Planning for an increased use of administrative data in censuses 2021 and beyond, with particular focus on the production of migration statistics

Dominik Rozkrut – President, Central Statistical Office of Poland.
Janusz Dygaszewicz – Director, Programming and Coordination of Statistical Survey Department.
Dorota Szałtys – Deputy Director, Demographic Surveys and Labour Market Department.
Gabriela Nowakowska – Head of Unit, Regional Statistical Office in Warsaw.

This paper presents an overview of the work carried out in the Central Statistical Office of Poland on the 2021 National Census of Population and Housing related to the use of administrative registers as data sources, including migration statistics. It shows the context of the implementation of the 2021 Census, the overall vision of the implementation of the census, as well as the possibilities and limitations that must be taken into account when choosing the sources and as regards data quality.

1. Introduction

To meet the challenges posed by the dynamic development of information and communication technologies, as well as recognizing the needs of users for more frequent, actual data on the lowest level of the spatial aggregation, the Central Statistical Office of Poland (CSO) carries out work on a wider use of administrative registers in statistical surveys, including the 2021 National Census of Population and Housing (2021 Census) as well as the censuses planned after the year 2021.

International mobility is recently an issue of particular interest. Timely, good quality and comprehensive information on migration and migrants is extremely important for the proper management of the scale and dynamics of migration flows. The population and housing censuses carried out usually every 10 years seem to be – compared to such a dynamic phenomenon as migration – an imperfect source of information, as the acquired data on migration is already outdated at the time of its publication. Hence, a number of countries, including Poland, are focusing on finding sources that contain up-to-date and systematically available information. Administrative data sets have the potential to become such a source. For obvious reasons (given primarily labor intensity), the most desirable situation for statisticians would be an existence of a single register in which the facts that allow to elaborate full social-demographic and economic characteristics of

a migrating person as well as conducting extensive migration analysis would be recorded for each migrating person (both in internal and foreign movement). However, paradoxically, a multitude of administrative sources containing information on migration can be considered as a strength. Integration of many heterogeneous sources can extend not only the coverage of the data in terms of content but also some qualitative aspects of the set. This is particularly important in the context of the inflow of short-term or seasonal immigrants, who are often registered in domain registers or records and not in reference registers (such as population register, taxpayer records, social security, etc.). Such situation occurs in Poland mainly in the case of citizens of Belarus, Moldova, the Russian Federation, Ukraine and the Republic of Armenia who may take up employment without the necessity of obtaining a work permit in any industry for a period not exceeding 6 months within 12 subsequent months.

In the following part of the paper, actions taken by the CSO regarding the use of systems and administrative registers as sources of information for 2021 Census will be discussed including surveys on migration.

1.1. Requirements for 2021 Census

In the preparatory work for the 2021 Census the following requirements were adopted:

- meeting identified and anticipated needs of users,
- collecting high-quality data: accurate, reliable, complete and share the results of Census no later than 12 months after the reference date,
- ensuring high quality results of Census within a limited budget,
- reducing the administrative burden of respondents,
- preservation of the essential characteristics of census,
- ensuring comparability of census results in time and space (national and international level).

High rank is given to the quality of census results. Collection of high-quality data in the census is important to fulfil the needs of both external, as well as internal users – producers of statistics. The results will serve as a standard against which the correctness of the results of other statistical surveys will be evaluated. The results of the census will also provide a reliable sampling frame for the current surveys.

In order to further understand the needs and priorities of the users on the results of 2021 Census, the Central Statistical Office carried out consultations with the users of statistics. Among them were: government and local administration units, research centers and private users. There was observed considerable interest from the parties involved in the consultations on issues related to migration, particularly international ones. The topics discussed in the consultation focused on: long-term and

short-term immigration for temporary residence; Long-term and short-term immigration for temporary stay; Country of previous residence; Country of departure, fact of ever resided abroad, place of usual residence one year prior to the census, as well as country of birth and citizenship.

The information obtained was the subject of analysis and the outcome has been included in further work, with particular emphasis on the possibility of obtaining the widest range of information from administrative sources.

1.2. Conditions

The main challenge is to achieve high-quality results of 2021 Census with a reduced burden on respondents. In view of the above, the creation of conditions for the use of multiple data sources becomes crucial.

Application of new methods in the 2021 Census based on multiple data sources faces many limitations, which must be removed or reduced. The key determinants of the use of administrative data in the Census are legal, methodological, organizational, technical and technological.

