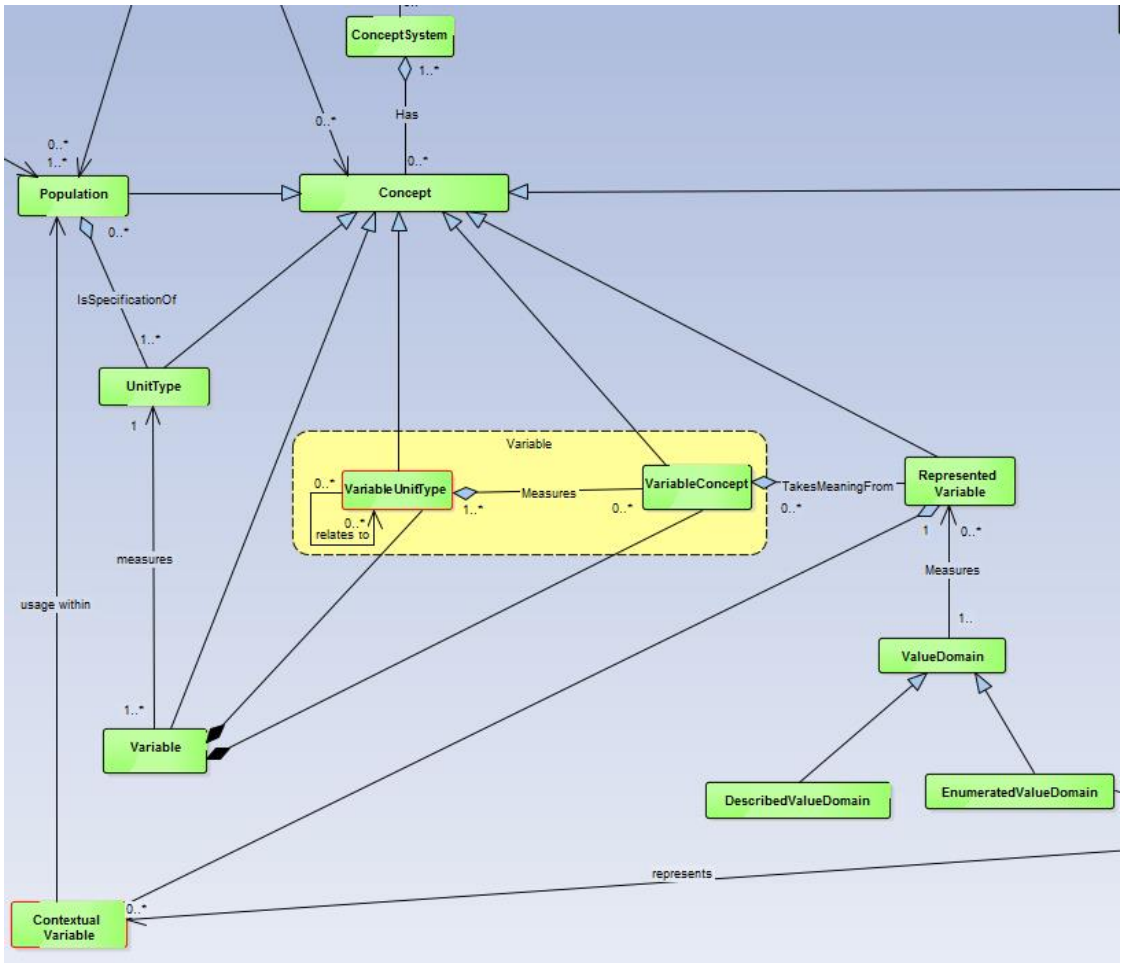

Fig 10 Contextual Variable

The Contextual Variable also plays a central role in the logical model where it holds the metadata events that defines the usage of a represented variable within a variable unit type and population at any given time.

3.2.4 Value domain – Node Set, Codelists and classifications

GSIM offers a strong support for our needs to hold information regarding the value domains, codelists and classifications that we need to manage in order to provide the researcher with information regarding the variables use of representation over time.

When presenting the representation used by a represented variable we use the Value Domain to hold the description of the codelist variant used by the represented variable. Regardless of whether the variant are based on a subset from a classification, which is often the case, or a separate codelist.

