ROMANIA

Reimbursable Advisory Services Agreement on
Romania Capacity Building for Statistics (P167217)

OUTPUT No. 7a

Report on advisory services provided to Recipient on the Methodology to assess and promote continuously and in real time the quality and coverage of the collected data PHC2021 and data protection/security

May 2022

Revised version December 2022

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

Project co-financed from the European Social Fund through the Operational Programme for Administrative Capacity 2014-2020

Disclaimer

This report is a product of the staff of the World Bank. The findings, interpretation, and conclusions expressed in this paper do not necessarily reflect the views of the Executive Directors of the World Bank or the governments they represent. The World Bank does not guarantee the accuracy of the data included in this work and does not assume responsibility for any errors, omissions, or discrepancies in the information, or liability with respect to the use of or failure to use the information, methods, processes, or conclusions set forth. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

This report does not necessarily represent the position of the European Union or the Romanian Government.

This report has been delivered in May 2022 and the revised version in December 2022, under the Reimbursable Advisory Services Agreement on Romania Capacity Building for Statistics (P167217) signed between the Romanian National Institute of Statistics and the International Bank for Reconstruction and Development on September 17, 2019. It is part of Output 7 under the above-mentioned agreement.

Acknowledgements

This report was prepared under the coordination of Michael Wild, Senior Statistician, World Bank with the support of the local team of experts. The team would also like to express its gratitude to government officials of the National Institute of Statistics (INS), Silvia Pisica (General Director and Project manager), Andoria Ioniță (Director), and their team of specialists for their constructive collaboration.

CONTENTS

Figures

Tables

List of Acronyms

AC&S	Address Canvassing and Sectorization
API	Application Programming Interface
ATU	Administrative Territorial Units
CAPI	Computer-assisted personal interview
CASS	Computer-assisted Survey System
CAWI	Computer-assisted web interviewing
DB	Database
DPS	Data Processing and Storage
DOP	Deepness of Data Processing
DTS	Territorial Statistics Directorates
EC	European Commission
ESS	European Statistical System
EU	European Union
EUINSPIRE	Infrastructure for Spatial Information in the European Community
FA	Factor Authentication
GEOLOC	Geographical Location Software Application
GIS	Geographical Information System
GPS	The Global Positioning System
ID	Identity Document
INS	National Institute of Statistics
IP	Internet Protocol
PES	Post Enumeration Survey
PII	Personally Identifiable Information
PIN	Personal Identification Number
PHC2021	Population and Housing Census in Romania 2021 round
RAS	Reimbursable Advisory Services
REGEX	Regular expression tester
SIRUTA	Information System of Administrative Territorial Units Register
SPA	Special Purpose Application
SSL	Secure Sockets Layer
STS	Special Telecommunication Service
Survey Solutions	Survey Solutions
TSE	Total Survey Error Model
TOTP	Time-based One Time Password
UJIR	Counties Units for Census Implementation
VPN	Virtual Private Network
WB	World Bank
WEBGIS	Web Geographical database

Introduction

The purpose of this report is to present **the methodology to assess and promote continuously and in real time the quality and coverage of the collected data PHC2021 and data protection/security**. This is part of the deliverables under the Reimbursable Advisory Services (RAS) Agreement on *Romania Capacity Building for Statistics (P167217).* The project is implemented by the National Institute of Statistics (INS) with support from the World Bank.

This report presents additional information on processes regarding quality and coverage of data collected during actual census, as they were documented during the Population and Housing Census (PHC) pilot (Output 10a[1]) and complemented by the sectorization process presented in Output 7c[2].

This document has five (5) sections.

The first section provides an overview of the data quality and coverage concerning the PHC, of the importance of data quality for statistical production, of the attributes of an effective quality assurance plan for the actual PHC data collection and the components and mechanisms for achieving quality data.

Section 2 provides a description of the operations, the steps of the data collection during a census and the approach applied in the case of PHC2021 in Romania, as well as the effects on data quality achieved after the implementation of the top-level processes: the Frame building, the preparation of the WEBGIS database, and the quality control process during census data collection including all its phases and steps.

The Special Purpose Applications (SPAs) prepared to support the data collection for the monitoring and evaluation of the coverage and the activity of the enumerators during PHC are presented in Section 3. The paradata reports for the questionnaires and enumerators are presented at the level of segments (sectors of census), counties and at national level.

The statistical support for data quality is presented in Section 4 and it considers the specific training materials, handbooks, and video-presentations by roles attributed for data collection, monitoring and evaluation during the actual census through Survey Solutions, respectively of the enumerator, supervisor and headquarter.

Section 5 presents the data protection/security specifics for the census process during self-registration, CAWI and CAPI processes in complying with the requirements for the Personal Identifiable Information (PII) protection and IT security. This section is detailed in other two reports under the RAS Agreement, respectively the Output 10c[3] regarding the actions applied for data protection and security during PHC data collection, and Output 10b[4] regarding the Statistical Disclosure Control process applied.

