Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Management of data processing processes using metadata

Prepared by Regional Statistical Office in Olsztyn, Poland

I. Introduction

1. Conducting statistical surveys requires performing a number of specific steps beginning from collecting data, through processing them to preparing results. Each step is a process. The processes are dependent on each other and must be performed in a specific sequence.
2. The large number of such processes and their interconnectedness leads to difficulties in determining what processing step we are currently in and what the next step should be. The solution is to describe these dependencies in the form of metadata and to manage them appropriately.
3. The MSPP (Process Control Module), part of the SPDS data processing system, is a software developed in Statistical Office in Olsztyn (Poland) that allows you to control data processing processes based on their description in metadata.
4. MSPP provides a tool for designing data flows within statistical surveys and between them. It allows to describe data sets, processing processes and dependencies between them.
5. MSPP also provides a tool for maintaining data integrity and running processes. This tool enables:
   (a) management of data processing processes
   (b) running processes (SQL Server stored procedures, SQL Server Integration Services ETL packages)
   (c) data flow tracking
   (d) informing selected users about the completion of processing of selected processes
   (e) checking the status of data sets
   (f) data browsing

II. Metadata structure

A. Basic objects of MSPP

6. The basic metadata objects used by MSPP are shown in the FIGURE 1.

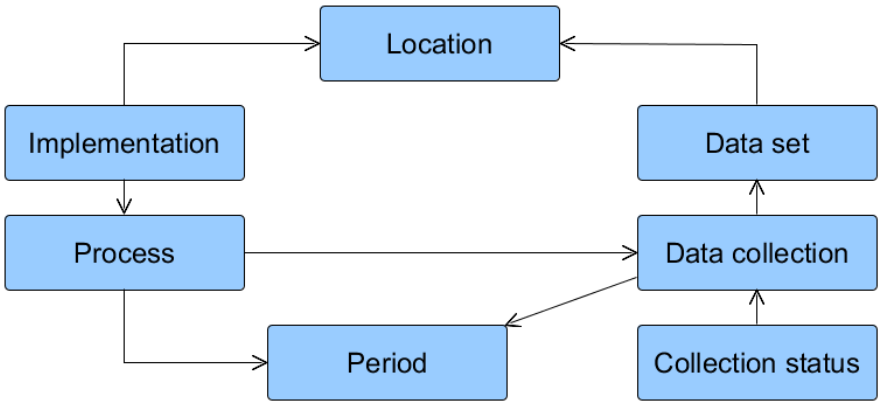

Figure 1

(a) Dataset - represents a single data set (table in the MS SQL database or file (XML, DBF, XLS, XLSX, TXT, CSV)

(b) Period - the time interval assigned to process implementations and data sets.

(c) Data collection - represents a pair: a data set with a selected period. A single data set can contain several data collections.

Dataset	Description	Assigned period / periods	Data collection
Set 1	Includes data for June 2013	June 2013	Collection 1
Set 2	Includes data for first half of 2012	first half of 2012	Collection 2
Set 3	Includes data for the first and second quarters of 2014	First quarter of 2014	Collection 3

(d) Process implementation - indicates an object that performs a certain process of processing statistical data. The implementation can be a MS SQL stored procedure, SSIS package or any other object performing physical data processing (eg external application, expert method).

(e) Process - is a representative of the process implementation for the selected data processing period. It is a single task of data flow / processing. Unlike its implementation, the process always applies to one period.

