Distr.
GENERAL

Working Paper
28 February 2014

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2014)
(Dublin, Ireland and Manila, Philippines 14-16 April 2014)

Topic (ii): Standards-based modernization

Capturing Metadata Objects in Statistical Business Processes and Using Them to Monitor the Process

Prepared by Deniz Özkan, Güneş İnan Vıcıl, Turkish Statistical Institute[1], Turkey

I. Introduction

1. International community put emphasis on creating common industrial standards in order to "industrialize" production of statistics. Modernising statistical outputs as well as production methods are the two key strategic objectives of The High-level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS). Conceptual standardisation and creating a common terminology was considered as essential in this process. GSBPM and the GSIM were developed and accepted as industry standards by the international community, to help to arrive at standardisation at the conceptual level.

2. Turkish Statistical Institute adopted GSBPM to define the statistical production processes and to identify the metadata using this standard model. The objective is to design a new metadata system around GSBPM for the efficient functioning of the business processes.

3. A draft statistical business process model was prepared by TurkStat and the metadata about the processes were collected from the subject matter departments. This process metadata is organised, analysed and the metadata objects flowing between the processes were tried to be standardised for TurkStat's statistical production processes.

4. The aim of this paper is to discuss the ways to standardise, capture, store, re-use and document the metadata to monitor the statistical business process. In the paper, the first issue that was discussed is; how could the types

[1] The content of this paper does not reflect the official opinion of Turkish Statistical Institute. Responsibility for the information and views expressed in the paper lies entirely with the authors.

of metadata in statistical processes be specified. Second, how could these specified process metadata be standardised under a common terminology. Finally, we discuss how we can capture the process metadata as it happens, keep track of it during the process flow to manage the process; and potentially re-use this process metadata in a metadata system or repository.

II. Process Metadata

A. Standards

5. Metadata is the key for managing the process. If organizations don't want to lose control over their data and their processes then the data and the metadata need to be documented in a systematic way. This was stated clearly in GSBPM document, paragraph no.115 (GSBPM Version 5.0, UNECE, 2013):

> *"115. Good metadata management is essential for the efficient operation of statistical business processes.... The key challenge is to ensure that these metadata are captured as early as possible, and stored and transferred from phase to phase alongside the data they refer to. Metadata management strategy and systems are therefore vital to the operation of this model and these can be facilitated by the GSIM."*

6. GSIM is designed to help improve communication by creating a common terminology across and between statistical organisations. GSIM is used to identify metadata flows within GSBPM sub processes, to describe and define the business processes needed to produce statistics and also to enable harmonizing statistical computing infrastructures and facilitating the sharing of software components.

7. Turkish Statistical Institute (TurkStat) adopted GSBPM to model the statistical production process and to identify the metadata flowing through the processes. The model was slightly modified in the sub process level (level 2) and sub sub-processes (third level) were added as well. The 7 phases of statistical production process and 33 sub-processes within each phase can be seen from figure 1.

Figure 1. TurkStat's Statistical Business Process Model

Quality Management / Metadata Management						
1.Specify Needs	**2.Design**	**3.Build**	**4.Collect**	**5.Process**	**6.Analyse**	**7.Disseminate**
1.1.Determine needs for information	2.1.Design statistical products and outputs	3.1.Build and develop production system components	4.1.Establish frame and registers, select sample	5.1.Clasify and code	6.1.Evaluate the information for its effect	7.1.Update dissemination systems
1.2.Consult and confirm needs	2.2.Design frame and sample methodology	3.2.Integrate production system with other systems	4.2.Set up collection	5.2.Micro-edit	6.2.Produce statistics	7.2.Produce dissemination products
1.3.Establish output objectives	2.3.Design data collection methodology	3.3.Test production system	4.3.Run collection	5.3.Macro-control	6.3. Ensure statistical quality	7.3.Manage release of dissemination products
1.4.Check data availabilty	2.4.Design statistical processing and analysis metodology	3.4. Finalize production system	4.4.Finalize collection	5.4.Impute	6.4.Examine and evaluate statistics	7.4.Manage user queries
1.5.Prepare business plan	2.5.Design production systems and workflows			5.5.Calculate weights and derive variables	6.5.Prepare statistics for dissemination	
					6.6.Finalize content	

8. Standardisation is the first step to be taken to modernise statistics and industrialise the production of statistics. In order to standardise the process flow and process metadata at the international level using GSBPM and GSIM, standardisations need to be done within the organisations first.