---

[1] Output 10a: Report on Recommendations to the Recipient on how to perform the PHC2021 piloting process

[2] Output 7c: Report on advisory services provided to Recipient on the Recommendations and best practices for implementing a data management system for PHC202 geo-spatial data and the actual sectorization

[3] Output 10c: Report on advisory services provided to Recipient on the Set of draft statistical census tools using multi-modal methods to promote data protection and security

[4] Output 10b: Report on advisory services provided to Recipient on the Technical assistance and best practice recommendations for SDC, data confidentiality, ways to secure micro-data and aggregate data

1. Overview of quality assurance framework for data collected during Population and Housing Census

Quality assurance is a process ensuring that quality goals:

- full coverage of target usual resident population - every person having the usual residence in Romania for at least previous 12 months before reference moment is enumerated including hard-to-find population categories;
- completeness - all the data requested (by international and national pieces of legislation) is provided through interviewing questionnaire ensuring the measurement of all other statistics; unicity - everyone is enumerated only once and in the right household and
- simultaneity - information obtained about individuals and housing in the census refers to the same point in time, the 2021 PHC reference date (December 1st, 2021)

are consistently met throughout the whole system of data production during census.

The major goal of a quality assurance framework for the census is to prevent and minimize potential errors at the design stage and detect errors as soon as possible so that timely remedial actions can be taken even as the census operations continue.

The quality assurance plan for the PHC should be comprehensive and should cover all activities in all its stages from planning, design, organization, development, data collection, processing, evaluation and validation, and dissemination of census results. The quality of statistical data is the result of the data generation processes, and deficiencies in data quality are usually the result of deficiencies in process rather than the actions of individuals working in that process. The key to quality assurance and improvement is to be able to regularly measure the timeliness and accuracy of a given process so that the process can be improved when deficiencies are detected.

Without such a plan, the census data may contain many errors which can severely diminish the usefulness and credibility of the results.

An effective PHC quality assurance plan should have the following attributes applied in different stages of it, respectively:

- a well-designed system or procedure tailored by type of census (traditional vs. combined or registered based census), data collection methods (mixed-mode data collection or single data collection mode) and adapted period for data collection, and to that to maximizes quality and efficiency);
- a well-designed data collection instrument, containing all the relevant validations, and adapted for easiness of accessing, understanding and using (to consider the content and language used, the complexity of variables to be collected, the user-friendly process and features for authentication, navigation and completion);
- a qualitative statistical infrastructure (data from administrative sources, dwelling database, street nomenclature, IT infrastructure and communications);
- an effective and adapted organization for a good comprehension of the instruments used and the data collection tasks, carried out by the respondent (during CAWI phase pre-registration and self-enumeration), by the enumerator and other census staff (during the selection, the training program and CAPI phase, regarding the structure of census sector and allocated quota of dwellings/persons, the responsibilities in relation with other census staff and/or public authorities) and by the other stakeholders of census (i.e. local public

administration, central administration, ethnical minorities, others), and a consistent and permanent information of society at large to build a positive attitude towards the census
- a quality control program to ensure the desired level of quality during the course of the census operations, respectively on data collection (including data from administrative sources and other census' databases), on data processing and validation (regarding logical conditions and their level of applicability for data control, the methods applied and deepness of imputation and SDC), and on data dissemination (the level, the way the results and methods used in data processing and data validation are presented to the public)
- an evaluation program to measure the accuracy of the census operation and identify areas where future improvements may be made.

Quality is therefore not just the outcome of mechanistic applications of predetermined measures but relies on a combination of:

- established, documented processes;
- systems to monitor the outcomes of these processes;
- active encouragement by management to involve staff undertaking the processes in identifying and resolving deficiencies with quality;

Quality may be accomplished by:

- setting standards and using data to improve the process;
- ensuring a better understanding of the overall process by the census staff and their involvement at all phases;
- including quality issues in training programmes;
- quality feedback for each census process for on the spot operational changes when necessary

The success of any quality control and improvement plan depends on:

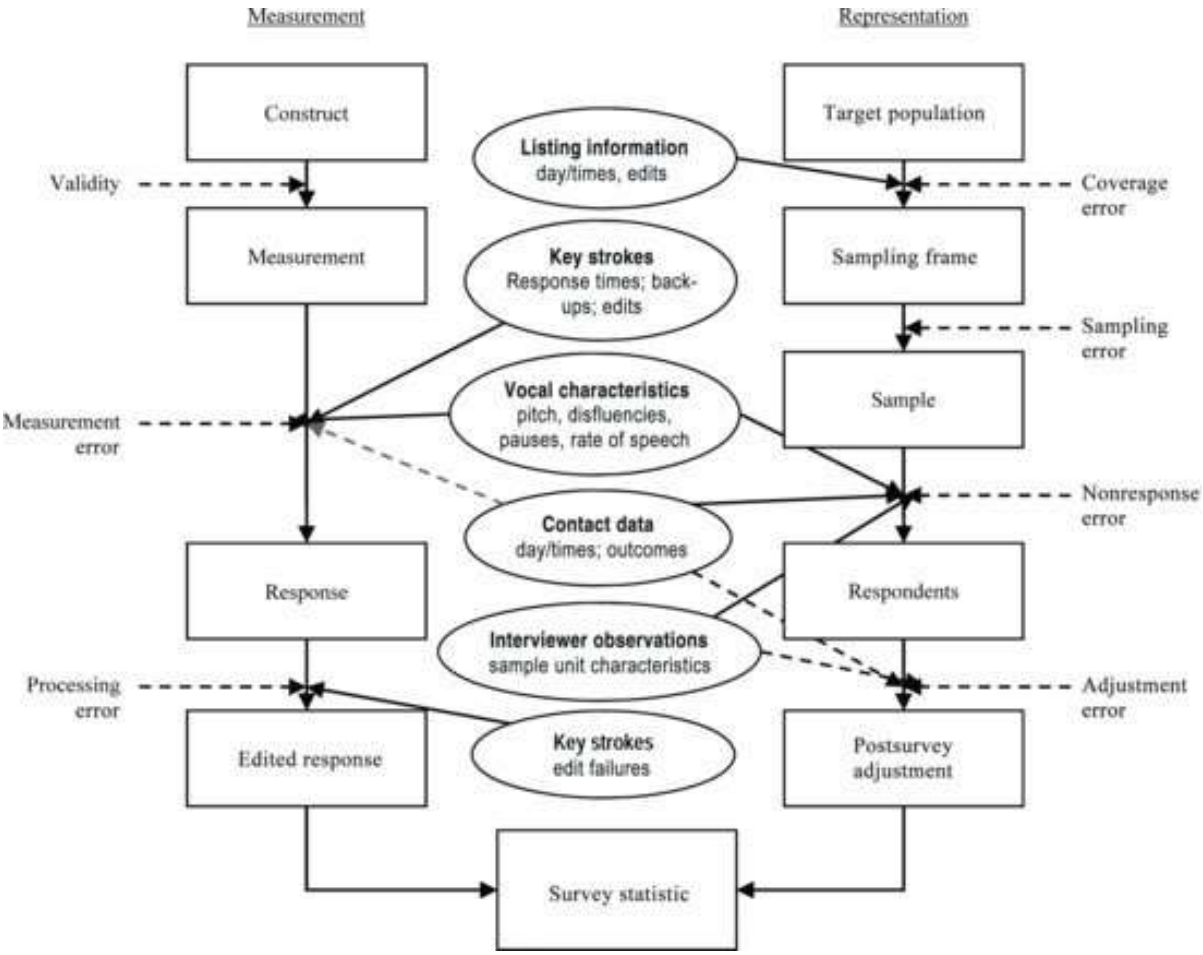
i. defining quality standards or requirements;
ii. determining appropriate verification techniques;
iii. measuring quality;
iv. providing for timely feedback from the results of the plan so that effective corrective action may be taken.

The main techniques that can be applied during data collection, and also at data processing phase, are complete verification, sample verification and post enumeration surveys.

2. Quality control in support of data operations for PHC

The core element driving the entire process of PHC2021 implementation and results obtained is the Quality control process. The quality control process for PHC is based on the Total Survey Error Model (TSE) model with the following structure – Figure 1:

Figure 1- Total Survey Error components and related paradata



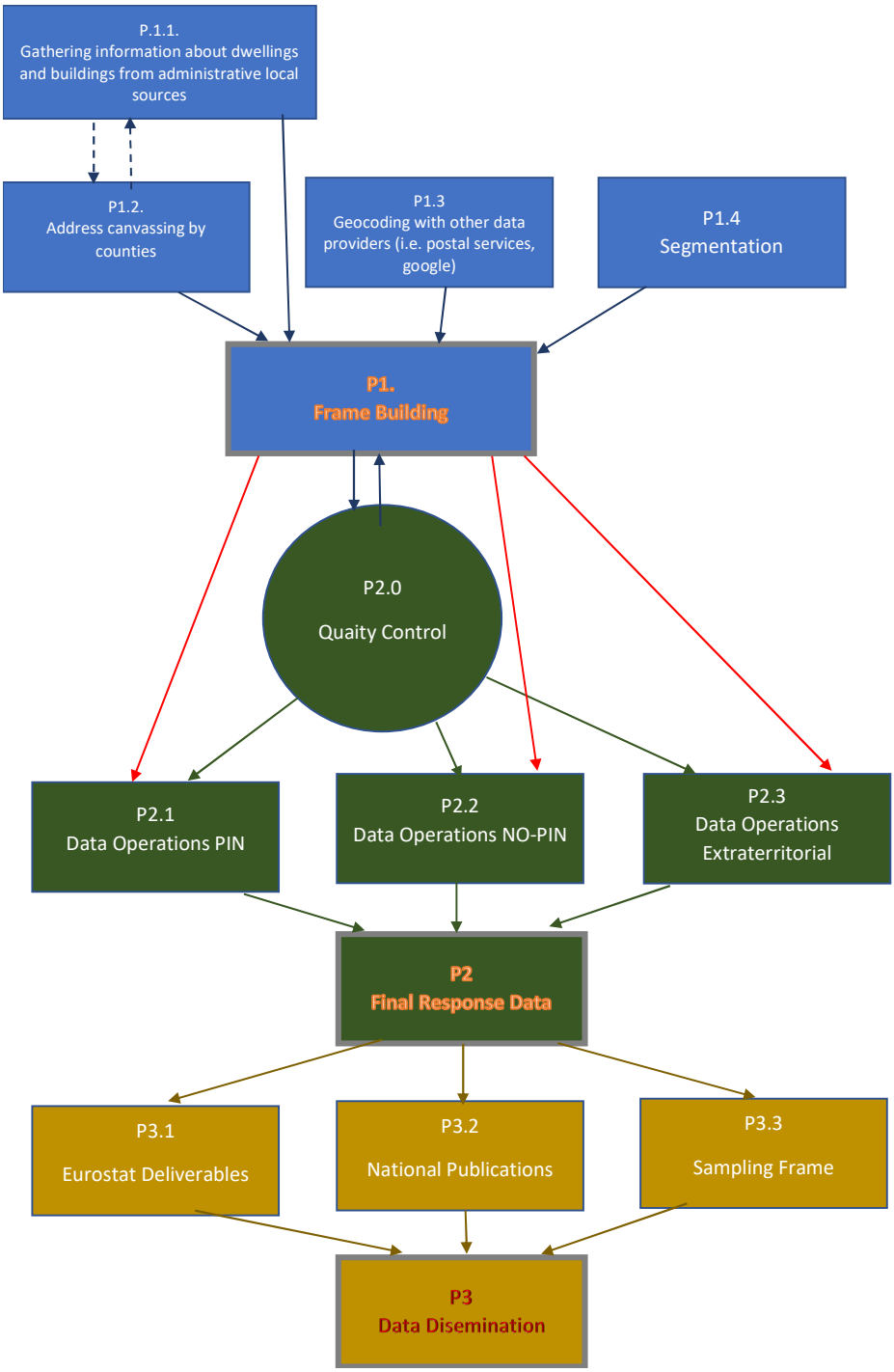
Source: (Kreuter and Casas-Cordero, 2010)