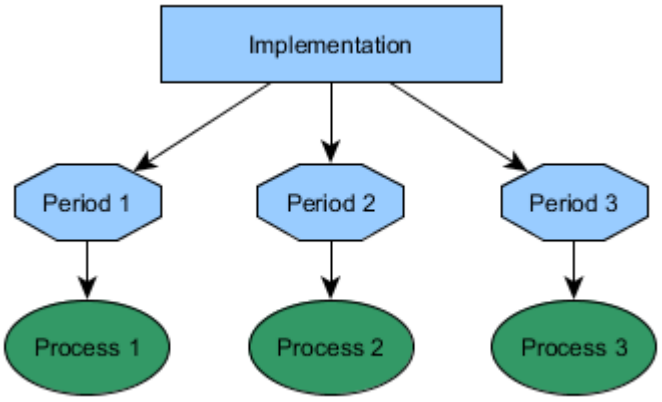

Figure 2

7. The state of the processed data is described by the statuses. Each data collection has a group of statuses.

8. Status [Timeliness of data] is one of the most important. It is involved in the mechanism of tracking data flows providing blockade of the process until the data it uses is not changed. This prevents running processes that perform operations on outdated data. As a result, processes are only performed when it is necessary.

9. The status [Timeliness of data] can take values:

Status [Timeliness of data] values	Description
Current	The data contained in the data collection is current and can be used to prepare other data collections
Out of date	Data contained in the data collection is NOT current and can NOT be used to prepare other data collections
Probably out of date	The data contained in the data collection is LIKELY NOT current and CAN NOT be used to prepare other data collections.

B. Relationships between data collections and processes

10. The data collection and process are treated as individual objects, between which dependencies are determined. The data collection can be a source or target collection for a specific process. The process may be associated with several source collections and several target collections. The diagram of dependencies between objects for a single process is presented in FIGURE 3. The implementation of the process uses three source collections and generates two data collections. These collections have certain statuses. This process can only be started if all data collections associated with it have statuses that meet the requirements defined in the metadata. Each status of the source and target collections described in the requirements must assume one of the specified values. The process during execution can change statuses of the source and target collections associated with it directly.

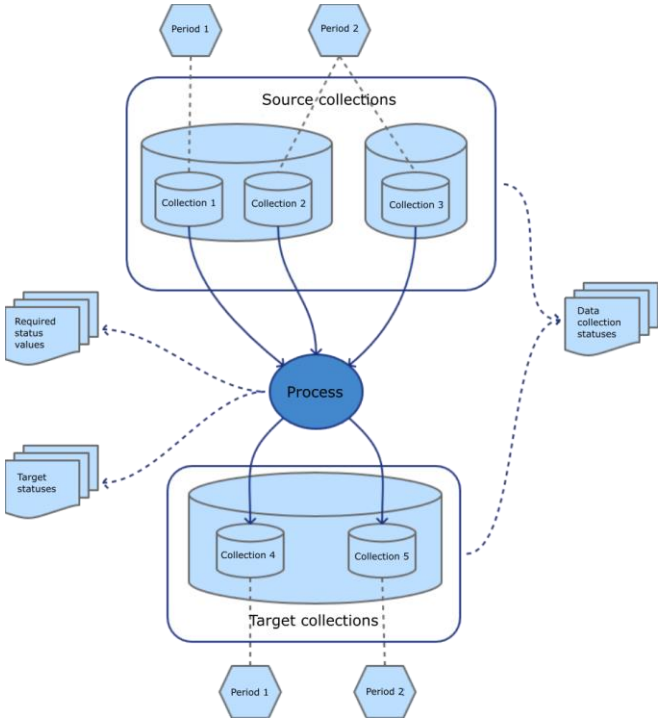

Figure 3

11. On the basis of requirements and statuses, the order of processes and their dependencies are determined.

12. This mechanism allows you to:

(a) manage the order of executing processes

(b) block the possibility of running multiple instances of the process at the same time

(c) manage the timeliness of the data stored in the collections

(d) track other important data properties (eg quality, availability, completeness)

13. FIGURE 4 is an example of ordering three processes. Their order is not defined explicitly but it is clear. Process 2 can be started only when the status [Timeliness of data] of the "Stage1" collection is "Current". This value can only be set by executing Process 1. In the same way, Process 3 is aligned with the process 2.