In terms of legal requirements in national law, changes in the Law of 29 June 1995 on Official Statistics (OJ 2016.1068, as amended) were introduced. The amended provisions strengthened the role of the CSO President in shaping the information content and quality of registers constructed and operated by other public authorities. The access of official statistics to personal data from administrative registers was also regulated.

In terms of methodological conditions, the work is carried out under the project "Improving the use of administrative data sources (ESS VIP ADMIN WP6 - Pilot studies and their applications)".

2. Work on the use of administrative registers in the 2021 Census

2.1. Evaluation of the quality of administrative registers as a source of data for 2021 Census

Ongoing work was focused on identifying the information needs of users, assessing the utility of data sources and the development of theoretical solutions, methodologies for the integration of data from different sources, imputation, calibration of data and generalising the results based on a combination of data from different sources.

Based on the available metainformation, analysis of reported and anticipated user needs and EU requirements in the area of the 2021 Census was carried out. In order to include the identified needs in a uniform manner, their structure was developed in tabular terms. Variable names have been included in the table - the needs of users grouped into thematic areas. The user needs have been identified as a result of the analysis and a list of variables has been drawn up. In the next step an

analysis of the available resources of metadata on administrative sources were carried out, as a result of which administrative registers that are sources of data for the specified variables have been pre-selected.

The scope of variables related to international migration that can be derived from the available systems and administrative registers is rather exhaustive from the point of view of meeting international needs. Nevertheless, it is important to be aware that registers reflect legal status, which means that the act constituting migration is not mobility itself but an official registration. In the case of emigration this procedure is abided by a small part of leaving the country due to departure abroad, whereas immigration is usually registered only after fulfilling the conditions set out in the laws. The analysis proved that there are variables that could be covered from administrative sources in a full scope (for all items in the classification) as well as variables with partial coverage, which means there is no possibility to retrieve data from the registers to develop individual items from the classification. Below a list of variables is presented for which the work on administrative files is planned:

1. country of birth,
2. country of citizenship,
3. way of acquisition of Polish citizenship – citizenship at birth,
4. country of previous citizenship (in case of acquisition of Polish citizenship),
5. fact of ever resided abroad,
6. year of arrival to Poland (for persons ever resided abroad),
7. place of residence abroad – country (for persons ever resided abroad),
8. number of years of the last stay abroad (for persons ever resided abroad),
9. place of previous residence,
10. date of residing in a current place,
11. place of residence one year prior to the census,
12. country of permanent residence (temporary immigration),
13. period of stay (temporary immigration),
14. year of arrival to Poland (temporary immigration),
15. does the immigrant stay in Poland for the first time (temporary immigration),
16. country od stay (temporary emigration),
17. period of absence (temporary emigration),
18. year of departure (temporary emigration).

As part of the work on the usefulness assessment of data from administrative registers for statistical purposes, methodology for assessing the quality of administrative data sources was developed,

along with the checklist. According to the assumptions made in the methodology, the quality assessment is carried out independently for each register. Methodology for assessing the quality of the register covers three areas: general information about the register, information about the quality of the register, the quality of data from the register.

The general information contains basic information about the source. Information about the quality of the register include the following criteria: accessibility and clarity, usefulness, comparability, timeliness, coherence, cost of using data from the register. Information about the quality of the data from the register include the assessment of the quality of the database and the data contained in the database. In this area two criteria, accuracy and comparability, were highlighted. Within the individual criteria indicators and the method of recognizing their value were established. Description of quality is supplemented with the result of the evaluation and / or conclusions.

The approach to the assessment of the quality of administrative registers and their usefulness in the 2021 Census are shown in Fig. 1 and Fig. 2.