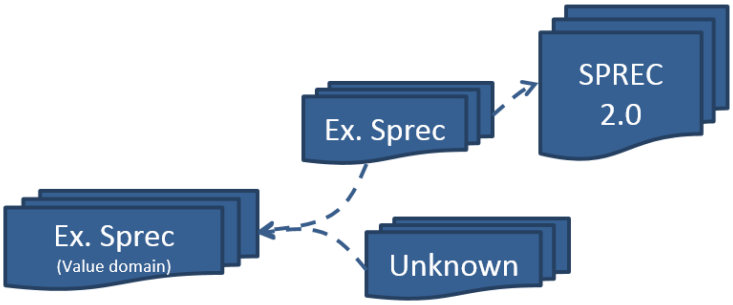

Fig 11 Value Domain

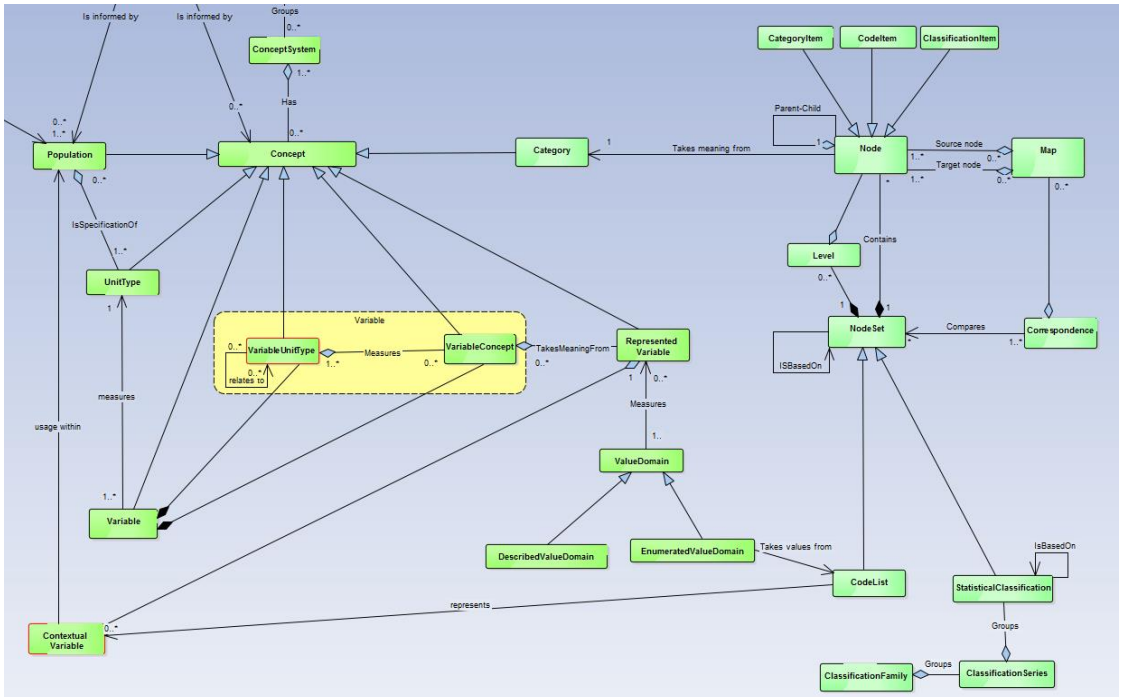

Fig 12 A subset of the selected GSIM objects in the context of the solution.

The codelist specialization of NodeSet that includes the version information represents the usage of the represented variable at a given time through the Contextual Variable.

4 Conclusions

4.1 GSIM – Enables researchers and enhances metadata maintainability

What was most striking during the initial work gathering core requirements from the researchers, and relating them to the register holders use of metadata to describe registers, was that there sometimes was a quite large gap between the researcher needs and the metadata created and used by the register holders.

The researchers placed a heavy focus on meaning and representation while the registers often had a larger focus on creating metadata on the variable level. Definitions/descriptions for variables lay on the variable instead of separated into concepts. Often researchers got extracts from registers with different variable names and similar definitions over time or the same variable names with different definitions.

Before being able to evaluate the variables in relation to the needs of the research project the researchers then needed to put in the work needed to separate the variable meaning from the variable names and also separate the representation over time.

The Concepts group in GSIM met the researchers need for separating the variable meaning into concepts and separating the representation from the (conceptual) variable. Initially the separation of the register holders metadata regarding meaning and representation from the variable name consumes a bit of work but, following that, the use of GSIM brings a metadata structure more easily maintained while in the same time allowing the researcher to allocate more time to research and less to (meta)data management.

4.2 GSIM – Enabling unambiguous communication by navigating increasing granularity

The inner workings of GSIM is of course hidden under the surface in the application but it supports the applications way of providing the researcher with increasingly higher granularity of metadata while navigating further into the application and selection and evaluation process. That process which starts with search, conceptual or through navigating registers and variants, and selection of a variable as an entry point follows the researchers wish to evaluate the variable from a conceptual point of view as stated by the reference group.
After initial time selection and evaluation of meaning and representation the researcher gets to evaluate the variable in detail within their selection. That is when the application moves from working with the variable to the represented variable, and when introducing time, contextual variable.
The selection of population, time and representation accompanies the increasingly higher granularity. When creating the export list that provides the means for communicating the design on a level of detail enough to remove most of the ambiguities from the researchers communication. At this stage the application provides the researcher with full GSIM support, from the concepts group, when requesting the data.