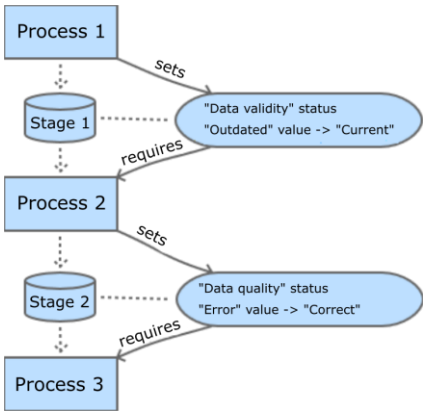

Figure 4

C. The sequence of starting the process

14. The following FIGURE 5 shows the sequence of activities performed by the MSPP as a part of a single process.

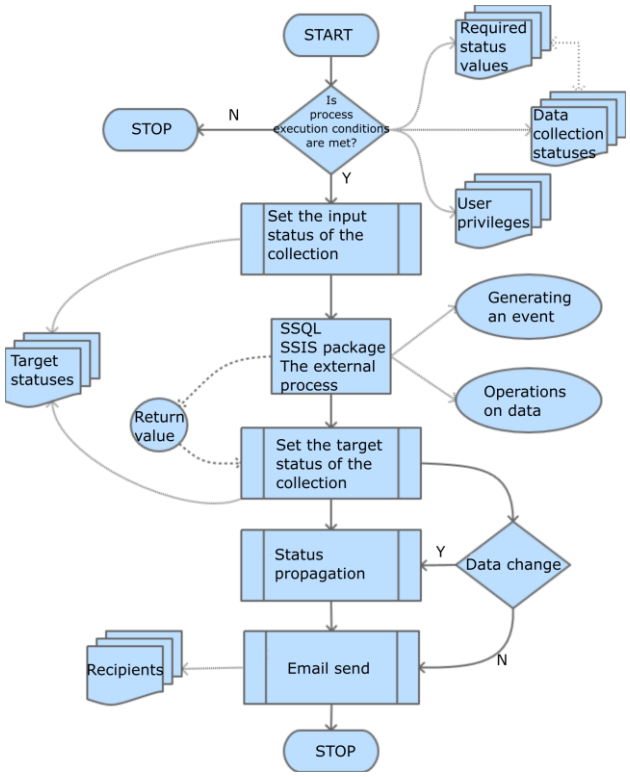

Figure 5

15. This sequence has the following steps:

(a) Checking whether the conditions of the process are met. Execution is stopped if one of the conditions is not met:

- the user has sufficient rights to run the given process
- all process requirements are met, the process is not blocked / started
- the process is implemented and ready for execution

(b) Application of input data status settings. Before the implementation of the process is started, the source and target collections receive the indicated statuses specified by the designer. Settings that impose write and read locks on the execution of the process are initiated automatically.

(c) Launching the process implementation with the period as a parameter. The implementation performs physical operations on data. It uses source collections and generates target collections. All errors and messages generated at this time are saved as a log in metadata.

(d) Set the output status of the data collection. This stage takes place after the operation has been properly completed on the data by the implementation and return of the resulting value. For each possible result value there is a separate set of output statuses defined. An example of the use of an output parameter is the implementation of completeness control of the data collection. If the implementation identifies that the data being processed is incomplete, it can return a value other than 0, which allows you to apply a separate set of statuses, including the status value "Data completeness" set to "Incomplete".

(e) Checking changes in the data collection. If the target collection has changed as a result of the execution of the process implementation, the "Status propagation" mechanism is activated, which changes the status of all dependent data collections to [Deprecated]. Thanks to this, it is possible to run subsequent processes in cases when it is necessary, for example in case of necessity to restart the process due to data correction in the source collection.

(f) Sending an email informing about the completion of the process. A short report on the execution of the process is sent to all users added to the distribution list.