Fig. 1 Approach to the assessment of the quality of administrative registers as a source of data for 2021 Census

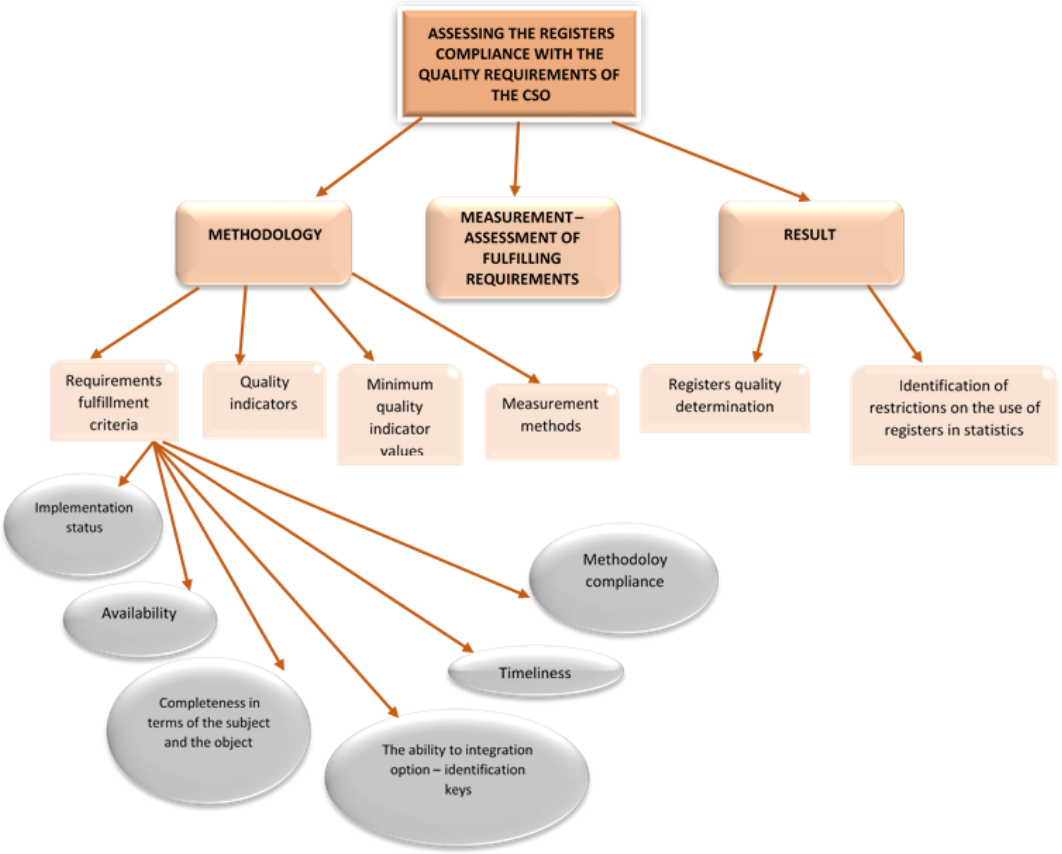
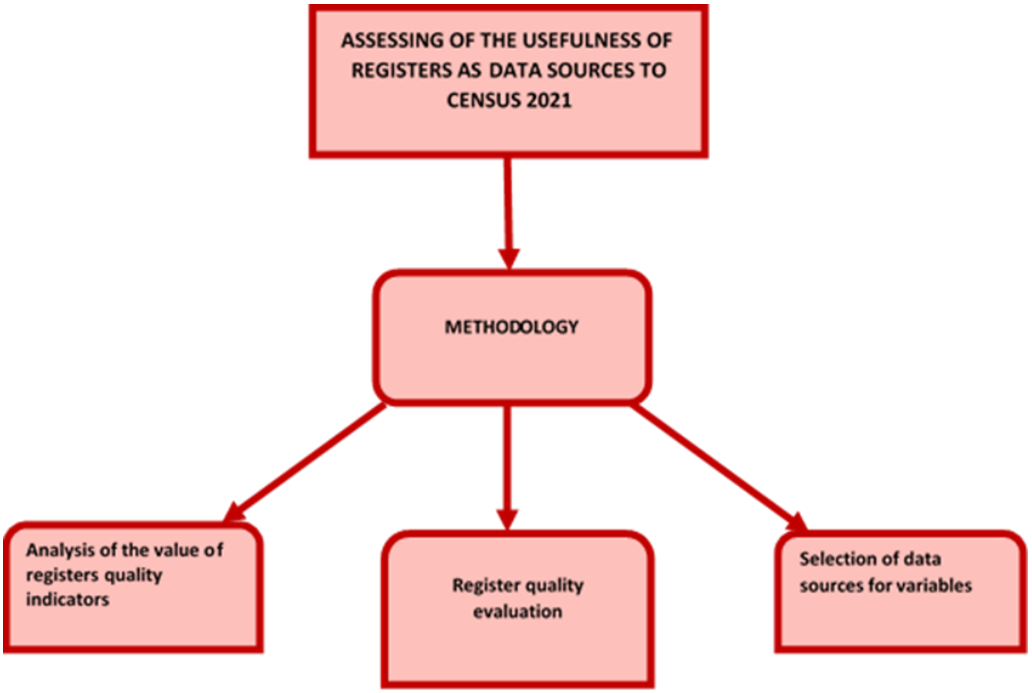


