CONFERENCE OF EUROPEAN STATISTICIANS

ModernStats World Workshop 2018
11-13 April 2018, Geneva, Switzerland

Status of statistical standards and ModernStats models in SURS

Julija Kutin (Statistical Office of the Republic of Slovenia, the Republic of Slovenia)
Email: Julija.Kutin@gov.si

Abstract

The Statistical Office of the Republic of Slovenia (SURS), uses ISO/IEC 11179 Metadata Registry (MDR) standard for describing variables and questions for web questionnaires.
We are obligated to use the SDMX standard for disseminating statistical data and metadata. Our system for reference metadata and quality indicators is based on ESMS and ESQRS (SIMS).

It was very appreciated to have an opportunity to implement the similar version of the GSBPM, because we did not have our own model for the statistical process. Statistical surveys documentation system, which has three levels of documents, is based on it. Modernisation of the statistical production systems is a demanding and certainly a long-term task, which includes the development of new IT solutions as well as a radical change in the whole production system at the institutional level. In recent years we have been putting a lot of our resources into the projects aiming at moving from a domain-oriented system to generalised, process-oriented solutions.

We got in touch with DDI and CSPA during ESSNet on "Sharing common functionalities in the ESS". We integrated "ENO Questionnaire Generator service" into our system.

1. SDMX

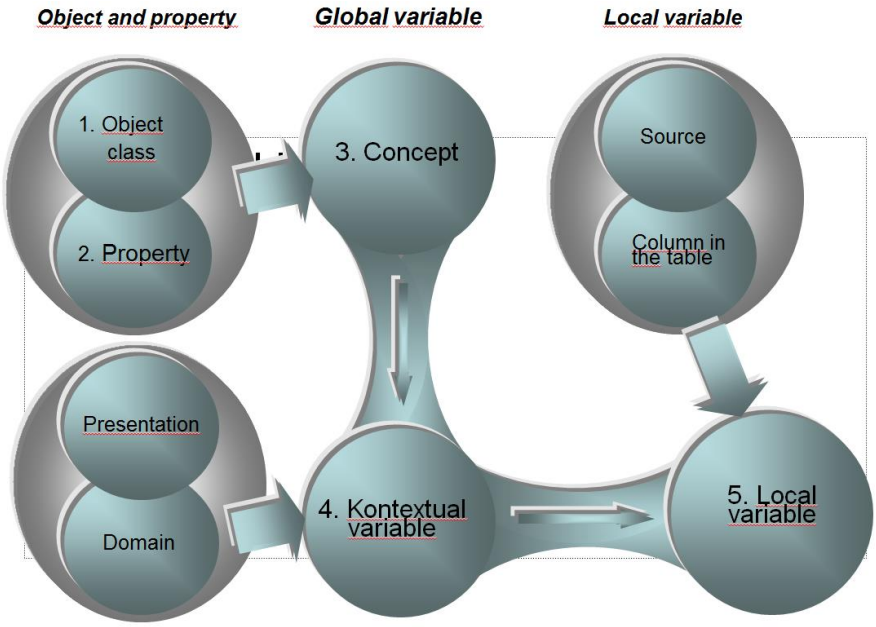
The most common statistical standard SDMX is implemented at SURS. We are using it for sending data (not all) and metadata (not all) to Eurostat and the IMF. Due to SDMX tools and standard format for sending data, the machine-to-machine data exchange is possible. In a machine-to-machine data exchange, standards are very important.

Sometimes we have problems because not all the data of all surveys are in databases. We still have a lot of manual work. The dissemination process is not metadata driven. On the other hand, the problem with metadata is even bigger. It is not a problem of the tool, but a problem of the content. There is not standardised Metadata Structure Definition (MSD) for all international organisations not really for SDMX sponsors. There are plenty of different MSD for each domain. At the beginning it was promised that we will use a common metadata reporting structure, which would allow us to disseminate harmonised metadata for all of our published datasets. But there are still different rules for metadata dissemination, and non-standard tables for indicators not really at Eurostat for all domains.

We have a new database for reference metadata and indicators, but we still have to prepare Metadata files for the IMF manually.

2. ISO/IEC 11179 Metadata Registry (MDR)

The second standard is ISO/IEC 11179 Metadata Registry (MDR) standard. We only use the part of the standard that refers to object class and variables. Some modifications and adaptations according to our system have been done. Our own needs were taken into account. It is now our global registry for variables and questions – SDM module (Picture 1.), as a part of Information System of Integrated Statistical Processing for Business Statistics. In SDM reusability is very important. It shall be guaranteed that two different data sets are using the same specific represented variable as instance variable, when the definition is identical. The aim is to prevent unnecessary duplication of variables within the organization.



