Metadata system meeting requirements of standardisation, quality and interaction and integrity with other metadata systems:
Case Variable Editor Statistics Finland

Sinikka Laurila
Statistics Finland, e-mail: sinikka.laurila@stat.fi

Abstract

What is Variable editor?
The ICT-strategy of Statistics Finland (updated in 2011) gives guidelines for developing metadata systems concerning both the content and the technology. Variable Editor is the first metadata application using the data model and technology according the strategy. Variable editor is an application for managing the metadata of data and variables, which are stored in an xml-database (eXist). The first version was created in 2010 and in April 2011, a project was established to coordinate the education and usage of the application. The project will end at the end of this year (2012) and after that, the work will continue in the Metadata service unit.

Standardisation of metadata
In designing the content of the metadata elements according to the CoSSI (Common Structure for Statistical Information) data model, it soon came obvious, that a lot of standardisation was needed. The more free text elements the application provides the more flexible the metadata element descriptions are. So standards were used whenever it was possible.

Controlling quality of metadata
The question of quality also became an important issue. The data and variable metadata descriptions have to follow a certain quality level so that they can then be further used for various purposes. In quality issues, the quality regulations of ESQRS are applied alongside with other quality frameworks.

Integration of metadata systems
Interaction and integration with other metadata systems at Statistics Finland followed the principle that information should be stored and maintained only once in one system from where it can then be copied elsewhere. Links were built in the Variable editor application to metadata databases like classifications, concepts, personal register, etc.

The European Statistical System (ESS) gives guidelines for managing and exchanging metadata. Although the metadata model used at Statistics Finland is our own CoSSI model and all our metadata systems follow it, we are prepared to exchange all metadata according ESMS or ESQRS using the SDMX data exchange concept. One way of doing this is using converters when making reports from the metadata database using Variable Editor.

Keywords: Metadata, Standardising Metadata, Metadata Information system, Quality in Metadata

1. What is Variable Editor?

The ICT-strategy was published in 2008 and updated in 2011. The strategy gave the policy guidelines and basis for developing metadata systems in future. There were two main policy issues which had to be fulfilled. The first was that the metadata model CoSSI-model should be used as a data model and secondly the metadata will be stored in an xml-database in the technical solution.

The CoSSI-data model is developed in Statistics Finland. It contains the data model for all metadata from publications to a single variable. When making the CoSSI-model existing metadata standards where applied broadly. The metadata model used at Statistics Finland cannot be separate. The conversions between our model and other standard metadata models must be possible. The CoSSI metadata model is a hierarchical and modular data model. The docmeta modul is almost the same for all metadata. The modularity helps when there are pressures to update or insert new metadata elements to the model.