Fig. 2. Approach to the assessment of the usefulness of administrative registers as a source of data for 2021 Census



After assessing registers, complemented tables enabled the identification of all available administrative sources for the identified information needs (variables) for the 2021 Census. 41 administrative registers kept by 20 administrators were identified. This enabled nominating 29 administrative, reference and complementary registers, providing values for 36 variables that are planned for use under 2021 Census.

Due to the assumption made regarding the use of multiple data sources in the 2021 Census, under methodological conditions, conducted work included the development of theoretical solutions and pilot test of the possibilities of using administrative registers to: integration of data from different sources, generalizing the results based on a combination of data from different sources at different levels, territorial cross-sections and data calibration and imputation. It is expected that the estimators based on the model will allow for greater precision than the direct estimators, so they will improve precision of estimation and increase the quality of the results of 2021 Census in significant way. The developed method of calibration of results will be important for the 2021 Census, in particular in case of non-response and distortion of random sample by systematic errors. As part of the work, methods of data imputation will be developed. New methods will be tested on a pilot basis, to enable evaluation of their usefulness in 2021 Census. After a practical pilot application of developed theoretical solutions conclusions will be formulated, as well as assessment

of the possibility of implementing innovative solutions for 2021 Census will be conducted and the recommendations for the needs of 2021 Census will be formulated.

2.2. Conceptual model of Statistical Population Register

Due to the assumption regarding the use of multiple data sources in the 2021 Census, the work carried out included the development of the conceptual model of Statistical Population Register. An important prerequisite for the construction of the Statistical Population Register is also the strategy adopted by Eurostat on transition after year 2021 to an annual system for the data transmission for a specific range of variables according to the accepted levels of detail used for classification and the need to develop a new methodology for balancing the population, based on the actual place of residence, and not as so far - permanent address.

The main goal of constructing Statistical Population Register is the generation of a broad spectrum of statistical data according to different cross-sections of territorial division and according to a defined reference period.

The specific objectives are:

- testing and implementing new solutions and methods of variables derivation from administrative registers,
- developing methodological assumptions (including the construction of algorithms) enabling to keep migration statistics and population estimates of the actual place of residence, as well as the structure of the population; construction of the new methodological environment may determine to abandon the method, elaborated by the CSO for national and international purposes, used to develop population size and structure,
- identification and implementation of quality indicators for data sources on the basis of created system of related statistical variables on the basis of administrative registers,
- defining the requirements and recommendations for the construction of:
    - a coherent statistical system based on administrative registers:
        - standardization and automation of the processing,
        - standardization of the model of data and metadata,
    - processes for data processing, which aims to transform the data acquired (i.e. collected sets of variables for the groups of statistics products) in the data set, constituting the basis for the development of the resulting information and conducting analysis.

The basic feature on which each data output within Statistical Population Register will be based (as regards the data / variables derived for a particular thematic area) is the list of persons connected with the address-dwelling list.

The purpose of building a list of persons is to create a wide and complete, coverage of the data set in terms of content and some qualitative aspects characterizing the population for many statistical surveys.

Constructing the list of persons is based on the integration of data from multiple heterogeneous sources in order to obtain a uniform, consistent image of the data collected. The solution assumes the existence of two types of registers:

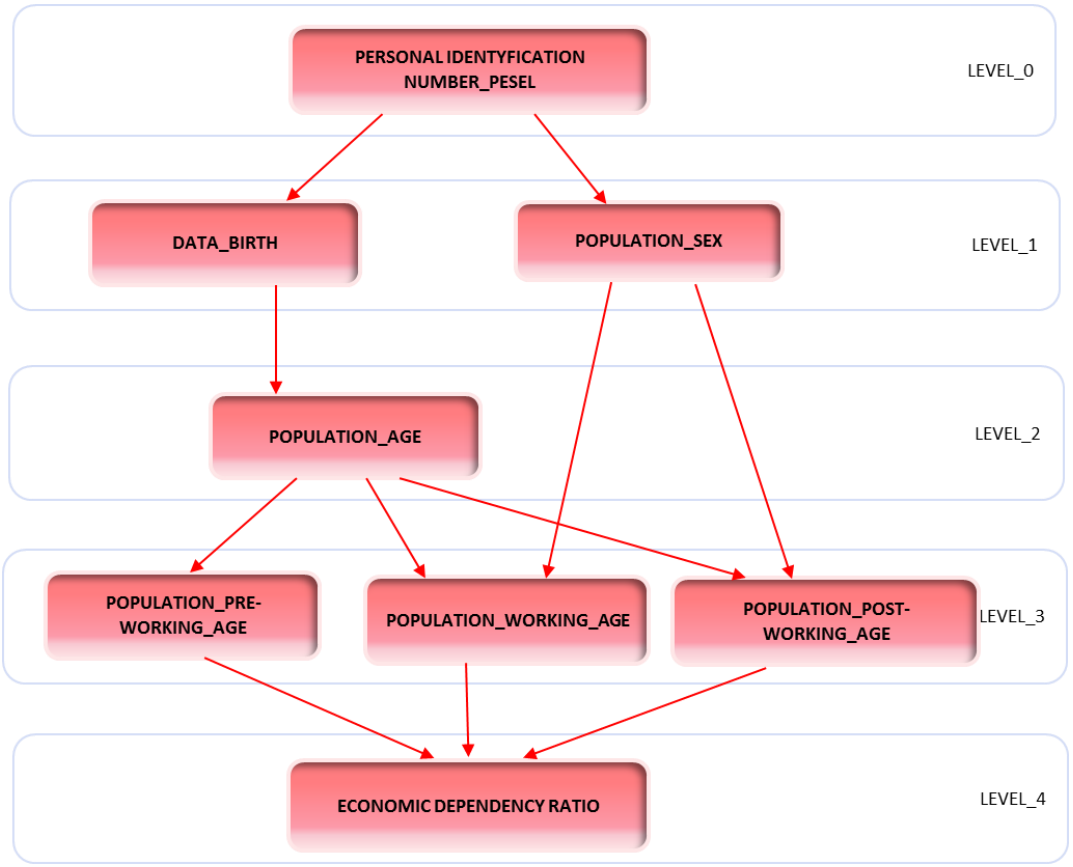
- reference registers ("cores") - characterized by the highest data quality, completeness and timeliness and a high subjective coverage so that they constitute the core, with which other registers will be integrated.
- supporting registers - other registers, the thematic scope of which meets the needs of statistics. Supplying the core with variables of the other registers requires choosing a variable value (based on the fixed rules) in the presence of the same variables in many sources, as well as allowing the calculation of the new variables, not occurring separately in any other source.

The target structure will be constructed in an evolutionary manner by implementing the individual modules of algorithms as needed. The concept involves the use of many registers for the construction of person list by finding unique Personal Identification Numbers (PESEL number) that passed correctness checks. Correctly verified PESEL numbers will be downloaded and included in the list according to the algorithm established on the basis of the referentiality of collections relating to the subjective scope of persons.

Construction of Statistical Population Register will be implemented in stages. The first step is transforming the datasets from administrative registers into statistical datasets. The second stage involves the integration of data and development of resulting information - calculating values of the variables and quality indicators. At each stage of processing it is assumed to obtain a new higher-level variable associated to variable or variables of the previous steps, for example: PESEL number (an eleven-digit numerical symbol that allows you to easily identify the person / inhabitant of Poland who owns it. PESEL number contains the date of birth, ordinal number, gender reference and control number) is a zero-level variable. By "decoding" the PESEL number, the first level variables are received: date of birth and sex. Second level variable - age will be calculated based on the first level variable - the date of birth. The aggregated variables of the third level indicating the

number of the population by economic groups of age will be derived on the basis of the related variables of lower levels: sex - first level and age - second level. In the last step there will be identified variable of the fourth level - economic dependency ratio.