Picture 1: SDM module, which is built on ISO/IEC 11179

That means that for the time being we are documenting only variables used for electronic reporting. Local variables are connected to questions and sub-questions of the electronic questionnaire.

Our SDM module allows that global variables are linked to different surveys and questions. For example, our questionnaires have some standard questions at the beginning and at the end. Those questions are the same for all surveys and linked to the same global variable. But there is a problem when one other – not standard - global variable is linked to two or more questions in different surveys. That means that we are asking our reporting units twice or more times the same question and we try to avoid that as much as possible.

Our system for e-reporting allows us to personalise the questionnaire. Question rooting is possible and some response validations are included: mandatory questions, multiple choice questions, questions with specific domain or rank of values. The response has to be inserted in defined format. It all depends how the variable connected to the question is defined.

Definitions of variables allow us to easily build the database for row data, they are used for data editing, etc.

The biggest problem of this system is that it allows us to define tables with more columns and lines on questionnaires, but the questionnaire generator is not able to generate them.

DDI is a standard which just come to our office. It is very complex, so we need time to study it. We have already mapped our ISO/IEC 11179 concepts with DDI concepts. DDI complex building blocks were not mapped, because we do not have such concepts in SDM.

For the time being we are testing how useful the DDI could be. We would like to use it for generating questionnaires.

3. GSBPM

We took the opportunity of the strength of the Modern Statistical Models to facilitate the statistical community to develop their own rigorous systems. That means that GSBPM is generic but not prescriptive. There is flexibility to allow for national implementations. So we developed the adapted version of GSBPM with different steps to produce our official statistics. It is **SURS's Generic Process Model**.

.

The biggest difference to GSBPM is that SURS's Process Model does not have the phase Build. We found out that it would be necessary to add this phase to SURS's Process Model.
And the second one is that we added the phase "The selection of observation units", which has four sub-processes: "Preparation of data source for constructing the sampling frame", "Constructing a sampling frame", "Selecting observation units" and "Creating the address list".

This is our standard for describing processes and individual statistical procedures based on foreign practices and standards. It provides standard terminology to describe the statistical process in a coherent way. That allows direct comparability and connectivity between phases and sub-processes of different statistical surveys and with other members of the international statistical community.

SURS's Process Model is mostly used to manage and document statistical production. On its basis the **Documentary system for statistical surveys** and two very important documents – guidelines for managing and documenting statistical production - **Guidelines for Quality Assurance (QA),** and **Descriptions of processes of the statistical survey** are prepared.

One of the basic policies of the Statistical Office of the Republic of Slovenia is the commitment to constantly monitor and improve the quality of statistical processes and products. A comprehensive and systematic description and guidelines for correct implementation of individual phases are given in **Guidelines for Quality Assurance**.

SURS's Process Model comprises eight phases that are broken down into different numbers of sub-processes. The chapters and sub-chapters of this document follow the structure of this process model. Each chapter consists of a short description of the entire phase; all sub-chapters are broken down into two sets: the first set contains a general description of the sub-process, whereas the second set provides QA guidelines to be applied by producers as a check list of quality elements to which attention must be devoted when conducting sub-processes. The manual Guidelines for Quality Assurance is intended for all producers of

national statistics that are in any manner involved in the statistical process, other producers of statistical surveys, and everyone who is not only interested in the final statistical outputs (statistical data and information) provided by producers of national statistics, but would also like to become familiar with the background and entire survey process ultimately leading to useful and reliable statistical information.

In addition to the Guidelines for quality assurance that are publicly announced, we have also prepared the internal document: **Descriptions of processes of the statistical survey**. The document completes the Guidelines for Quality Assurance with guidelines helpful for SURS's employees in getting acquainted with the process of the statistical survey. To the definition of each sub-process two chapters are added: Implementation steps, and who performs them or (after) cares about the implementation, and Related Documents / Links. The document represents a standard framework and contains the standard terminology required for coordinated operation and training of SURS's employees, and at the same time also a basis for systematic introduction of improvements in individual processes and statistical surveys.

The documentation of statistical surveys is collected in a new documentary system STATDOC and is an important part of the statistical process. A good documentation system improves the quality of data, statistics and metadata. Currently, a new documentation management system is in place. Documents are standardised and structured on the basis of SURS's Process Model.

Before the documentary system was implemented, there was a problem that only a few statistical surveys were documented. If there was the documentation, it was a problem to find it and hard to know which is the newest version. It was sometimes lost on old computers or somewhere in drawers. If we found the documentation, it was not in structured form.