9. While standardising the metadata objects we used an internal jargon and tried to standardise the objects around the terminology commonly used by the organisation; not relying on the GSIM glossary.

10. The Euro SDMX Metadata Structure (ESMS) is an ESS standard for reference metadata. The ESMS concepts mainly refer to the statistical concepts which are describing and documenting the statistical business processes and the outputs for data users. (Consoli A., Götzfried A., Linden H., Rychel B., 2013)

B. Challenges

11. TurkStat collected process metadata from the subject matter departments including the information such as: process steps, inputs and outputs of the process steps, suppliers/users of the inputs and outputs and executers of the work (agents), as well as the systems and tools used in the processes. The information was collected in a free text format under these topics.

12. In order to organize this free text information and model them, a detailed analysis of the process steps, inputs, outputs and the systems used have to be done. Based on these analyses, metadata objects were identified in each sub-process. It was seen that every department used their own technical jargon to state the process steps, work undertaken and input and output of each process step. These different expressions and jargons used by different departments needed to be brought together under standard object names.

13. One of the challenges that were encountered during this identification and standardisation process was the difference between the levels of these objects, i.e. hierarchy from a general level to the more specific levels. Some of the information provided by the departments was more general and some of them were more specific.

14. Solution for this issue could be creating a list to cover objects at each level, and bring out the hierarchical relationship between these objects. Then standardise each specific object within its own group of objects. At this point a common vocabulary was needed. This does not mean that every concept is identified by a single unique name, but concepts were given a set of terms when a single name could not be used, the goal is to create a reference almost like a classification of terms or thesaurus.

15. Standardisation exercise for metadata objects was done following a bottom up approach in three steps:
    Step 1. Identify the metadata objects at the most detailed level using process information (Figure 2a).
    Step 2. Create clusters of the metadata objects (Figure 2b).
    Step 3. Define the hierarchical relationships between these clusters (Figure 2c).

Figure 2a

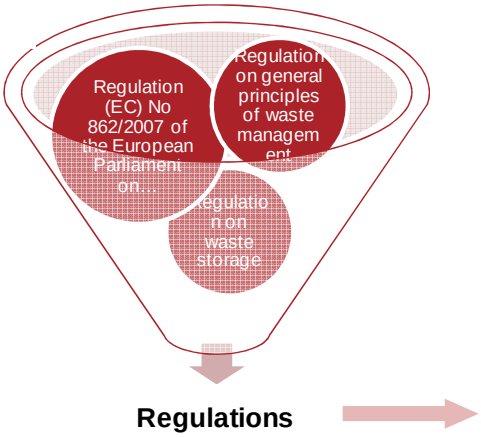


Figure 2b

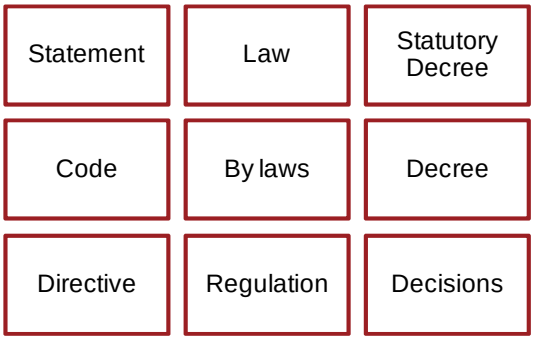
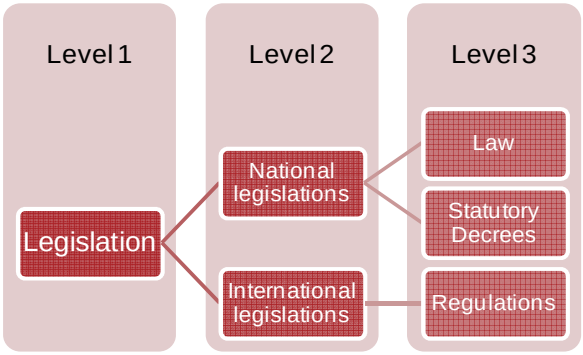