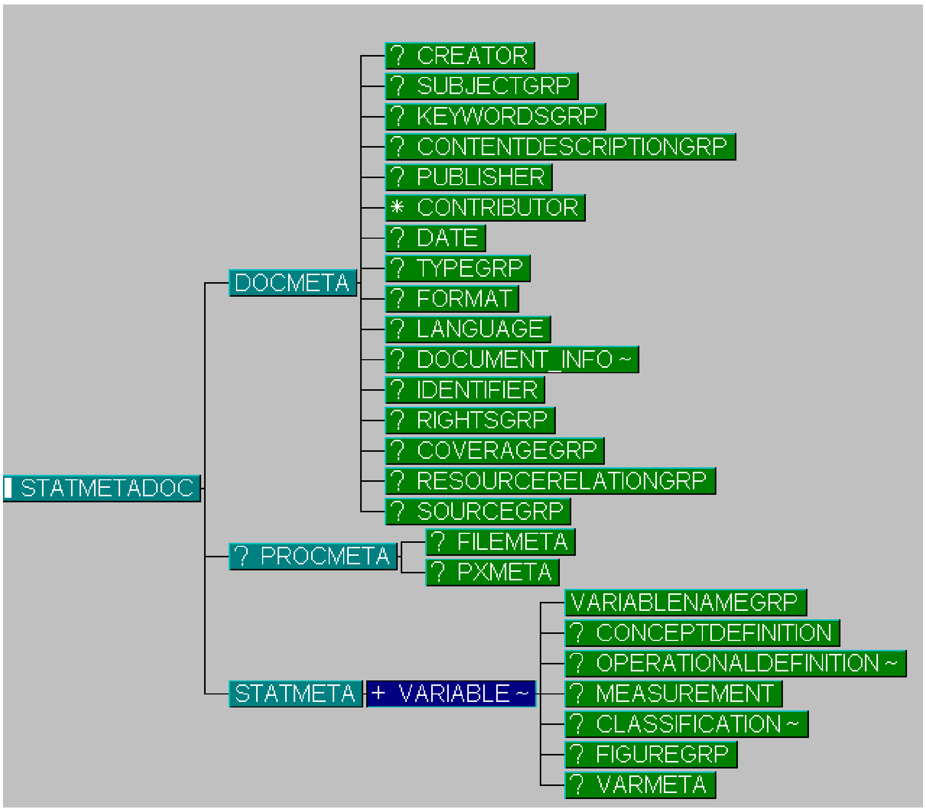


Figure 1. The CoSSI metadata model for statistical data and variables.

The current recommendation for metadata databases is an xml-database. The eXist-xml database is used at Statistics Finland. Today all published metadata is in xml-format. According to the ICT-strategy all current metadata systems, classifications, definitions and data and variable metadata will be maintained in an xml-database in coming years.

The Variable Editor application was made by MicroSoft.net. The application was ready in 2010. The implementation of Variable Editor was organised in a project, which lasted from April 2011 until the end of 2012. As a result of the implementation project about 190 statisticians were educated and today the meta database consists of about 700 metadata descriptions of data and its variables.

2. Using standards in Variable Editor

As a principle standards were always used in metadata elements when possible. Standards were usually in lists from where the user could choose the right item.
In document metadata such standard lists are available for:
- subject field and division
- maintenance
- keywords
- creator and contributor

In statistical meta-data lists are available for:
- concepts
- classifications
- measurement units

These lists are all maintained centrally.

The free text metadata elements are accurately instructed in the user guide. The instructions are based on a common agreement of the content of these instructions. There are still some metadata elements that are to be standardised in near future. These are statistical unit, population and data source.

The technical information of the variables is automatically transferred from the data file. The format and information vary according to the format of the database. For instance in SQL Server table it is crucial to know the key variables and indexed variables. The most common formats used today are SAS, SQL Server, PcAxis and various text formats.

3. Quality control in metadata database

The importance of quality issues has risen when developing the harmonisation and integration of different databases. Efforts are made to create understanding between national databases when they are combined into common databases. The information used in comparing the data is metadata. Metadata can be used only if it reaches good quality standards.

In case of the Variable Editor, good quality is seen as a necessity. Before using a centralised metadata system, statistics had their own metadata, usually stored in word- or excel-documents, which were maintained and used only by statisticians dealing with their specific subject field. Today, Variable Editor and it'-s metadata database are available to all, at least in reading mode. The effectiveness of the common metadata database also lies in the fact that certain statistics are responsible for defining certain metadata. For instance, population statistics determines the metadata of population phenomena. This also means that the correct basic definitions are found in the metadata of the "mother-statistics". The principle is that the correct metadata is stored and maintained only once, from where it can then be copied for various purposes, for instance in commissions.

The use of Variable Editor is promoted through standards and good guidelines, but this is not enough. The quality of the data and variable descriptions must be guaranteed also by inspections. These are carried out by using a quality matrix and producing quality reports. The quality matrix is used to see the quality of a single description. The matrix is filled out by the inspector, and it contains several checking points like the following:
- is the name of the description made using the instructions
- are all the mandatory elements described
- for free text elements, are the descriptions made as instructed

The reports are regurlarly produced from the metadata database. They produce information on statistics in the metadata database, important elements lacking information and logical errors. The reports are fixed and can be added when needed. Systematic inspection is crucial for maintaining the quality of the system.

In the near future we will implement the Eurostat Code of Practice to the quality checking of the metadata database.

4. Integrating metadata systems

Variable Editor is the first metadata system made according to the new policy guidelines. In few years time, the rest of the metadata systems are going to follow the same concept. However, the integration of metadata systems is important and the Variable Editor solution uses other metadata databases whenever possible. The figure below shows the metadata systems that have been integrated to the Variable Editor metadata application.