III. MSPP architecture

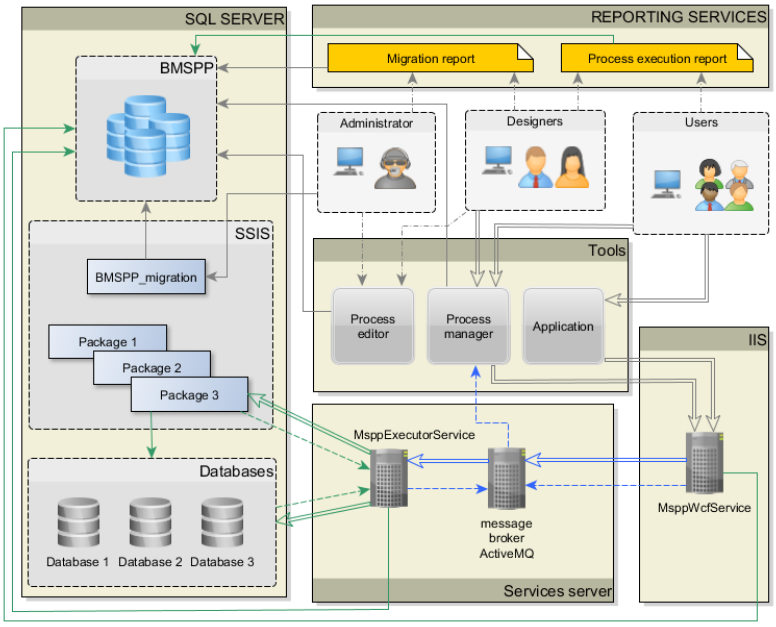

Figure 6

16. The structure of MSPP is presented in FIGURE 6. MSPP consists of the following components:

(a) MSPP Database (BMSPP) - stores MSPP metadata.

(b) MSPP tools

- Process editor - used to describe the metadata of processes, is the BMSPP user interface.

- Process control - used to review, execute and monitor data processing processes
- Application - any application that uses MSPP libraries that allows to view, execute and monitor data processing processes

(c) BMSPP administrative tools

- BMSPP_migration - SSIS package for data migration between individual BMSPP instances.

(d) MsppWcfService - interface between client applications and executive mechanisms for MSPP processes. It allows you to download metadata about individual objects described in the MSPP and order the execution of processes.

(e) MsppExecutorService - manages the execution of individual implementations of MSPP processes.

(f) ActiveMQ message broker service - allows to monitor the processes performed. Transfers messages (progress, information, warnings, errors) from client processes to MSPP. Mediates in ordering the execution of the process.

(g) Reporting Services Reports - Report tool

IV. Usage example

17. The MSPP in action is presented in the example below. All images come from the "Process control" application.

18. The following FIGURE 7 shows the initial state of data processing, in which most data collections have the status "Out of date" which means that the processes that accrue these collections have not yet been launched. Only one collection [Bilans_ludności] for the fourth quarter of 2013 has the status "Current" and is the source for the process [Przygotuj bilans ludności] for the first quarter of 2014. The only processes allowed to run are [Urodzenia – Przygotuj dane z SIB], [Urodzenia – Przygotuj dane z SIB] and [Migracje – Przygotuj dane z SIB]. Other processes are not allowed to run due to unfulfilled requirements.

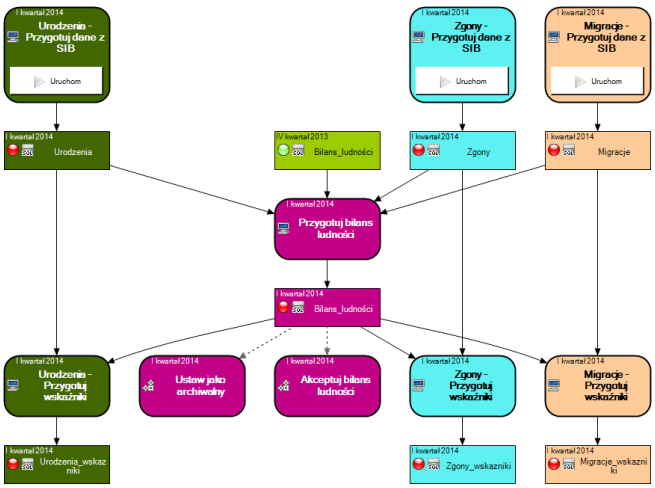

Figure 7

19. After starting and completing the processes [Urodzenia – Przygotuj dane z SIB] and [Zgony – Przygotuj dane z SIB] for the first quarter of 2014, data collections [Urodzenia] and [Zgony] will be given the status "Current". The [Przygotuj bilans ludności] process is still not available because the data collection [Migracje] has the status "Out of date" (requirements for this process are not met).