Figure 2c

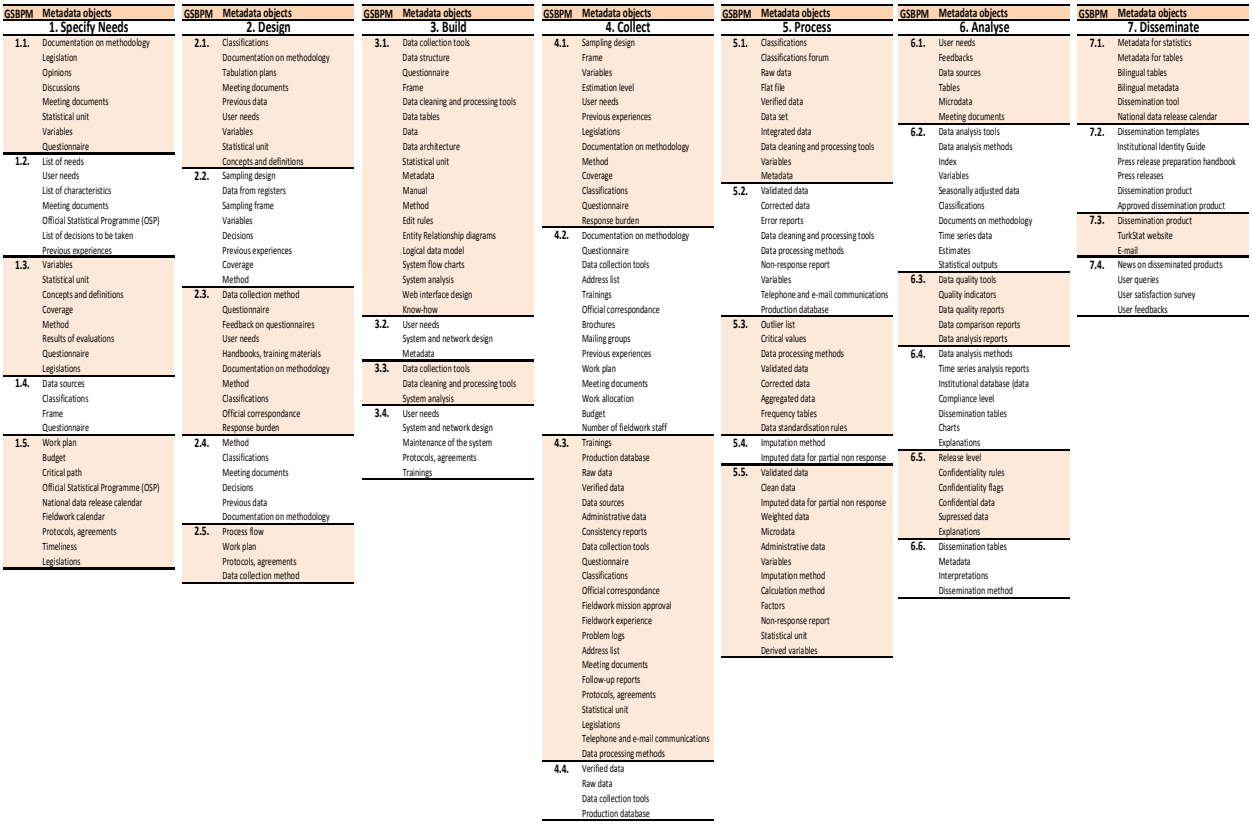


16. The levels of hierarchy between the objects could go down as low as to five or six levels, depending on the detail of information provided by the departments. After the standardisation exercise is complete, we would be able to see how these standard metadata objects flow through the process.

III. Metadata Flow

17. When the metadata objects were standardised at various levels, then these objects could be summarized for each sub process at different levels too. Metadata flowing through the processes as either input or output can be linked back to the GSBPM processes and what sub processes were impacted by what metadata object can be traced. For the sake of example we included a sample of high level metadata objects that could be captured in process flow at GSBPM sub processes. While doing this exercise we have seen that it would be possible to detect the metadata objects as soon as they first appear in a certain sub process; and then re-use in the following sub processes when needed.