If no sub-sampling takes place, the model applies to all the non-sampling components, and if sub-sampling takes place for some variables, then also the sampling part applies, i.e., Sampling Error. Also, some of the other components may only apply for some of the data items but not all of them, since they are predetermined by European Legislation, like Validity, or Measurement Error. Nevertheless, every effort should be made through instrument testing and the pilot data collection, to verify these concepts as well.

Besides the process addressing the TSE through the design of the instruments the quality control process also requires its own organizational structure, with the team of territorial quality support being directly assigned to the central quality control group. This facilitates communication and workflow, which is an important element of the intra-census quality control process.

Thus, the quality control is part of all census's operations covering all stages (design, organization, data collection, processing and validation, and dissemination – as briefly presented in Section 1), the related data operations' quality control steps (Figure 2) are the most critical and are presented further.

Figure 2 - PHC Quality control process of data operations



There are other quality controls, not only in data collection phase. One especially important takes place in data aggregation phase, before dissemination.

Due to Romania's integration into the ESS, several of the steps are either already completed, or are predetermined by European legislation. It is therefore, that these guidelines start with the Preparation of frames and mapping in the process group P1. Frame Building and part of the 3 main process groups.

1. Frame Building (P1.)
2. Data Collection (P2.)
3. Data Dissemination (P3.)

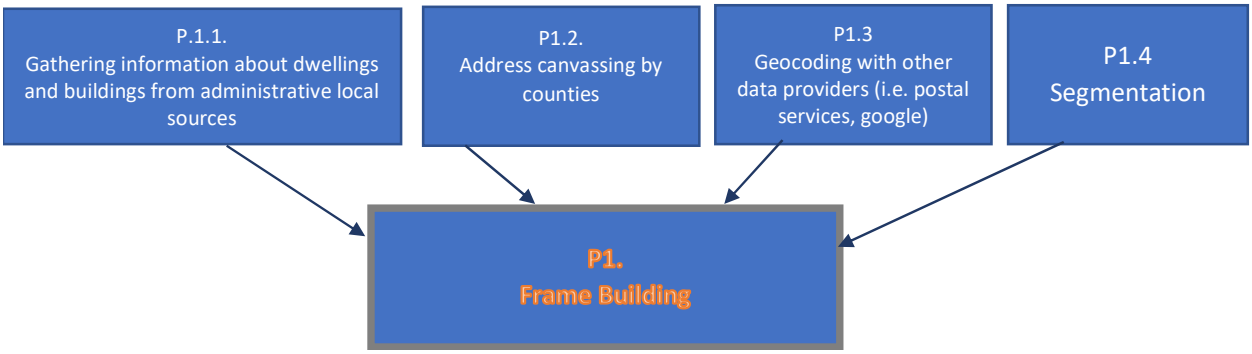
Each of these process groups contains several interacting sub-processes, which produce the input for the next process group. From a process management perspective, these process groups also constitute the "gates" for the transition to the next process group, and without a successful completion, meeting all the required success criteria, this transition would not be permitted.

The sections below summarize the concepts and transpose them into the specific of quality control requirements for tools applied to actual PHC.

2.1. Frame building - Preparation of WEBGIS Database (P1)

The construction of the census frame is one of the most important preparatory activities for any census, as it should contain all housing units to be enumerated during the census, preparing by that the WebGIS database (see Figure 3).

Figure 3 - Process groups for Frame Building and WebGIS Database



The WebGIS database has two purposes, namely to:

i. provide the complete housing-based census enumeration units down to the dwelling level
ii. to be used for sample survey purposes and in particular demographic surveys in combination with administrative records to update the population statistics in the post-censual period.

P1.1. Gathering information about dwellings and buildings from local administrative sources
For the actual census the database for dwellings and buildings has been developed using the administrative data sources from local level: the streets registry managed by city halls, the address lists of buildings and dwellings from the National Agency for Cadaster and Land Registration. In the cases of missing the available information gathered from local level, the address lists were updated by visiting the census sectors by counties' staff of INS (from UJIRs) before the start of data collection phase. The updated databases were used for Frame Building, respectively developing and updating the WebGIS database (see details of process further).