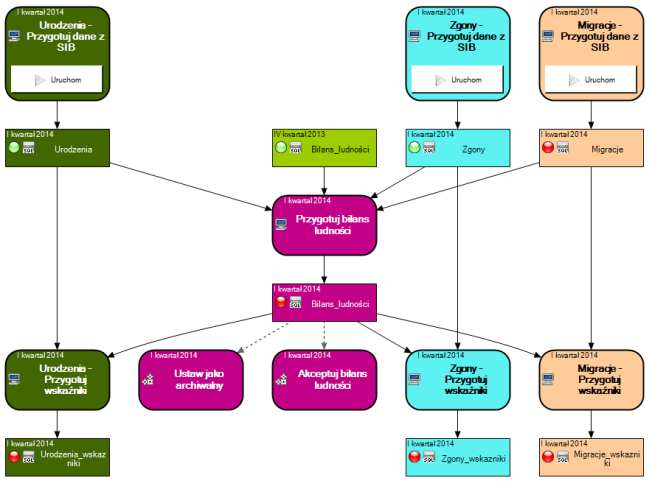

Figure 8

20. Processes [Urodzenia – Przygotuj dane z SIB], [Zgony – Przygotuj dane z SIB] and [Migracje – Przygotuj dane z SIB] have been defined as long-term external processes. Processes of this type represent a series of operations which are executed / performed by MSPP directly and may take a long time (where appropriate, up to several weeks). In this case, the MSPP allows the user to inform the system about the start of the process and separately about the completion of the process. After running the long-term process, the diagram presents a progress bar representing the activity of this process and the "Finish" button, thanks to which the user can inform the system about the process's completion. After starting the [Migracje – Przygotuj dane z SIB] process, the diagram will look as follows:

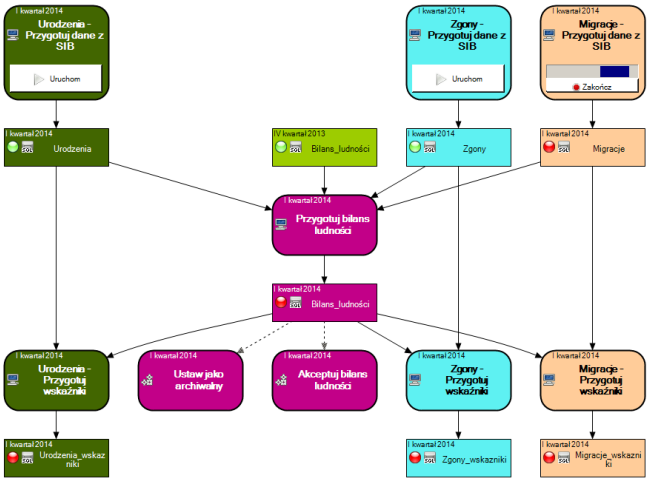

Figure 9

21. Completing the process [Migracje – Przygotuj dane z SIB] will change the status of the [Migracje] collection to "Current", thanks to which the [Przygotuj bilans ludności] process will be made available for launch.

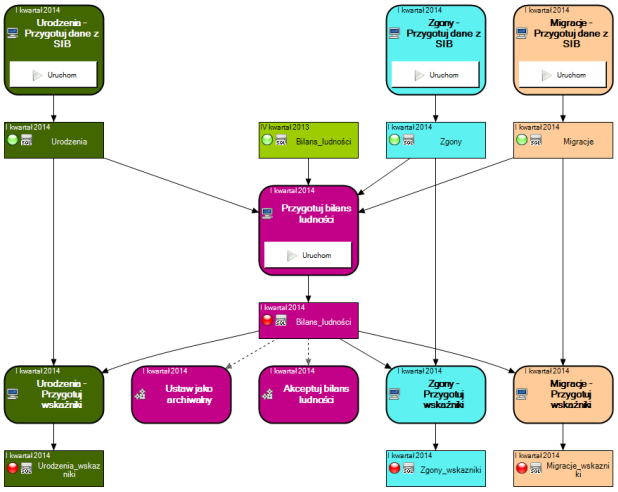

Figure 10

22. After completing the [Przygotuj bilans ludności] process, the status of the [Bilans_ludności] collection will be updated.