Figure 3. Sample of high level metadata objects potentially exist in GSBPM processes



GSBPM	Metadata objects
1. Specify Needs	
1.1.	Documentation on methodology
	Legislation
	Opinions
	Discussions
	Meeting documents
	Statistical unit
	Variables
	Questionnaire
1.2.	List of needs
	User needs
	List of characteristics
	Meeting documents
	Official Statistical Programme (OSP)
	List of decisions to be taken
	Previous experiences
1.3.	Variables
	Statistical unit
	Concepts and definitions
	Coverage
	Method
	Results of evaluations
	Questionnaire
	Legislations
1.4.	Data sources
	Classifications
	Frame
	Questionnaire
1.5.	Work plan
	Budget
	Critical path
	Official Statistical Programme (OSP)
	National data release calendar
	Fieldwork calendar
	Protocols, agreements
	Timeliness
	Legislations

GSBPM	Metadata objects
2. Design	
2.1.	Classifications
	Documentation on methodology
	Tabulation plans
	Meeting documents
	Previous data
	User needs
	Variables
	Statistical unit
	Concepts and definitions
2.2.	Sampling design
	Data from registers
	Sampling frame
	Variables
	Decisions
	Previous experiences
	Coverage
	Method
2.3.	Data collection method
	Questionnaire
	Feedback on questionnaires
	User needs
	Handbooks, training materials
	Documentation on methodology
	Method
	Classifications
	Official correspondance
	Response burden
2.4.	Method
	Classifications
	Meeting documents
	Decisions
	Previous data
	Documentation on methodology
2.5.	Process flow
	Work plan
	Protocols, agreements
	Data collection method

GSBPM	Metadata objects
3. Build	
3.1.	Data collection tools
	Data structure
	Questionnaire
	Frame
	Data cleaning and processing tools
	Data tables
	Data
	Data architecture
	Statistical unit
	Metadata
	Manual
	Method
	Edit rules
	Entity Relationship diagrams
	Logical data model
	System flow charts
	System analysis
	Web interface design
	Know-how
3.2.	User needs
	System and network design
	Metadata
3.3.	Data collection tools
	Data cleaning and processing tools
	System analysis
3.4.	User needs
	System and network design
	Maintenance of the system
	Protocols, agreements
	Trainings

GSBPM	Metadata objects
4. Collect	
4.1.	Sampling design
	Frame
	Variables
	Estimation level
	User needs
	Previous experiences
	Legislations
	Documentation on methodology
	Method
	Coverage
	Classifications
	Questionnaire
	Response burden
4.2.	Documentation on methodology
	Questionnaire
	Data collection tools
	Address list
	Trainings
	Official correspondance
	Brochures
	Mailing groups
	Previous experiences
	Work plan
	Meeting documents
	Work allocation
	Budget
	Number of fieldwork staff
4.3.	Trainings
	Production database
	Raw data
	Verified data
	Data sources
	Administrative data
	Consistency reports
	Data collection tools
	Questionnaire
	Classifications
	Official correspondance
	Fieldwork mission approval
	Fieldwork experience
	Problem logs
	Address list
	Meeting documents
	Follow-up reports
	Protocols, agreements
	Statistical unit
	Legislations
	Telephone and e-mail communications
	Data processing methods
4.4.	Verified data
	Raw data
	Data collection tools
	Production database

GSBPM	Metadata objects
5. Process	
5.1.	Classifications
	Classifications forum
	Raw data
	Flat file
	Verified data
	Data set
	Integrated data
	Data cleaning and processing tools
	Variables
	Metadata
5.2.	Validated data
	Corrected data
	Error reports
	Data cleaning and processing tools
	Data processing methods
	Non-response report
	Variables
	Telephone and e-mail communications
	Production database
5.3.	Outlier list
	Critical values
	Data processing methods
	Validated data
	Corrected data
	Aggregated data
	Frequency tables
	Data standardisation rules
5.4.	Imputation method
	Imputed data for partial non response
5.5.	Validated data
	Clean data
	Imputed data for partial non response
	Weighted data
	Microdata
	Administrative data
	Variables
	Imputation method
	Calculation method
	Factors
	Non-response report
	Statistical unit
	Derived variables