The process of Frame Building preparedness could consider sample surveys for database verification specifically designed to improve gridded coverage estimates. These surveys are designed to verify primarily the coverage of the frame data, before, during and after the actual census data collection process. The sample is based on the WebGIS database as well as any auxiliary information which can be allocated in space. The sampling design is based on 1 km^2 grid cells and well stratified, by any auxiliary information available in geo-referenced form.

In addition, and depending on the quality of the continuous sample survey process, a Post Enumeration Survey (PES) may also be relevant if no other means of evaluation do not exist.

In cases where these sample surveys are not possible, because of time or resource constraints, but administrative data sources are available, if known to be accurate, the verification of the database can also be carried out through such sources.

P1.2. Address Canvassing by Counties
The counties provide addresses themselves retrieved from local county administration. The activity is interlinked with P1.1 and it continues until the database is ready for use in census CAWI phase. This activity is carried out by the INS WebGIS team.

P1.3. Geo-coding with other Data Providers (i.e., Postal Service, Google)
The geocoding of buildings/dwellings addresses is a continuous process during entire census. For the CAWI phase the geocoding was prepared by INS and provided as nomenclature from the WebGIS database.

In cases where addresses can not be identified with own resources, a commercial service like the Google geo-coding API, or the HERE geo-coding API may be used in conjunction with the existing database, to provide the most comprehensive coverage.

P1.4. Segmentation (Sectorization)
Segmentation is conducted to divide the total number of dwellings into segments of equal size, for logistical and statistical purposes. Segmentation for the PHC2021 is conducted based with a target size of approximatively 600 dwellings inside the segment. Within the segment, the distribution commonly doesn't matter, so if it is a single building with 600 dwellings or 600 buildings with 1 dwelling each, was not considered in the segmentation process. Supplementary information is provided under Output 7c, dedicated to sectorization process.

The segmentation is done based on the known dwellings on known streets. Since the completeness of the original database in terms of the street names and building/dwelling numbers can't be relied on, the self-registration form should permit free entry of such information. Where the information is entered using free entry, subsequent encoding should be done to resolve inevitable spelling errors or erroneous additions of known (encoded) streets.

WebGIG database preparedness

The WebGIS database represents the core for Frame Building and is used in CAWI and CAPI data collection phases of the census. The steps recommended for preparedness of such database and the actual process done for the census are presented further.

While WebGIS database can be utilized for assisting in coding of the streets during self-registration phase (for main and secondary addresses), it can't be utilized for the CAPI phase of enumeration. This is due to limitations of the technology applied for the CAPI interviewing:

- the Survey Solutions software has a limit of 15,000 items for the categorical selection, meaning that the full list of streets from which a street can be selected may not be more than 15,000;
- the Survey Solutions software does not have a possibility for connection to a remote database for live querying since Survey Solutions works autonomously (disconnected) in the CAPI mode.

Correspondingly, for the CAPI mode:

- if self-registration has been done properly, but the actual census questionnaires not filled out totally, the enumerators should start with a set of interviews with preloaded main and secondary addresses (obtained during the self-registration), which should be protected from accidental or intentional change by the enumerators during the face-to-face interviewing.
- if self-registration has not been done, the enumerators will have to start from an empty interview completing all fields in the CAPI questionnaire on the tablet, with the exception when the the full address of the residence is available in WebGIS. The address will be prefilled in the questionnaire when are prepared the tasks of each enumerator. By that the field of address is filled in the CAPI phase, also.

In the latter case there will be a non-trivial number of cases, where the street names and building numbers are entered manually by typing, not selection. Exact matching to the reference database may establish the link with only part of them, while more elaborate algorithms of error correction and inexact matching may be required for establishing a full link.

Any new streets or buildings reported by the respondents during the self-registration phase or by the enumerators during the CAPI-enumeration should be validated before the ending of the PHC, so that the results of the census provide viable data for the construction of the new Census Frame. In order to ensure the correct coding of addresses, during the census data collection phase the enumerators must register, for the new addresses, the coordinates of the building by pressing the GPS button in the census questionnaire installed on their tablet.

The nomenclature of the country's administrative division should be utilized in identification variables as codes to simplify data entry and make it more robust to typos. This is provided that each such level of administrative division fits within the above-mentioned limit on the total number of the units in a selection question. The nomenclature should be the same for all components of the system, including the questionnaire, the reference databases, digital maps, etc. Where it is not possible to achieve consistent coding, the recoding using a crosswalk file shall be applied.

To guarantee classification consistency, the self-registration portal should and will directly connect to INS WebGIS database (most likely a replicate). This will guarantee, that complete address IDs are observed and, in this way, hopefully, will be avoided any mis-localization.

For this purpose, two important preparatory steps are required:

i. have a publicly accessible database (e.g., see here: https://docs.bitnami.com/virtual-machine/infrastructure/postgresql/administration/connect-remotely/)
ii. have all codes in the questionnaire harmonized with the codes from the WebGIS database.

For the interviewee to have a streamlined experience and only choose streets in his/her locality, a cascading filtering system was applied. This was created, using the primary administrative levels – county, ATU (administrative-territorial unit) and locality (components of ATUs). The filtering used lists (databases) generated from the open-source file released by INS and available through data.gov.ro (administrative units including SIRUTA codes), respectively:

- The first level (level 1) of the SIRUTA classification includes the counties and the Bucharest Municipality and their coding system.
- The second level (level 2) of the SIRUTA classification includes municipalities, cities, communes in each county and their coding system.
- The third level (level 3) of the SIRUTA classification includes the component localities, villages and sectors of the Municipality of Bucharest and their coding system.

