Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

Modern, process oriented and metadata driven statistical production

Prepared by Anna Długosz, Statistics Poland

I. Introduction

1. The traditional approach to the organisation of surveys in official statistics is domain-based. Each statistical survey is carried out from the beginning to the end by an authorial unit. Surveys are carried out in different ways, as a result of the implementation of original and specific methodologies of the statistical process, often with a significant blurring of the boundaries between individual stages of statistical production or individual understanding of what should be done within a given stage.
2. Collected data are poorly shared between different surveys - data collection for specific surveys dominate. And tasks connected with data processing, validation and editing are duplicated. These processes are inefficient, unproductive and cost-ineffective. Metadata is not used to manage data processing. This approach is neither productive nor efficient.
3. The existing drawbacks and problems of the current way of conducting statistical surveys result in the practice of generating and maintaining dispersed, autonomous, inconsistent and ineffective solutions for IT support of statistical surveys and the absence of a process-oriented and meta-data driven approach to statistical production. Identified disadvantages (problems) indicate a lack of a coherent approach to modelling and designing IT solutions supporting statistical production. A modern approach to statistical production is based on the use of good, worldwide accepted solutions, which are connected with the developed international standards.

II. Modern, process oriented and metadata driven statistical production

A. The Model of the Statistical Production Process

4. Polish statistics as a starting point for the idea of improving the process of statistical production has adopted the Generic Statistical Business Process Model – the GSBPM. The Model of the Statistical Production Process (based on the GSBPM) is a holistic approach to the production process, which allowed to separate systems and applications necessary for the implementation of the processes, from the *specify needs* phase to the *evaluate* phase.
5. There is not a lot of differences between GSBPM v. 5.1 and MPPS. We added some sub-processes connected with our yearly statistical programme, plans and schedules. They have been separated because they are very important in our organization. The annual programme defines the scope of the output information to be provided and imposes obligations on respondents and data providers to provide data. Other added sub-processes depict a usage of geospatial data which are omitted in the traditional GSBPM model. We also took into account sub-processes connected with assessment,

verifying and quality indicators. These sub-processes are added to assess the quality and take possible corrective actions. It is very important because positive verification of products (outputs) of previous phase allows us going to the next phase. Control at this stage is crucial, independent of the controls carried out at the outputs of individual sub-processes. Its outcome is connected with the plans verification.

6. The implementation of the MPPS is essential to modernise (build) IT solutions supporting the realisation of public statistics tasks. First of all, the activities will focus on the processes concerning the four main production phases: *collect*, *process*, *analyse* and *disseminate*. The next move is to create a reference architecture framework based on a process-oriented model of statistical production and create a performant metainformation system.

7. The presented solution is a logical architecture. We take into account that there may be a difference between a logical architecture and a physical architecture. We do not expect that logical architecture will map one to one to physical architecture.

8. Vital elements of the new architecture are data repositories for all phases, which will store the current "states" (entities) of the processed data and a separate metadata repository. This approach facilitates access to statistical data for subsequent "states". Subsequent "states" of data result from the successive steps of their processing in the successive phases of the statistical production process.

B. Data repositories

9. In the phase *collect*, data is collected and saved in the Raw Data Repository (SDS) to make them available for further processing in a production environment. Data collection is preceded by the determination of the data collection method, the creation of a list for a survey based on the content of the Registry Data Repository (SJS) and the selection of a random sample in the case of a representative survey. Additional two geocoding sub-processes use the contents of the Geospatial Data Repository (SDG). Also, an additional sub-process v*erification of plans* should be performed to assess the quality and take possible corrective actions. Positive verification of plans and conducting geocoding of collected unit data will finalize the phase allowing to go to the phase p*rocess*.

10. The main effect of the phase *collect* is a set of unit data, correctly verified and collected in the Raw Data Repository (SDS). It should be emphasized that the phase *collect* is the first of the four phases of the actual production cycle. It is crucial from completeness and quality of input data necessary to calculate final products expected by public statistics users. For this reason, the implementation of each of the sub-processes should be consistent with the assumed qualitative indicators, located in the Metadata Repository. After completing the sub-process, the effects of its implementation should be described with appropriate metadata and stored in the Metadata Repository. Achieved values of qualitative indicators should decide on admission to the next phase of the production cycle or in other case to start procedures for improving the quality of collected data.

11. In the next phase, *process*, the data are transformed to prepare them for statistical analyses. The unit data collected in the phase *collect*, stored in the Raw Data Repository (SDS), is subject to transformation following pre-designed rules that are stored in the Metadata Repository. Then the data is a subject to statistical classification and coding under established dictionaries and classifiers. Then the data is reviewed for verification and validation and later followed by their final transformation to a set of statistical unit data. At the same time, it should be possible to calculate new complex variables and statistical units and to supplement missing data, i.e. imputation, as well as final anonymisation or pseudonymisation of identifiable unit data. In the case of sample surveys based on a random sample, weights are determined, the results are elaborated on the whole population survey, and their necessary calibration is carried out and the essential aggregates calculated. After positive verification of qualitative indicators for this phase, final data sets are created, which should be loaded into the Operational Data Repository (SDO). For statistical surveys carried out regularly, this phase is carried

out each time. Sub-processes of this phase may take place as many times as necessary (with the possible exclusion of sub-process c*alculation of weights*, which is specific only for representative surveys).