GSBPM	Metadata objects
6. Analyse	
6.1.	User needs
	Feedbacks
	Data sources
	Tables
	Microdata
	Meeting documents
6.2.	Data analysis tools
	Data analysis methods
	Index
	Variables
	Seasonally adjusted data
	Classifications
	Documents on methodology
	Time series data
	Estimates
	Statistical outputs
6.3.	Data quality tools
	Quality indicators
	Data quality reports
	Data comparison reports
	Data analysis reports
6.4.	Data analysis methods
	Time series analysis reports
	Institutional database (data
	Compliance level
	Dissemination tables
	Charts
	Explanations
6.5.	Release level
	Confidentiality rules
	Confidentiality flags
	Confidential data
	Supressed data
	Explanations
6.6.	Dissemination tables
	Metadata
	Interpretations
	Dissemination method

GSBPM	Metadata objects
7. Disseminate	
7.1.	Metadata for statistics
	Metadata for tables
	Bilingual tables
	Bilingual metadata
	Dissemination tool
	National data release calendar
7.2.	Dissemination templates
	Institutional Identity Guide
	Press release preparation handbook
	Press releases
	Dissemination product
	Approved dissemination product
7.3.	Dissemination product
	TurkStat website
	E-mail
7.4.	News on disseminated products
	User queries
	User satisfaction survey
	User feedbacks

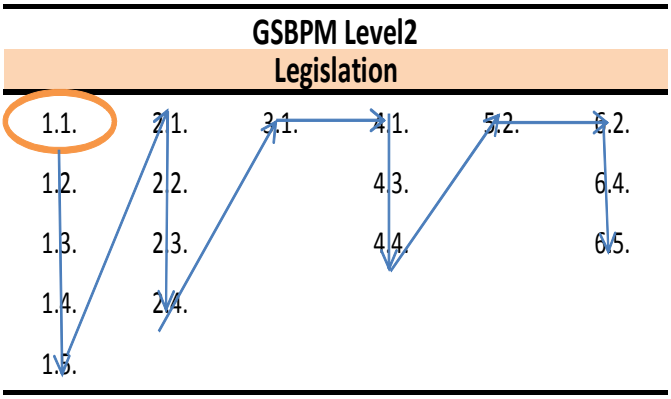
18. If we zoom into the high level metadata objects in GSBPM sub process 1.1. "Determine needs for information" we see that most of the subject matter departments use and study methodology documents, legislations, questionnaires from past surveys or from international sources etc. to identify the needs. The outputs of this sub process may be systematically documented, partially documented or may not be documented at all by the departments depending on their preference. It was seen that, a lot of the data or metadata objects that showed up for the first time at this sub process were either used as is in the following steps or they were transformed and modified into new metadata objects in the following processes.

19. In the following example, legislation appears first in GSBPM sub process 1.1., and then it gets used all the way through the sub process 6.5. Legislation or parts of the legislation is used as input in one way or another in these sub processes. If the information required by legislation could be captured and documented at this first step then the information will be readily available for the users in following steps.

Figure 4a. A sample of high level metadata objects in GSBPM sub process 1.1. "Determine needs for information"