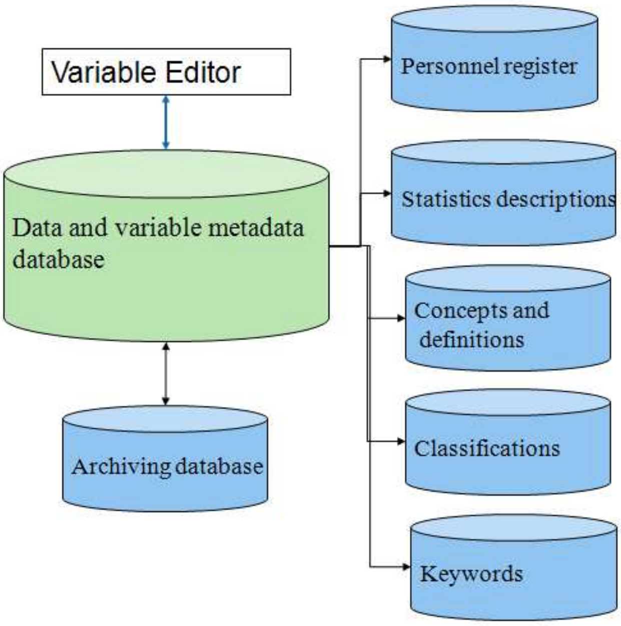


Figure 2. Data and variable metadata database and the integrated common metadata databases.

5. Metadata in statistical processes

Creating a common metadata database for data and variables was a crucial target in modernising the metadata databases. Another major target was integrating metadata into statistical processes from data collecting to data dissemination and archiving data. Metadata has been used in various ways in statistical processes. This has mainly been done differently in different statistical systems, sometimes common practices have been used, but usually different solutions have been used in different statistical systems.

Today Statistics Finland applies the Generic Statistical Business Process Model (GSBPM) in process management. Our goal is to make the xml-metadata database a common source for statistical processes at each process level. It is also important that the solutions which are made, can be commonly used. There is no point in using separate integration methods for different statistical systems.

This task is not realised quickly. The starting point is usually a real need to modernise a statistical system. In the modernising project both the IT-experts and metadata experts work together with statisticians. The processes of the data collection and data dissemination have been prioritised at the start of the project. In the coming years, this work will continue and it will result in statistical processes, which use metadata databases effectively in every phase.

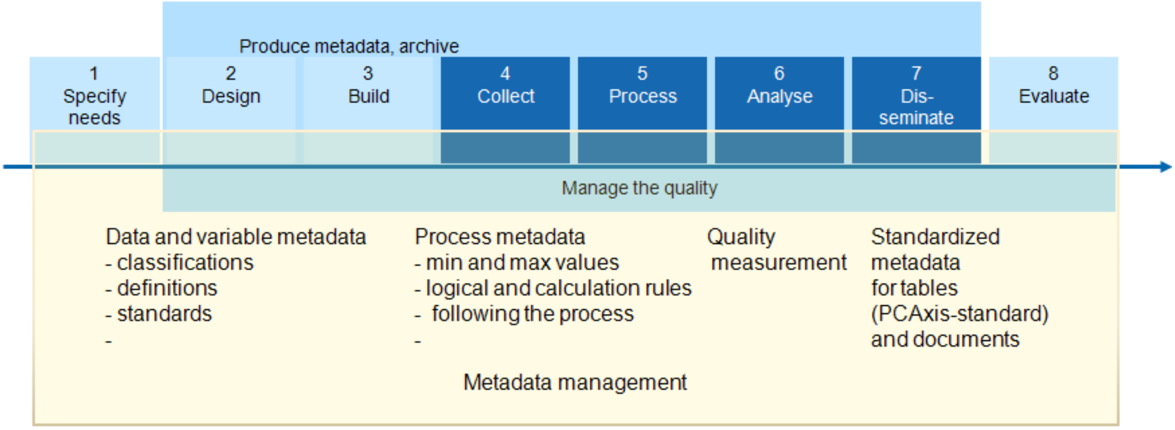


Figure 3. Statistical processes (The Generic Statistical Business Process Model) and the need for metadata in each process phase.