12. The phase *process* and the next *analyse* are usually implemented sequentially.

13. In some cases, however, it is allowed to implement both phases, i.e. *process* and *analyse*, iteratively or in parallel. During analytical works there may be an urgent need to provide additional data, which may mean that another processing cycle will be needed for all or part of the collected data, which may also lead to a restart of phase *collect*. It is allowed that the *process* and *analyse* phases may start before the end of the *collect* phase, which in justified cases is necessary for the preparation of preliminary results. Making initial results available is very important for some users of statistical data, expecting quick statistics to make management or investment decisions. All effects of the sub-processes of this phase should be described in the metadata stored in the Metadata Repository. It is particularly important to include the metainformation regarding the quality of data gathered in the SDO, i.e. the degree of data processing, compliance of definitions of concepts and classification, the level of an imputation and deduplication (removal of unnecessary repetitions) and the precision of the final results.

14. In the *analysis* phase, analytical statistical products are created and thoroughly checked. They are the entrance to the next sub-process in which the resulting statistical information is prepared and developed. The phase also includes the preparation of necessary metainformation and additional explanatory information in the form of comments, descriptions or methodological notes accompanying statistical publications. Before making results available, it is necessary to ensure that the analytical products comply with the data received and meeting the needs of users. This phase begins with the data downloading from the Operational Data Repository (SDO). During this phase is conducted preparation of preliminary results, and then their validation following the rules stored in the Metadata Repository.

15. Before the final verification of the resulting data, the fundamental spatial analyses should be carried out and prepared for the presentation on the maps. The resulting data supplemented with spatial analyses should be subject to final review and explanation. Such prepared and checked data before their publication is checked for statistical confidentiality, so that it is not possible to disclose the identity of subjects participating in the statistical survey.

16. The *disseminate* phase includes all activities related to rules and mechanisms of sharing and preparation of a group of products to make them available in various forms as well as in multiple channels of sharing, both traditional and electronic.

17. Activities connected with data sharing include a support of users. This support takes place through proper communication and promotion activities. An additional sub-process of satisfaction surveys was created to facilitate the assessment of whether and to what extent the designed and built statistical production process produces products meet the expectations of final recipients. The acquired knowledge, following the TQM methodology, is crucial for improving satisfaction in the next cycle of the research and allows to assess the relevance of identifying the needs of users in phase *specify needs* and, if necessary, preparing an essential improvement plan. All effects of the sub-processes of this phase that are important for the quality of the products being made available should be described by the metadata stored in the Metadata Repository. It is particularly important to include in the Metadata Repository the metadata about the quality of products placed in the SDP, i.e. the degree of their compliance with the designed sets of publication products and the degree of consumer satisfaction of the resulting statistical information.

C. The metainformation system and the metadata repository

18. To improve all steps of the statistical production process and to make data validation as well as editing more efficient, introduction of process-oriented and metadata driven approach is essential.

19. The metainformation system with the metadata repository is necessary to monitor and manage the production process. The system is to control the overall statistical production process. Data sets that

are not provided with the appropriate metadata have a much lower value, and may even become completely useless. In a quest to build more efficient processes, one of our top priorities today is to create and consume rich and powerful metadata, which is basically "data about data". In addition to the metadata describing the data set and data itself, information contained therein should include other metadata, for example metadata connected with data editing. But the value of metadata is not to be overrated.

20. The statistical production process should be based on a well-designed metainformatiom system. It helps to improve the quality and efficiency of a process and quality of statistical products. Without the ability to manage metadata efficiently we lose time and productivity.

21. The metainformation system needs metadata repository that collects metadata as well as communicates it to other systems. The metadata repository contains metadata describing data and processes as well as metadata generated in the statistical production process. It should enable the use of metadata to support the statistical production as well as to monitor and drive processes.

22. Thus, the primary role of the metainformation system and the metadata repository is to collect and make available metadata for the statistical process. Ensuring functionality of metadata collection in the system requires preparation of appropriate metadata storage structures as well as implementation of a procedure for a life cycle management of stored metadata. The functionality of making available metadata requires the development of metadata exchange interfaces, in particular ensuring the possibility of automatic access.

23. The system must also provide the statistical production process management functionality, i.e. it should enable the processes to be controlled by making available the relevant metadata to control the processing operations, and enable their states to be monitored on the basis of feedback from subsequent data processing steps.

D. Data editing

24. Data editing is to detect and treat errors and missing values in a data set. In this sense, when we talk about data editing we also mean data validation. Editing is all the activities on data that we have to do in case of incorrect, incomplete, unreliable or outdated data. Editing data means also reviewing of data for outliers – data which are totally larger or smaller than other data. We can say that data editing is a process of improving data and data set.

25. Data editing's target is to improve the quality, accuracy and adequacy of the data and make it useful. This process is very important and crucial for quality of statistical outputs. But on the other hand this process is very difficult, time and labour consuming.

26. One of the ways to make this process more efficient is to automate it. The goal of automatic editing is to accurately detect and treat errors and missing values in a data file in a fully or half-fully automated manner, i.e., without or less human intervention.

27. But in order to achieve significant improvements in the efficiency and quality of data editing processes, a comprehensive modernisation of the entire statistical production process is necessary.

E. Use of Administrative Data Sources

28. For processing large data sets we use the dedicated IT environment - the Operational Database of Microdata.

29. The Operational Database of Microdata has the ability to communicate with other IT systems of official statistics, thanks to which it is possible to integrate data from statistical surveys, statistical operations with data from administrative sources.

30. It communicates with the metainformation subsystem, where all information about the processed data sets is placed.

31. It is administered and managed by the Managment Console.
32. The converted data is exported to the Microdata Analytical Database, in order to compile results data and create analyses.
33. The OBM operates three interconnected systems, which are used for systematic work on data from administrative sources.
34. The systems that make up the processing environment: System Processing of Administrative Data, Variable Quality System, Statistical Operations System.
35. The System Processing of Administrative Data is used to undergo processes of improving data quality.
36. The Variable Quality System allows to create quality reports and view variables from administrative registers.
37. The Statistical Operations System enables data integration and the creation of statistical research results, which are grouped according to belonging to the relevant variable domain.