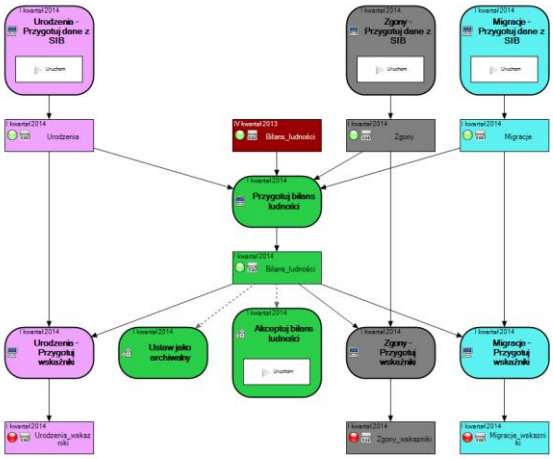

Figure 11

23. The process [Akceptuj bilans ludności] is a virtual process that changes only the status value "Data acceptance status" of the data collection [Bilans_ludności] to "Accepted". Processes [Urodzenia – przygotuj wskaźniki], [Zgony – przygotuj wskaźniki] and [Migracje – przygotuj wskaźniki] can be made available to run only after accepting [Bilans_ludności]. These processes can only be started after the user accepts the data.

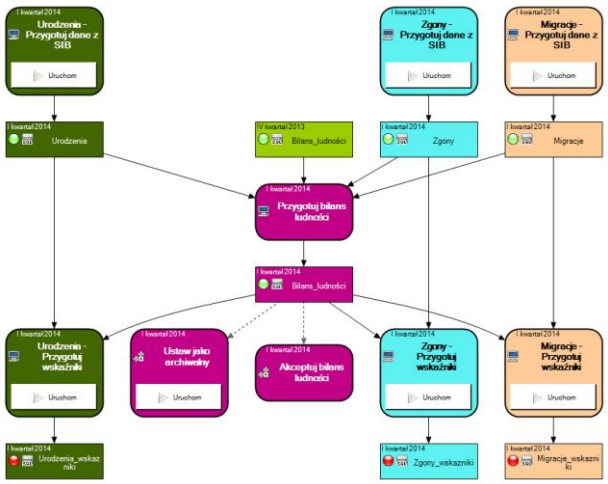

Figure 12

24. The following FIGURE 13 shows the state after the completion of all processes that calculate the indicators. All collections have the status “Data validity” set to "Current".

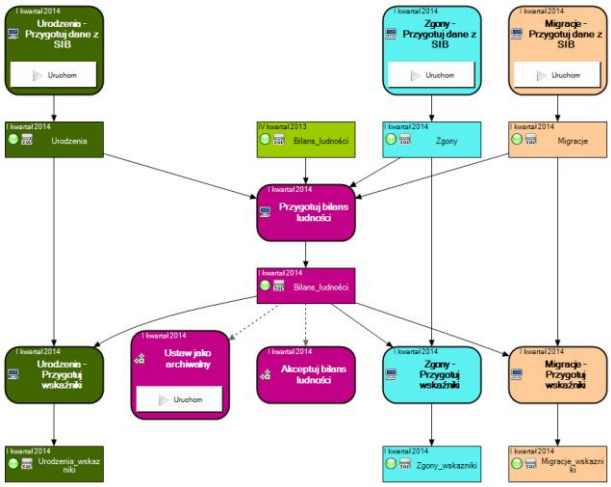

Figure 13

25. Restarting the [Zgony – Przygotuj dane z SIB] process will cause the data in the [Zgony] collection for the first quarter of 2014 to be changed. The mechanism for the propagation of the status of "Data validity" will be automatically launched, and therefore all dependent data collections [Bilans_ludności] for the first quarter of 2014, [Urodzenia_wskaźniki] for the first quarter of 2014, [Migracje_wskaźniki] for the first quarter of 2014 and [Zgony_wskaźniki] for the first quarter of 2014 will be marked as out of date.