For the actual census use of WebGIS, the export provided by INS included 6,141,532 observations, i.e., address points. The initial database was used as the base for the database for the self-registration phase. Along with the POINT coordinates (lat/long format) of each address (point), the database included several additional variables used for the self-registration stage: street ID, address ID, street category, street name, number, block, and the total number of dwellings in the building.

Previously launching in production, the WebGIS database required some cleaning of the variables and concatenation. This method was chosen to increase the number of unique addresses with point coordinates that can be then used to populate the database for the CAPI phase - e.g., several address points had the same street name and number but different street IDs inside the same ATU. Since street IDs could not be used in the form, it was chosen to expand the prefilled address name with the block variable since this variable could help differentiate between the address points.

The cleanup and standardization phase for the WebGIS database included the following steps:

- Cleaning and concatenation of the street name and street category variables. This step was included to make the street names more familiar to the interviewee and help differentiate between streets with the same name but different categories, e.g., *Strada* Viitorului vs *Intrarea* Viitorului.
- Following the street name cleaning and wrangling, a fourth database was created that included the localities' street names, street IDs, and corresponding inferior SIRUTA codes.
- Several cleanups were required to create a unique field based on the street number and block fields in the database. Because there was no standardized way of filling out the number and block fields, these were sometimes used interchangeably (e.g., block details in the number field) or included additional information about the data point. These details were not appropriate for these fields. As a result, these needed to be standardized and cleaned in a way that made sense in terms of an address and made sense for the interviewee. Approximately 160 REGEX patterns were created and used to standardize these variables.

In addition, commas (,) and quotes (“) were removed as these are used by the database server as separators.

- The number and block variables were concatenated after the cleanup into a single variable.
- The addresses were then exported into a fifth database that included the address IDs, street names and corresponding street IDs.

After the cleaning phase, additional issues were identified in the addresses database, making some of the addresses unusable in the self-registration stage. Summed up, these include:

- Addresses missing details. These addresses included different address IDs; however, based on the values of the street number and block variables, it was impossible to differentiate between the two or more addresses. These summed up 250,169 addresses.
- Addresses missing numbers. These addresses included different address IDs but no street numbers or blocks. This makes it impossible to be used in the self-registration phase because no number can be attached to these IDs. There are 208 cases in total.
- Addresses missing street names or street IDs. There were 8,805 cases where either the street ID or street name was missing from the address. Even if these addresses had IDs and point coordinates, it would have been impossible to correlate these with a street name in the stage of self-registration.

The total final database used in the self-registration phase includes 5,864,084 address points entries. The excluded addresses can still be used in the CAPI stage of the census but will lack details regarding the self-registration phase. In these cases, manual fixes are required to clarify the link between the street address and address IDs.

To not face serious issues regarding street database and with the geo-referencing process it is recommended to access alternative databases for development of a reference dictionary to match INS street names to the database. With such a dataset, it would be easier to harmonize the street names INS has included in their database, as well as to create a link table with the different Synonyms, which makes a text-based search easier.

Alternatives could be considered like an example from the Romanian Postal Service (Poșta Română) which should have the most up-to-date database of streets and street numbers. They have a website where anyone can search the postal number of their address: https://www.posta-romana.ro/cauta-cod-postal.html or https://www.rozip.info/, or from the one of largest courier company, (https://www.tnt.com/dam/tnt_express_media/ro_ro/download_documents/services/OOA_zone%20domestic_apr2016_v3_new.pdf).

The actions taken, described above, could be applied by case in any similar situations of preparing the WebGIS database for future surveys.

2.2. Quality Control (P2.0)

The quality control process for PHC is based on the Total Survey Error Model (TSE) (Figure 1) which expose two groups of errors of a statistical survey: the representation errors (coverage, sampling, nonresponse, adjustment); and the measurement errors (validity, measurement, processing). Considering the design, organization, data collection, data processing and validation, and dissemination as stages of PHC and map them with TSE it can be observed the distribution of errors by stages (Table 1) that should be identified and correct to have good quality data as a result of census.

Table 1 - Type of errors by PHC stages

Error Type / **PHC Stages**	*Coverage*	*Nonresponse*	*Adjustment*	*Measurement*	*Validity*	*Processing*
Design	✓					
Organization	✓					
Data collection	✓	✓				
Data processing and validation	✓		✓	✓	✓	✓
Dissemination					✓	✓

Even though Quality Control can and should be an independent support process of its own, we decided to pursue a more integrated approach, given the degree of involvement into the different process groups (based on Figure 2), and in particular the importance of the census frame generation for the actual data collection process. We therefore embedded it into the main process groups. Moreover, its position in the top-level processes also underlines its importance as being an interface between data collection and the overall frame generation process.

P2.0.0 - Tabulation and Instrument Design

Although quality control is placed in P2. Final Response data, its work starts already much earlier, as it needs to also provide inputs to the creation of the WebGIS database, as well as the development and testing of the resulting quality control framework including the questionnaire.

P2.0.0.1 - Questionnaire Development Process

A considerable part of the non-sampling error can be eliminated by addressing the left-hand side of Figure 2 through a well-designed data collection instrument. If well designed, it addresses:

- Measurement Error by a well-tested set of global validations, as well as
- Processing Error by a well-tested set of data entry conditions.