6. The impact of European standards ESS and SIMS on the national metadata system

The European metadata standard ESMS (European Standard for Metadata System) and it'-s extension SIMS (Single Integrated Metadata Structure) are used in Eurostat. - Eurostat also offers it'-s own metadata application for storing metadata (National Reference Metadata Editor, NRME). The Finnish metadata model consists almost the same of the elements as SIMS, but not fully.

A challenge for integrating the European and national metadata systems is to build good converters between them. The policy guideline at Statistics Finland is to firstly use the national metadata system Variable Editor and secondly other outside metadata systems. The European and the national metadata systems are both stored in xml-databases. The converters have been made using SDMX concept for Eurostat'-s ESMS and ESQRS meta data standards.

The European standards also define quality issues in producing data and metadata. The Code of Practice will be applied at the national level in 2013. The indicators of each principle in the Code Practice have been translated and estimated. The next step is to operationalise them further to suitable quality indicators.

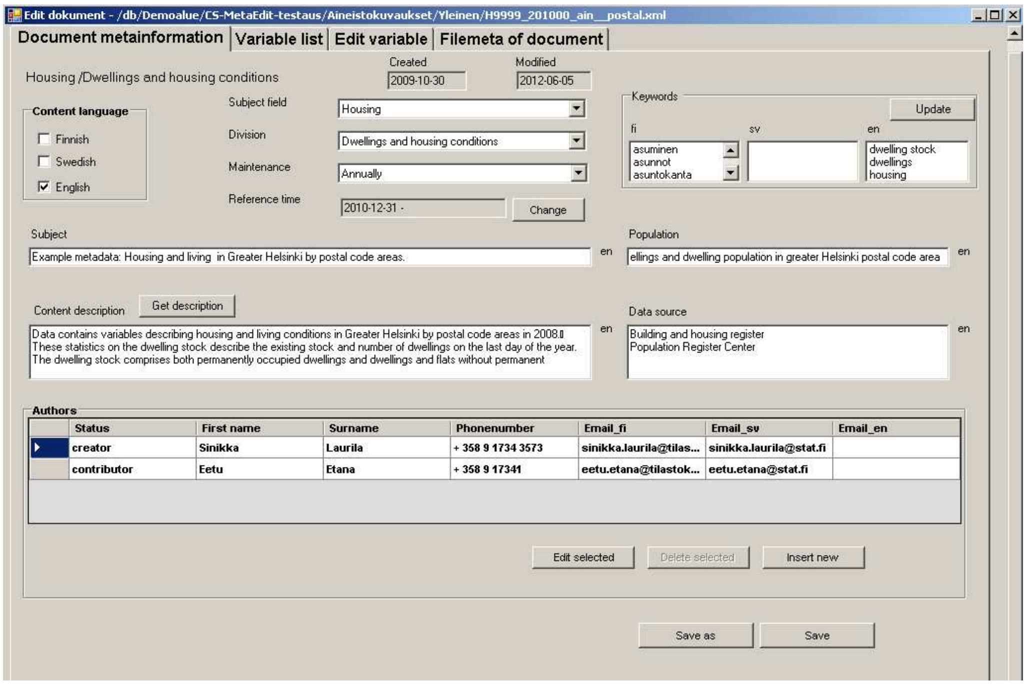
References

UNECE Secreretiat (2009) **Generic Statistical Business Model**. Version 4.
Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata
(METIS) in Geneve, 2009.

The CoSSI data model. Common Structure of Statistical Information. Version 1.6.
http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html

European Statistics Code of Practice.
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice

Appendix. The interface of -Variable Editor



Housing /Dwellings and housing conditions

Created: 2009-10-30
Modified: 2012-06-05

Content language
- ☐ Finnish
- ☐ Swedish
- ☑ English

Subject field: Housing
Division: Dwellings and housing conditions
Maintenance: Annually
Reference time: 2010-12-31 -

Keywords

fi	sv	en
asuminen asunnot asuntokanta		dwelling stock dwellings housing

Subject: Example metadata: Housing and living in Greater Helsinki by postal code areas. en

Population: ellings and dwelling population in greater Helsinki postal code area en