Finally, the preparation of the survey documentation is a planned part of the survey implementation. Documents have to be updated for each survey instance. A new documentary system STATDOC allows us to have the documentation of statistical surveys in one place for all surveys. The information cannot be lost, because unauthorised access is not possible. Standard form of the documents allows comparability between phases and sub-processes of different surveys.

Documentary system STATDOK is internal. It works on SharePoint and is hierarchically built at three levels. Standardized templates for each level and guidelines for fulfilling the templates are prepared.

Documentation for each survey is stable at the first two levels. It contains survey name, responsible person, responsible department, level of documentation, version of the documentation, the validity of the document, process which it belongs to.

At the first level we have standard Excel template with key information on conducting the survey. The template has seven sheets - one sheet for each phase. The first two phases are documented together. Sub-processes at the third level of SURS's Process Model are briefly described. It is the basis on which it is possible to compare the key steps of the implementation of the sub-processes of surveys. The Metadata are standardized as much as possible, which means that we use classifications to define a lot of metadata of the template.

The second level of the documentation is intended to describe sub-processes in greater detail. There are seven standardised Word templates - one for each phase. Each sub-process at the second and third levels of SURS's Process Model is described with all details. They include a list of all documents at the third level.

At the third level the implementation documents are stored with a “recipe” for carrying out the survey – more detailed instructions, which are often changing with each survey instance. They are variable and versionable on an annual basis. There are more documents for each phase. They are not all standardized and they are not all in the same format.

STATDOK is useful for Methodologists, because they can get answers for the questions such as How did we do it the last time?, How should it be done?, How is it done in other surveys?, Do we have data for 19XX?, Are they comparable?, Where can we find them? Also for the head of the organisation unit this system could be very useful. For example, it helps us to teach a new co-worker, when someone is absent, is getting another job or is getting retired. The assignment is faster and easier. When we want to optimise the sub-process of the survey, we can check how other better optimized surveys in or outside our organization are documented and implemented.

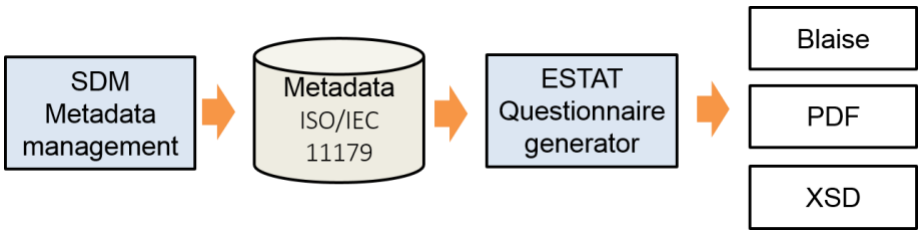
Phase / sub-process administrator and leadership also use documents in STATDOC, because it allows us to prepare all kinds of analyses: Where we could standardize the survey? Which surveys are similar? There is the similarity according to what? Which surveys include the handling with personal data? Are there critical points in the process X?

4. CSPA

Applying CSPA as an architectural blueprint might enhance the idea of sharing tools and services in the global community of official statistics. We tried to implement the Questionnaire generator, developed in INSEE, because we are not fully satisfied with ours which has quite a few shortcomings. Our process of designing and generating the questionnaire is centralised, complicated, with a lot of manual work. It does not allow us to make tables in questionnaire.

The idea was to enable statisticians to design and view questionnaires before the questionnaire is visible in the test environment of the ESTAT data collection portal.

Our web data collection portal ESTAT is used for interviewing enterprises. Questionnaires are generated with metadata driven questionnaire generation on the basis of the metadata in our metadata base – SDM module. ESTAT questionnaire generator transforms metadata, prepared in ISO/IEC11179 standard, to Blaise questionnaire (Picture 2.). Some manual modifications are needed to have it ready for data collection.



Picture 2: The existing SURS questionnaire generating system with SDM module

There was the decision to take advantage of an opportunity CSPA in bringing and try with INSEE questionnaire generator – ENO questionnaire generator, which is CSPA complained. On the other hand our system is not CSPA complained, but it is modular and that was enough to replace our questionnaire generator with a new one. It happened in June 2017, which was a major milestone in this reuse case.