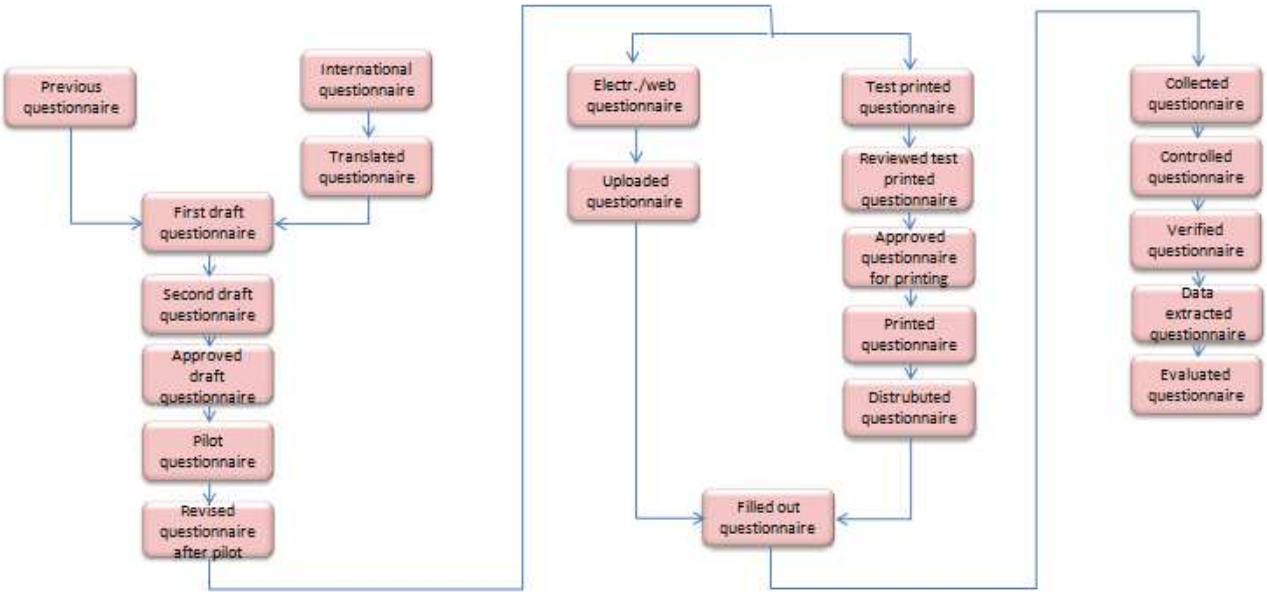
GSBPM	Metadata objects
1. Specify Needs	
1.1.	Documentation on methodology
	Legislation→
	Opinions
	Discussions
	Meeting documents
	Statistical unit
	Variables
	Questionnaire

Figure 4b. Flow of legislation through the GSBPM sub processes



20. Another finding from the process metadata standardisation work is the information on how the metadata objects are modified and transformed into new objects. A good example for the transformed metadata objects is “questionnaire”. By looking at the inputs and outputs of various sub processes, one can see how the questionnaire was transformed along the process, from the sample questionnaire, to draft, pilot, final, collected and verified questionnaire. If we keep track of this metadata we can follow the process path it goes through, and we can predict in which process it becomes an input and in which process it becomes a modified output of the process, simply based on the metadata.