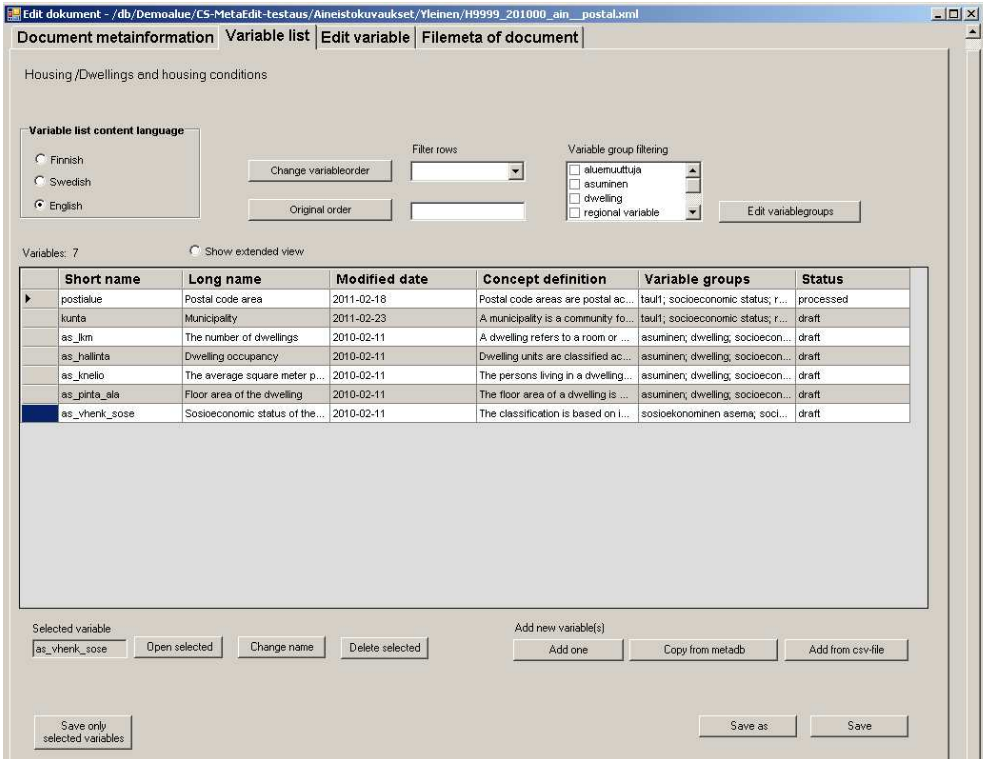
Content description:
Data contains variables describing housing and living conditions in Greater Helsinki by postal code areas in 2008.
These statistics on the dwelling stock describe the existing stock and number of dwellings on the last day of the year.
The dwelling stock comprises both permanently occupied dwellings and dwellings and flats without permanent en

Data source:
Building and housing register
Population Register Center en

Authors

	Status	First name	Surname	Phonenumber	Email_fi	Email_sv	Email_en
▶	creator	Sinikka	Laurila	+ 358 9 1734 3573	sinikka.laurila@tilas...	sinikka.laurila@stat.fi	
	contributor	Eetu	Etana	+ 358 9 17341	eetu.etana@tilastok...	eetu.etana@stat.fi	

Figure 1. The Docmeta metadata model of data.



Housing /Dwellings and housing conditions

Variable list content language
- ○ Finnish
- ○ Swedish
- ● English

Filter rows

Variable group filtering
- ☐ aluemuuttuja
- ☐ asuminen
- ☐ dwelling
- ☐ regional variable