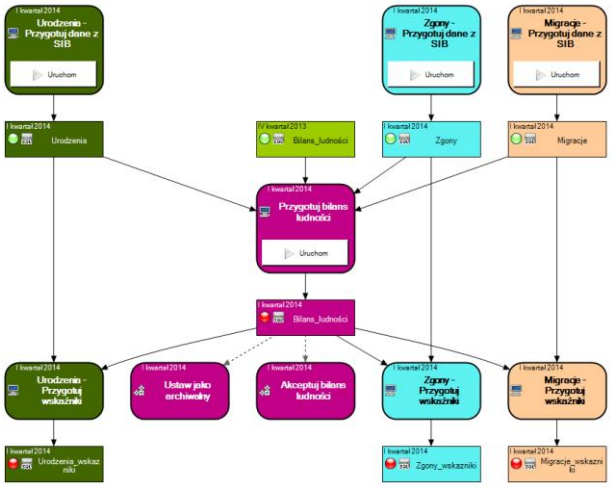

Figure 14

26. Changing the status of "Data validity" of these data collections to "Outdated" means that processes [Przygotuj bilans ludności], [Akceptuj bilans ludności], [Urodzenia – przygotuj wskaźniki], [Migracje – przygotuj wskaźniki] and [Zgony – przygotuj wskaźniki] must be restarted. After starting and completing the processes [Przygotuj bilans ludności] and [Akceptuj bilans ludności] we will get the status:

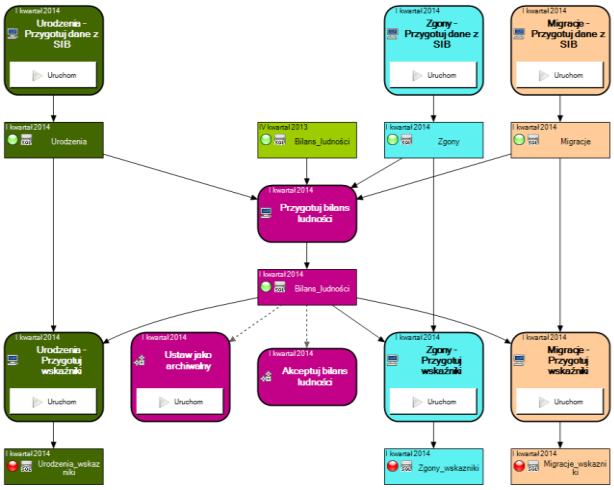

Figure 15

27. Data collections that have the status value "Data acceptance status" set to "Archival" cannot be changed, which means that they do not change the value of their status "Data validity" and do not transfer propagation to the dependent collection. The [Ustaw jako archiwalny] process is the second virtual process in this example. This process sets the status "Data acceptance status" of the [Bilans_ludności] collection to "Archival", so the data in this collection will be protected from changes. Archival collections are marked on the diagram with a yellow flag. After running the [Ustaw jako archiwalny] process, the diagram representing the processing state will look like this:

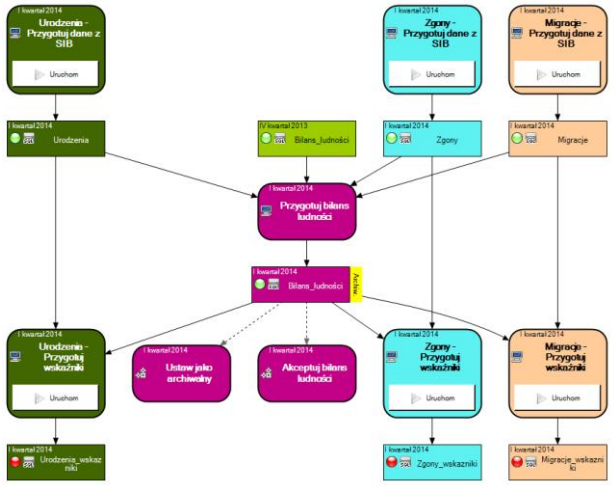

Figure 16

28. The following FIGURE 17 shows the state after the completion of all processes that calculate the indicators. All collections have the status [Data validity] set to "Current".

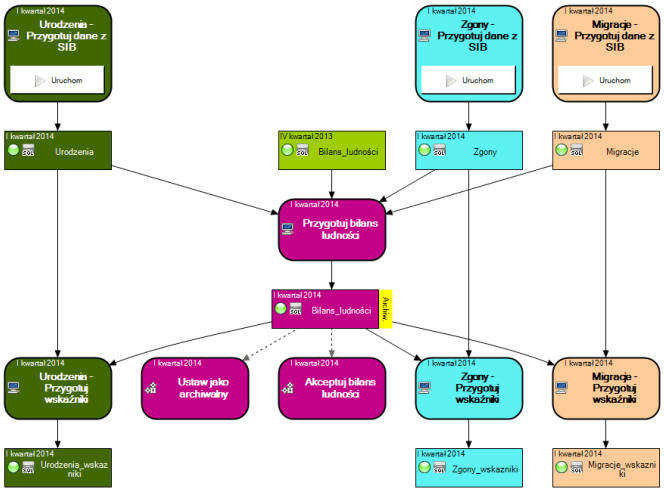

Figure 17

29. Restarting the process [Migracje – Przygotuj dane z SIB] will cause the data in the [Migracje] collection for the first quarter of 2014 to be changed. The mechanism for propagating the status of "Data validity" will be automatically launched. Since the [Bilans_ludności] collection for the first quarter of 2014 has the status "Archival", the status "Data validity" of the data collection [Bilans_ludności] for the first quarter of 2014, [Urodzenia_wskaźniki] for the first quarter of 2014 and [Zgony_wskaźniki] for the first quarter of 2014 will not be changed. Only the data collection [Migracje_wskaźniki] for the first quarter of 2014 will be marked as "Outdated".

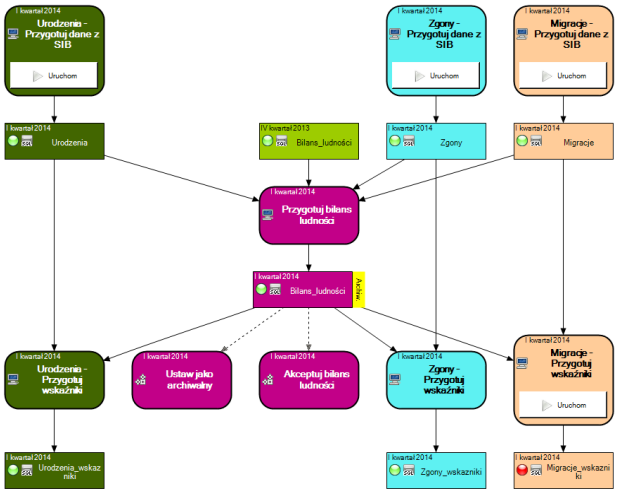

Figure 18

30. Processes [Urodzenia – Przygotuj dane z SIB], [Zgony – Przygotuj dane z SIB] and [Migracje– Przygotuj dane z SIB] will not affect the data collection [Bilans_ludności], which has the status of "Archival" and "Current". The [Przygotuj bilans ludności] process will not be available for launch.

V. Conclusion

31. The software developed in the US Olsztyn allows you to manage processes data through metadata. MSPP has a wide range of applications from simple processing to management of complex statistical surveys.

32. MSPP will be used for data processing in the Polish agricultural census in 2020.