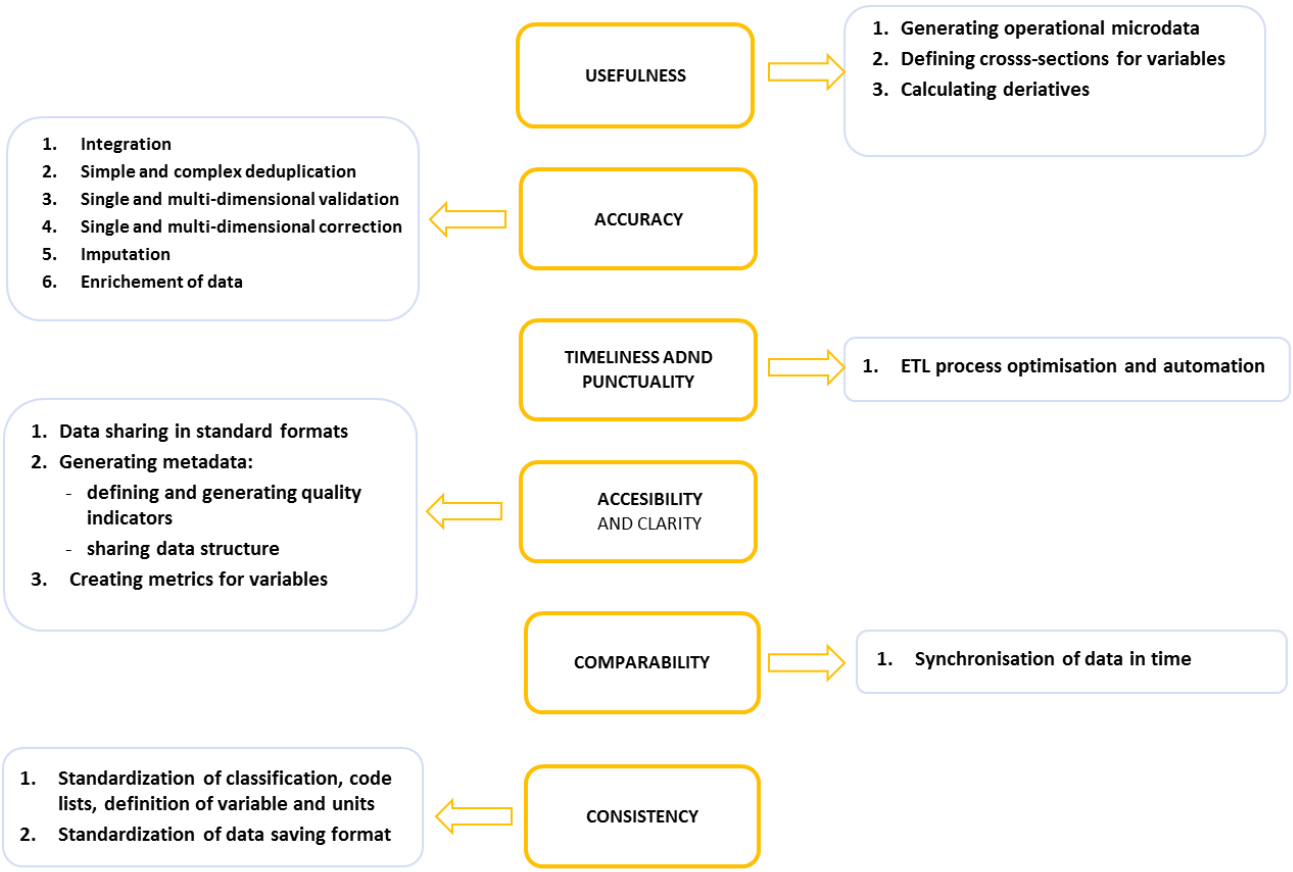
Fig. 3. Multi-level approach to obtain the values of exemplary variables



The adopted multi-level approach has a direct impact on the quality of the variables and indicators. Good quality of zero-level variables guarantees good quality of variables of higher levels. Together with the high level of detail of the input microdata it enables receiving high-quality output statistics at low territorial level with high frequency (usefulness).

The third stage involves the analysis of quality indicators of all statistical processes involving the construction of Statistical Population Register.

Fig. 4. Quality assessment of constructing processes of Statistical Population Register



At the stage of Census implementation, the Statistical Population Register will play a key role. Determining the subjective scope of Statistical Register will allow to determine the population covered by the Census according to place of residence before the census critical moment. The combination of information from the building-dwelling database, created after the Census 2011 and modernized for the needs of current surveys will in turn enable the preparation of characteristics of building and dwelling resources.