Variables: 7

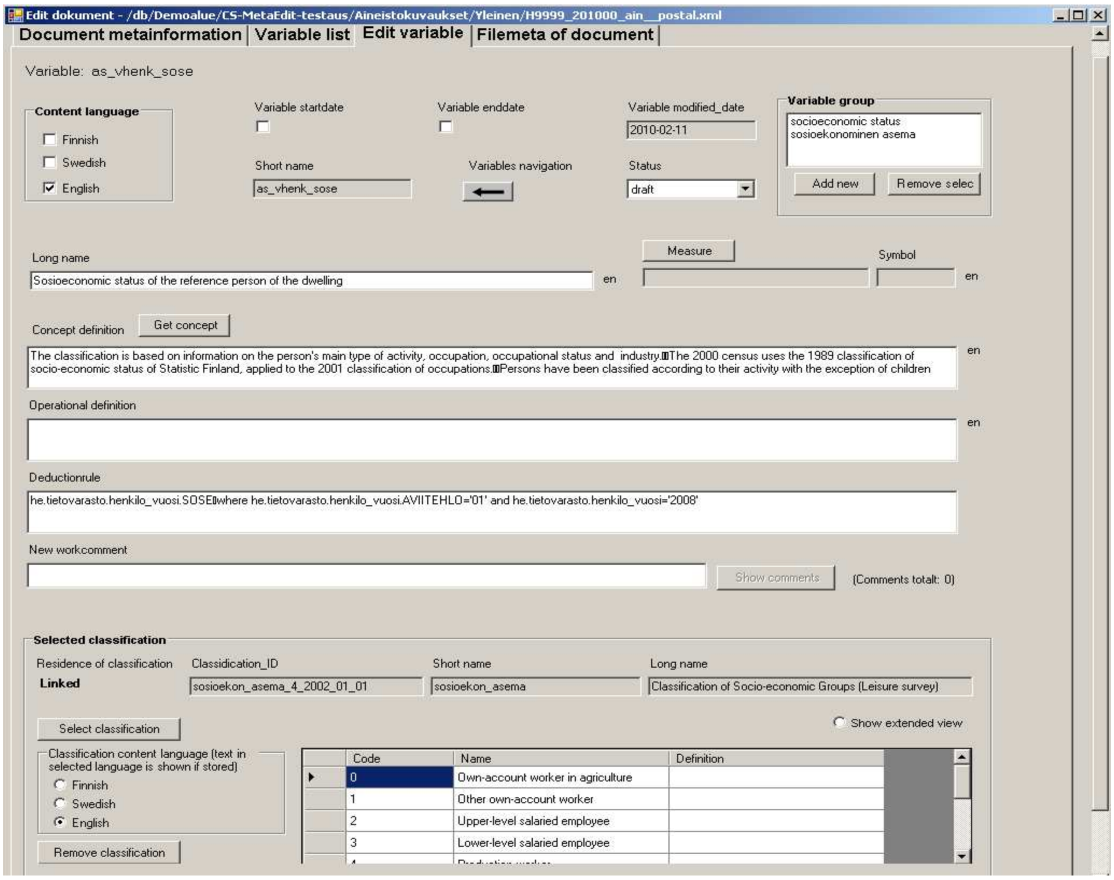
○ Show extended view

	Short name	Long name	Modified date	Concept definition	Variable groups	Status
▶	postialue	Postal code area	2011-02-18	Postal code areas are postal ac...	taul1; socioeconomic status; r...	processed
	kunta	Municipality	2011-02-23	A municipality is a community fo...	taul1; socioeconomic status; r...	draft
	as_lkm	The number of dwellings	2010-02-11	A dwelling refers to a room or ...	asuminen; dwelling; socioecon...	draft
	as_hallinta	Dwelling occupancy	2010-02-11	Dwelling units are classified ac...	asuminen; dwelling; socioecon...	draft
	as_knelio	The average square meter p...	2010-02-11	The persons living in a dwelling...	asuminen; dwelling; socioecon...	draft
	as_pinta_ala	Floor area of the dwelling	2010-02-11	The floor area of a dwelling is ...	asuminen; dwelling; socioecon...	draft
	as_vhenk_sose	Sosioeconomic status of the...	2010-02-11	The classification is based on i...	sosioekonominen asema; soci...	draft

Selected variable: as_vhenk_sose

Add new variable(s)

Figure 2. The Variable list.



Edit dokument - /db/Demoalue/CS-MetaEdit-testaus/Aineistokuvaukset/Yleinen/H9999_201000_ain__postal.xml
Document metainformation
Variable list
Edit variable
Filemeta of document

Variable: as_vhenk_sose

Content language

- □ Finnish
- □ Swedish
- ☑ English

Variable startdate □

Variable enddate □

Variable modified_date: 2010-02-11

Variable group: socioeconomic status
sosioekonominen asema

Short name: as_vhenk_sose

Status: draft

Long name: Sosioeconomic status of the reference person of the dwelling en

Symbol en

Concept definition

The classification is based on information on the person's main type of activity, occupation, occupational status and industry.▯The 2000 census uses the 1989 classification of socio-economic status of Statistic Finland, applied to the 2001 classification of occupations.▯Persons have been classified according to their activity with the exception of children en