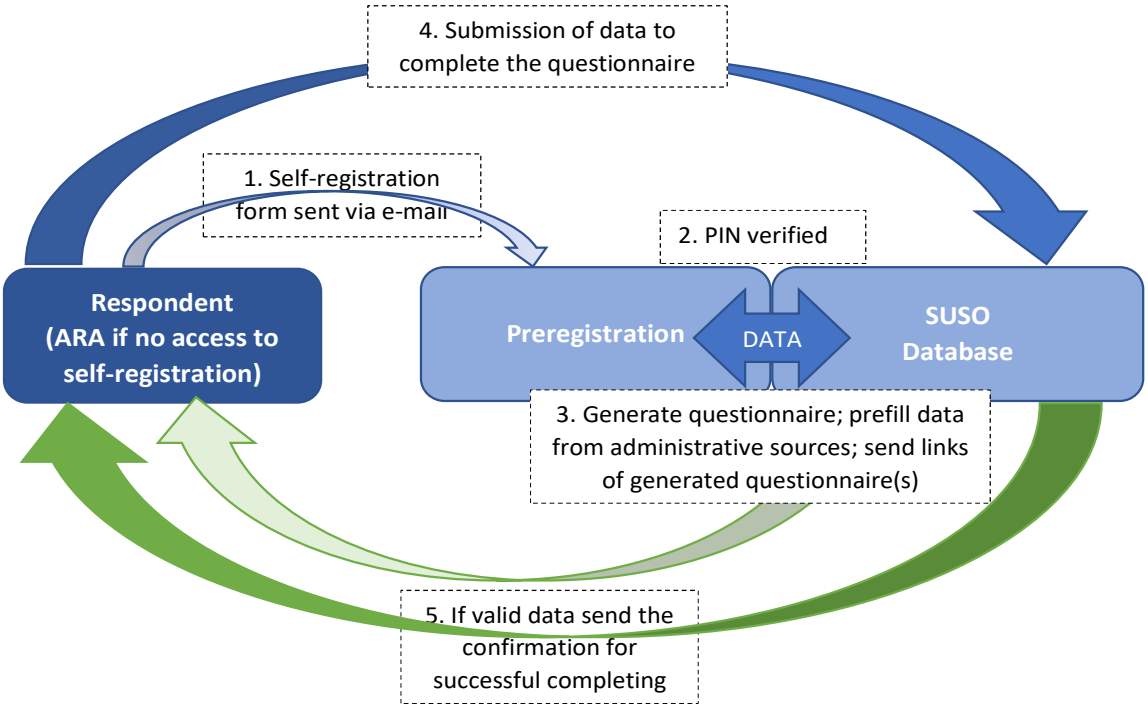
The questionnaire design starts with elaborate the set of variables necessary to be collected for PHC2021. Based on previous experience of structure of the questionnaire the sections/chapters were adapted to the Survey Solutions designer, preserving the main structure, from statistical point of view, related to the categories, especially. The questionnaire was tested in three phases, one, inside the headquarter of INS, the second phase in the territorial offices, and third one, in the Pilot of PHC2021. Each phase ends with improvements of questionnaire design regarding including

better explanations for better understanding of the questions and possible answers, especially the boundaries of the numeric answers; data validation conditions; including the validation variables for measurements of completion and statuses of the questionnaire completion process.

P2.0.0.2 - Electronic Submission System

The CAWI process applied for data collection contains preliminary quality control check due to flow designed and the automated validations of questionnaires that checked the completeness of the answers and generate the self-enumeration proof (see Figure 4). In the case of outside country respondents, Romanians who stay abroad for less than 12 months and who have their habitual residence in Romania, they could self-register as a person from inside country by accessing recensamantromania.ro (https://autorecenzare.insse.ro/)

Figure 4 - PHC Questionnaire validation workflow in Survey Solutions CAWI



Self-registration

Household head/Reference person (or any other 'authorized member/ARA by case) visits census website, and enters required self-registration, which is:

- Personal ID (PIN) and Email of her-/himself
- Personal ID and email of all other eligible household members (or at least one email address for all members),

- An temporarily artificial personal ID, generated only for census purposes, which included gender and date of birth for foreign citizens with usual residence in Romania, if they do not have a PIN
- The full address, including street name, administrative number, block, floor and number of apartment and where available, of the household, through intelligent drop down
- Street address of the secondary dwellings, if the case

PIN validation
Starts with uniqueness of PIN(s) and continues with basic and specific validation of PIN (formally, existence). Validates address and checks over-/under-coverage for provided address.

Questionnaires generated, prefilled, and sent:
, After validating the pre-registered PINs, a person questionnaire is created for each registered household member, pre-filled with existing information from administrative sources used for validation/navigation purposes. For each household head also a dwelling section in his/her own questionnaire is created.

Response by PIN is monitored on regular basis, with reminders/in-person follow ups

Submission of Data requested through questionnaire
Each person fills the data and submit the complete questionnaire. Multiple checks are applied (built into the questionnaire) to control the logic (e.g. presence or absence of answers at the compulsory variables; correlation between answers of several variables; normal limits for quantitative variables;) and the completeness of answers (all questions have answers in the format required through questionnaire: figure, text, selection from a list of items), that together validate the completion of the questionnaire and allow/inform the respondent about the end of the process.