Figure 5. Questionnaire metadata flow through the statistical production process



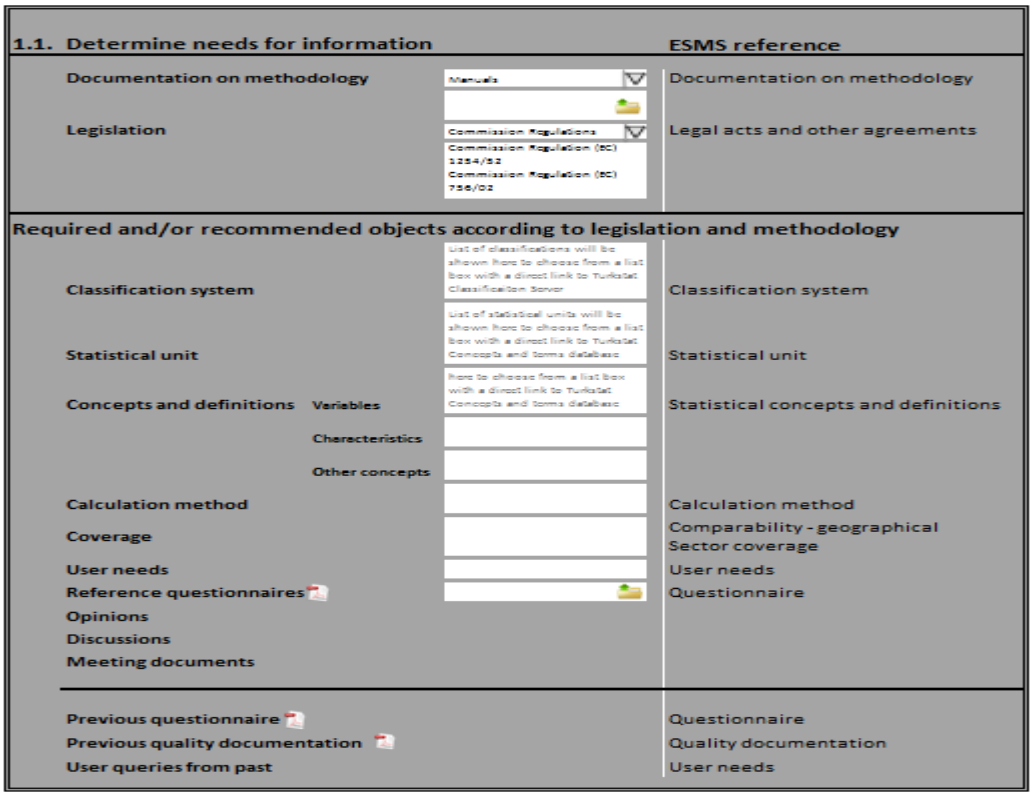
IV. Process Flow Monitoring

21. If we want to carry the findings from this study one step further, and try to design a metadata system based on the metadata flows in the process we could potentially create a metadata capturing system something similar to the one below. Based on the information obtained and the metadata objects determined in sub process 1.1. we can collect the metadata in a format similar to the example shown in Figure 6. Most of the work in sub process 1.1. is concentrated on reviewing the documentation on methodology and legislation, reviewing the user needs etc.

22. To show an example of mapping between the process metadata and dissemination metadata (Figure 6), a standard dissemination metadata structure, ESMS was used. The reason why ESMS was chosen is because it is well known to many NSIs in Europe and TurkStat is sending reference metadata to Eurostat according to ESMS for selected domains and statistics.

23. We could design a system to capture the metadata of the sub process 1.1. in a more structured manner, and link the metadata structure to ESMS structure starting at this point just as a reference, because the dissemination metadata is not final yet. The following scheme is an example to show how the metadata could be captured for a particular study based on methodology documents and legislation and decisions taken in sub process 1.1.

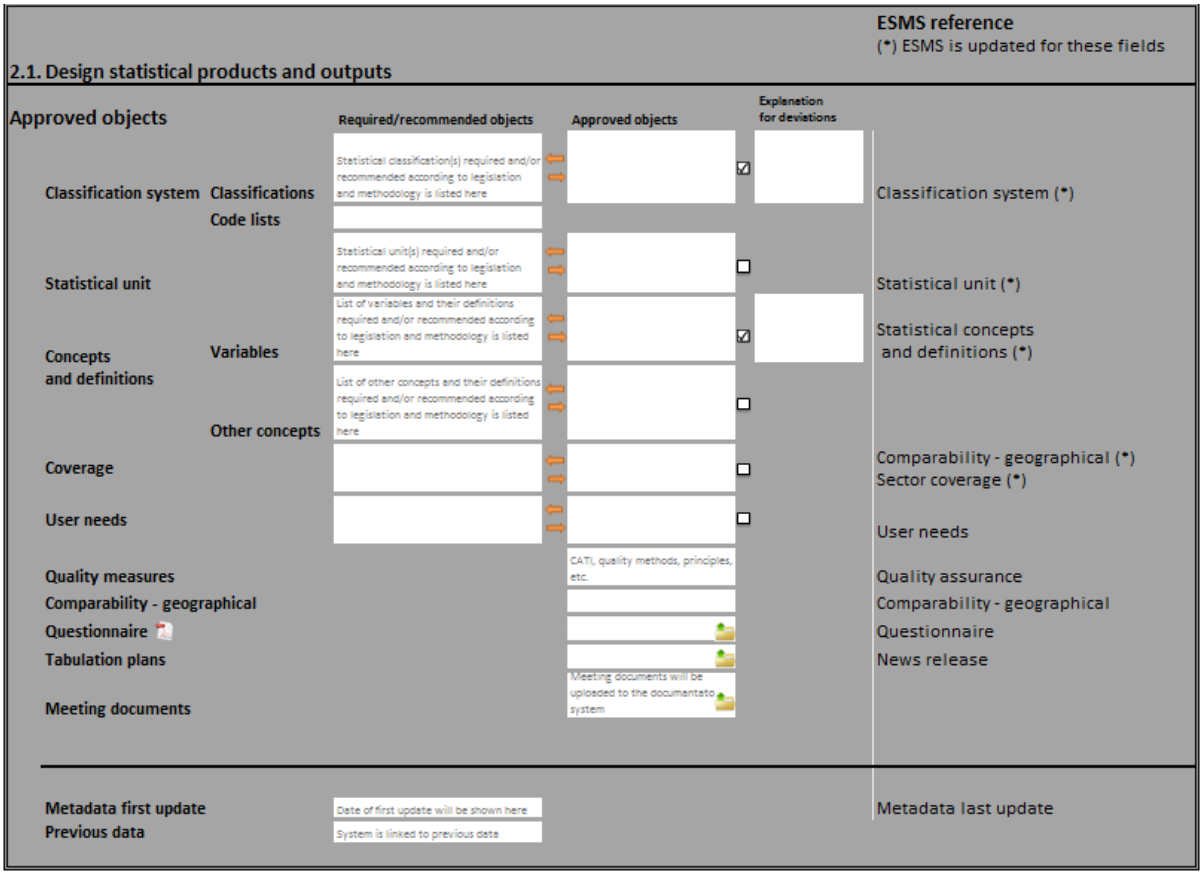
Figure 6. Metadata identified in sub process 1.1. and mapped to ESMS reference metadata structure