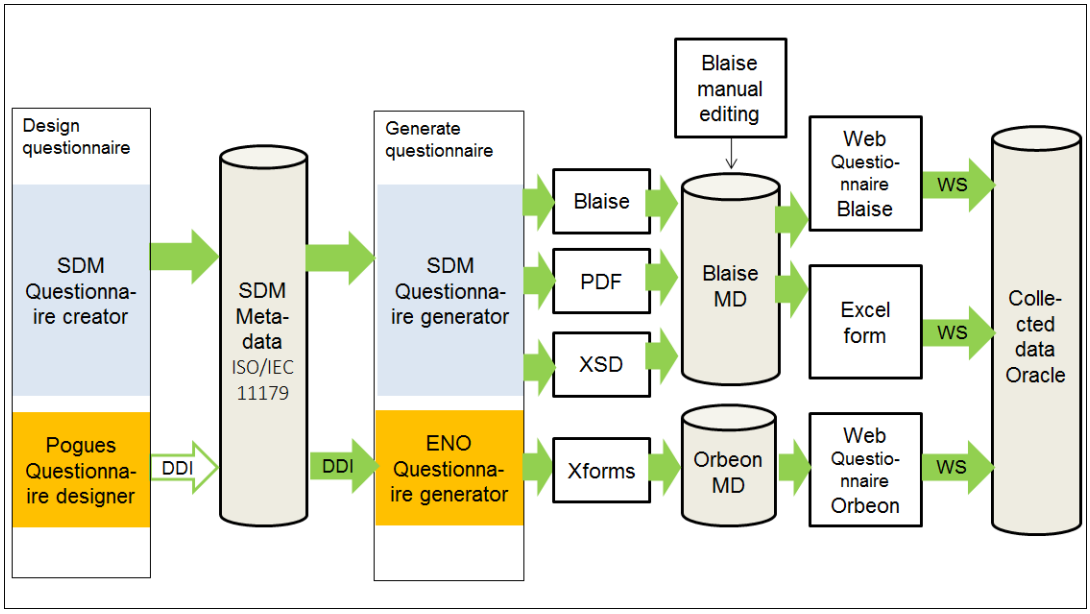
ENO questionnaire generator use DDI concepts for generating questionnaires. That is why, SURS started to analyse the requirements for generating complex questionnaires using DDI.

The analysis showed that most of the necessary metadata for new questionnaire generator already exist in the ESTAT SDM metadata repository from which Blaise questionnaires are generated. That is why we had to map our ISO/IES11179 concepts to DDI concepts. We used our own made ENO DDI mapper. It was not so hard, because standards are very similar and concepts, codes/representations, categories, etc., are similar to DDI concepts. DDI is oriented toward the ISO/IEC 11179 model.

DDI file with all necessary metadata for simple questionnaire was successfully exported from our SDM metadata module and imported to ENO questionnaire generator, which successfully made a web questionnaire. It was a simple one – without tables with hierarchical row and header codes and it was not linked to our metadata of surveys, activities and classifications (that means that we had to prepare classifications manually).

The questionnaire was successfully integrated into existing ESTAT web portal and data were successfully collected to our database as with our other web questionnaires (Picture 3.).

The reuse of the ENO generator was more successful with the simultaneous reuse of Pogues designer designed of INSEE. We started DDI questionnaire modelisation. The file exported from Poques could not be imported to our SDM module.



Picture 3: The existing SURS questionnaire generating system and new shared components

At the top of the picture there is our process so far. Below there are two new boxes: ENO Questionnaire generator and Poques Questionnaire designer. The connection between our SDM metadata base and ENO Questionnaire generator was made successfully. But the connection between Poques Questionnaire designer and SDM metadata base was not. Finally, we managed to make a direct connection from Poques Questionnaire designer to ENO Questionnaire generator. This is not in production, but ENO could be.

5. Challenges and next steps

SURS decided that not only users of our data and our web site are important, but it is also necessary to think about the modernization and digitalization of internal processes. With the higher quality and efficiency of processes, we can save time and money, which can be dedicated to providing high quality and more reliable data for our users.

The modernisation of internal processes is faster and easier by applying international standards and using modern statistical models as much as possible, so we really appreciate them. We will choose methods, standards and technologies that are more innovative, and thus have an important organizational impact. By using modern statistical models, we can avoid the duplication of efforts - avoid overlapping activities.

We are going on with using SDMX.

The analysis showed that the existing SDM metadata module should be redesigned in order to support complete DDI specification and integration of Pogues as a visual tool for designing and managing questionnaires metadata. That is why we have to meet with DDI deeper even if it is complex. It is planned to use ENO DDI compliant SDM as common metadata repository for questionnaire design & generation (Blaise, Xform, Pdf, etc.…).

It was also decided that redesign of the SDM module will be started after the review of internal questionnaire design standard (complex questionnaires generation).

We are working on expanding the use of GSBPM to all statistical surveys. GSBPM should promote the benefits and synergies that can be obtained by also implementing the other modernisation models (GAMSO, GSIM, CSPA).

And there are other models (CSPA) coming to the office to be tested and implemented. Reusability is never just a technical issue: semantic, organizational and even legal interoperability levels are impacted. The use of CSPA should be considered for the specification and development of services for applications according to the CSPA guidelines.