Register information resources based on data derived exclusively from the administrative registers will supply initially Census questionnaire, directed to the respondents in order to update the data. Electronic Census questionnaire will also include logical control and validation of data, and transition paths will enable the efficient filling in the questionnaire, which in turn gives protection of receiving good quality data.

Construction of Statistical Population Register before the reference date of the Census is related to numerous benefits, among which the most important are:

- gaining "on the input" the subjective scope of the so-called "target population", representing a set of all statistical units in a defined geographical area on a reference day that qualify for providing information on at least one specific topic,
- obtaining dwellings resources; inhabited, uninhabited, as well as non-residential house premises (including geographic coordinates x, y),
- obtaining resource of buildings in which there are dwellings with geographical coordinates x, y,
- linking people with specific dwellings (understood as a reference addresses of residence),
- identifying, before carrying out a proper Census, the lack of connections and other anomalies when assigning people to dwellings / addresses, e.g. the presence of a large group of people / clusters at the same address, and therefore taking actions to solve the problem and develop appropriate procedures.

Statistical Population Register will be divided into several thematic areas (domains) resulting from the commitments related to the implementation of the 2021 Census and beyond, but also in relation with currently run surveys:

- status and demographic characteristics of the population,
- the population in collective living quarters and the homeless persons
- education,
- economic activity of persons,
- commuting to work,
- disability,
- country of birth and citizenship,
- internal and international migration,
- nationality, language and religious affiliation (denomination),
- households and families
- status and characteristics of the dwelling stock (dwellings and buildings).

As mentioned before the basic element of the Statistical Population Register is the list of persons, which will constitute the reference population. The key that connects the population to the various domains (modules) will be the Unique Statistical Number (USN) assigned to each person, both the one who has PESEL number and the one that was included in the list of people without PESEL number. The USN will also be the key to anonymize PESEL numbers. Introduced domain variables

in conjunction with basic list of persons will constitute the so-called domain table. The number of records of such a table will always overlap with the list.

It is particularly important to select existing registers in Poland which, due to their scope, can be valuable in the development of results in particular thematic areas. The activities on establishing which registers are useful for which topics are focused on the acquisition of metainformation elements which are of key importance from the point of view of set suitability.

The intention of the CSO is to prepare for each thematic area (domain) the complete documentation concerning the suitability of the selected administrative sets, elaborating domain variables describing the unit on the list of selected thematic area and a pilot generation of statistics according to the established structure together with the characteristics.
The activities related to the construction of a domain list covering migration will aim at using a fairly wide spectrum of available administrative sources in Poland. Such an approach may seem controversial. Practice has proven that use of too many sets can consequently contribute to lowering the quality of the result data. However, migrations are a population phenomena which, unlike many of others (births, deaths, civil status changes) are not events recorded to hour or day, but complex, multi-step processes subject to change over time and recorded by different sources. The issue of residence, employment, education of foreigners in Poland is governed by a series of laws, for implementation of which different entities are responsible. The result of such a law making process is the creation of separate autonomous subject and scope registers in which data redundancy relating to foreigners is quite common.

It should also be emphasized that immigrants - especially short-term ones - do not have to own a PESEL number (these issues are regulated in detail by the relevant national law) and on the lists or in the databases they are registered according to internal procedures of the given authority.
For the purposes of the Census implementation and the construction of the list of domain "International Migration", administrative sources were identified, which will provide information source on migrants (immigrants and emigrants). As in the case of building a basic list of persons, also in this case, it was decided to use a solution that envisage the existence of two types of registers: reference and support.
As reference registers for features associated with migration, the following were indicated: PESEL register, National Taxpayers Record, Social Insurance System, Health Insurance System the POBYT system (national collection of registers, accounting records and lists which contain data on foreigners' cases).

. In turn, as supportive administrative sources, the data sets run by Minister responsible for labor (including information on the employment of foreigners), the Minister responsible for science and higher education (containing information on foreign students and university graduates), and the Minister responsible for education (data on pupils of foreigners in primary and lower secondary schools secondary) were identified.

The approach taken should allow the collection of data on both the number and basic demographic and socio-economic structure of emigrants and immigrants. It should be borne in mind, however, that the information available from administrative sources is of purely formal and legal nature, which significantly limits the assessment of the actual migration phenomenon. Hence, the role of statisticians, who should support data owners in improving the quality of registries and administrative systems, is extremely important and is beyond question.