Operational definition en

Deductionrule

he.tietovarasto.henkilo_vuosi.SOSE▯where he.tietovarasto.henkilo_vuosi.AVIITEHLO='01' and he.tietovarasto.henkilo_vuosi='2008'

New workcomment

(Comments totalt: 0)

Selected classification

Residence of classification	Classidication_ID	Short name	Long name
Linked	sosioekon_asema_4_2002_01_01	sosioekon_asema	Classification of Socio-economic Groups (Leisure survey)

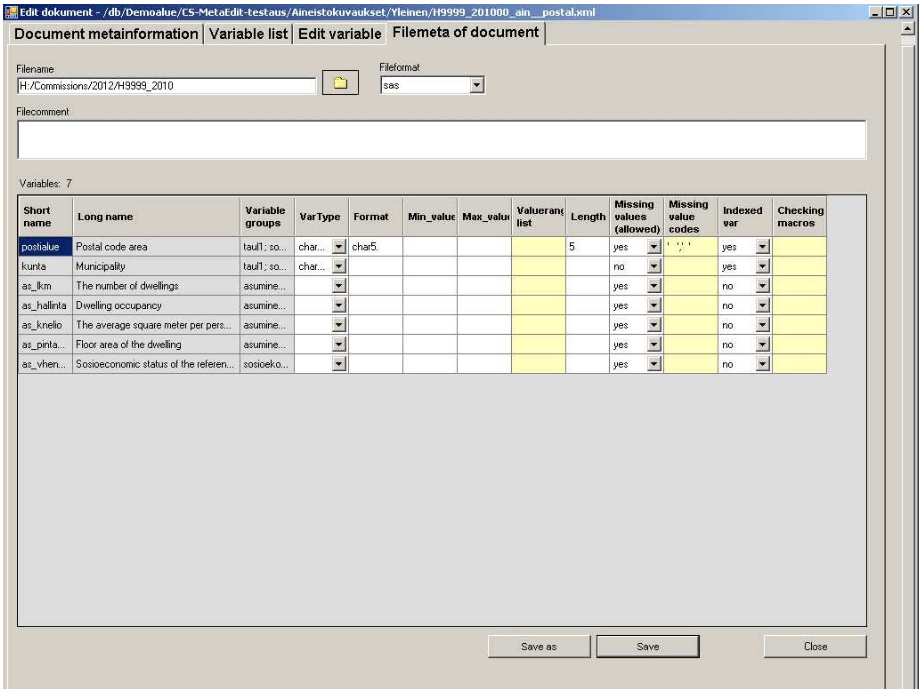
○ Show extended view

Classification content language (text in selected language is shown if stored)

- ○ Finnish
- ○ Swedish
- ● English

Code	Name	Definition
0	Own-account worker in agriculture	
1	Other own-account worker	
2	Upper-level salaried employee	
3	Lower-level salaried employee	

Figure 3. The Statmeta metadata model of variables.



Edit dokument - /db/Demoalue/CS-MetaEdit-testaus/Aineistokuvaukset/Yleinen/H9999_201000_ain__postal.xml
Document metainformation
Variable list
Edit variable
Filemeta of document

Filename: H:/Commissions/2012/H9999_2010

Fileformat: sas

Filecomment

Variables: 7

Short name	Long name	Variable groups	VarType	Format	Min_value	Max_valu	Valuerang list	Length	Missing values (allowed)	Missing value codes	Indexed var	Checking macros
postialue	Postal code area	taul1; so...	char...	char5.				5	yes	' ';' '	yes	
kunta	Municipality	taul1; so...	char...						no		yes	
as_lkm	The number of dwellings	asumine...							yes		no	
as_hallinta	Dwelling occupancy	asumine...							yes		no	
as_knelio	The average square meter per pers...	asumine...							yes		no	
as_pinta...	Floor area of the dwelling	asumine...							yes		no	
as_vhen...	Sosioeconomic status of the referen...	sosioeko...							yes		no	

Figure 4. The Varmeta metadata model of data.