24. When we move along the process, we see that the new metadata objects come into the process and the metadata already captured in previous processes would be carried forward. The new objects showing up for the first time need to be captured as the process goes on, and the objects carried forward and involved in a particular process may need to be revised or updated where necessary (Figure 7). If we review the list of metadata objects involved in sub process 2.1.we can see that some of the metadata like classification, statistical unit, list of characteristics etc. were already obtained from process 1.1.

25. In process 1.1. the metadata objects were either recommended or required based on certain legislations or methodology, however in process 2.1. some of these requirements may not be met based on the decisions taken along the process. Therefore, another field was needed to capture the approved metadata and objects. So, the reference metadata need to be updated here based on what has changed. Fields shown with a (*) beside their names under the ESMS reference section are expected to be updated in process 2.1. (Figure 7). If a metadata report based on ESMS was going to be produced, then it could have been produced after this updating was done in the starred fields.

Figure 7. Metadata identified in sub process 2.1. and metadata carried forward from previous processes to be updated



26. What this exercise shows is that when the metadata is identified in a process it needs to be captured and documented in a structured way. This will allow organizations to watch the process and get the metadata in a more accurate way and also to re-use the metadata where and when it is needed.

V. Conclusion

27. The discussion on process flow monitoring in section IV is rather theoretical at this point. TurkStat is working on how the process metadata could be captured, what kind of systems could be developed to store the metadata and how this metadata could be used to monitor the processes.

28. Metadata management is one of the overarching processes in GSBPM. The key here is to identify, standardise and capture the metadata as early as possible in the process and keep track of the process using the metadata. If there was a system to hold the process metadata information in a structured manner, then it would have been possible to extract and re-use these metadata wherever needed, i.e. in monitoring the process, in quality management or in dissemination metadata systems.

VI. References

Consoli A., Götzfried A., Linden H., Rychel B. (2013), "Better documenting statistical business processes: the Euro Process Metadata Structure", Work Session on Statistical Metadata (METIS), Geneva, 6-8 May 2014

Özkan D., Dorsan N., Eminkahyagil G.,(2013). "Implementation of GSBPM and metadata studies in the Turkish Statistical Institute", Work Session on Statistical Metadata (METIS), Geneva, 6-8 May 2014 www.unece.org/stats/documents/2013.05.metis.html

Strategic vision of the HLG (2013). (High-Level Group for the Modernisation of Statistical Production and Services), http://www1.unece.org/stat/platform/display/hlgbas/Strategic+vision+of+the+HLG

UNECE Secretariat (2013). "Generic Statistical Business Process Model: Version 5.0, December 2013", Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS)

UNECE Secretariat (2013). "Generic Statistical Information Model (GSIM): Version 1.1, December 2013" http://www1.unece.org/stat/platform/display/gsim/Generic+Statistical+Information+Model