The completed questionnaires are sent to the server where the automated validation is applied for other checks, including unicity of the person in the collected data. The automated validation of the questionnaire is a post-collection process, included in the validation phase.

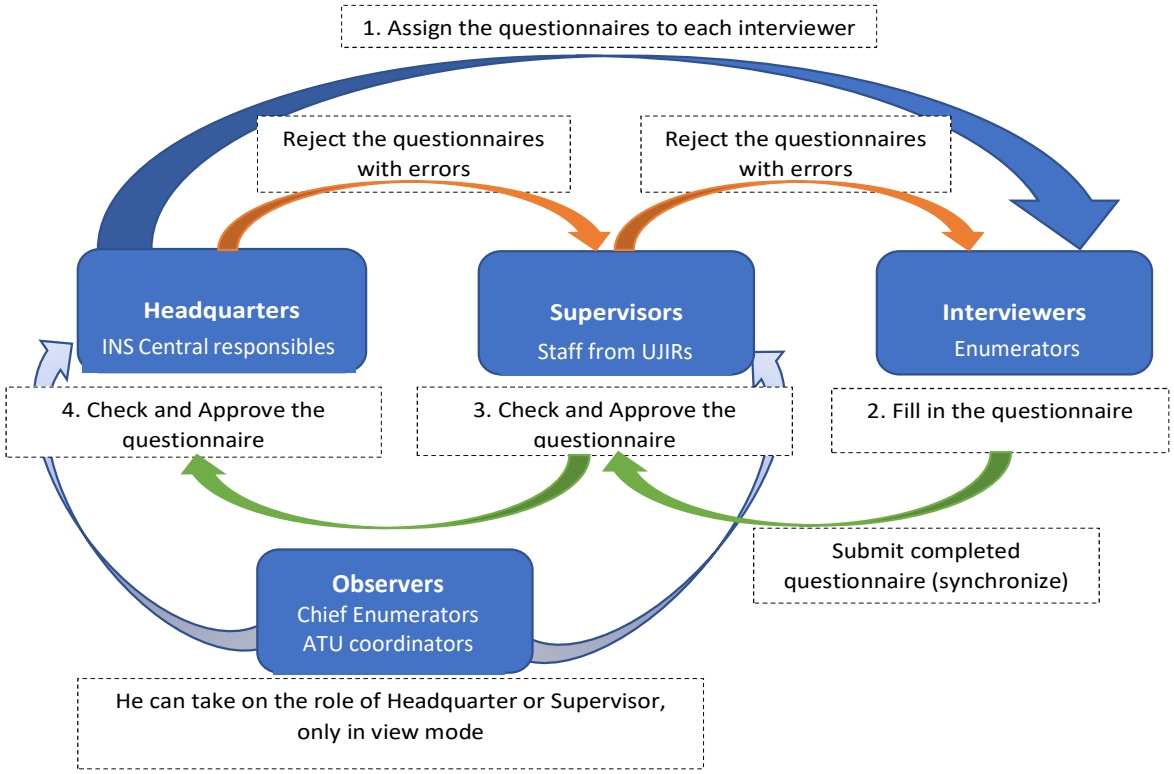
Once the automated validation of questionnaire is done, the respondent can ask for the proof of self-enumeration on a dedicated platform, included in the census portal. If not validated, the person had to restart the process of pre-registration and filling in the census form.

Assistant for self-registration and self-enumeration (ARA)
In the CAWI data collection phase, fixed points were established in rural localities and in some large cities where enumerators were present to assist the self-enumeration, called ARA enumerators, offering help and guidance to persons that want to self-enumerate but do not have access to internet or not have e-mail address or are not handling the tools for this process. The ARA role was, upon request, to guide the head of household/reference person to follow the same and entire process of pre-registration using a dedicated email for CAWI process to receive the links for questionnaires of persons which are providing the PIN and addresses and assist the persons during the data filling. Also, a call center established at the INS headquarters provided information on how the self-enumeration should be done.

The CAPI process of PHC data collection benefits of preliminary quality control checks due to flow designed and attributions allocated in Survey solutions to enumerators (interviewers),

supervisors (staff from UJIRs), headquarters (INS Central PHC team - coordinator of PHC) and observers (chief enumerators, AUT Coordinators) – see Figure 5 below.

Figure 5 – PHC Questionnaire validation workflow in Survey Solutions CAPI



Headquarters (**INS Central** – a dedicated team of INS) have the role of coordinator of entire census, as follows:

- import the questionnaire from the designer platform (Survey Setup)
- create survey assignments (Survey Setup) at the start of the census, but also during the PHC2021 as request of the Supervisors from the county level
- track the overall progress of the census (Reports)
- check the completed questionnaires,
- Approve/Reject the questionnaires (Interviews), but only the ones checked by the Supervisors. The Rejected questionnaires go back to the Supervisors for rejection to the Enumerators, for correction or complete the unanswered questions.
- manage the users of the data collection system that have roles in SuSo, accordingly with the statuses of the assignments, the average speed of the completion and overall job of the teams
- to export the data collected from these assignments (Data Export), which is the data source for quality reports during the census, and in the end of the PHC2021, the source data for post-processing process.

- to find and check the status of particular case assignments (Troubleshooting) during the PHC2021, accordingly with the situation impossible to be solved at the Supervisors level.

Interviewers (Enumerators - during CAPI phase) have the important role of completing the questionnaire, as the face-to-face process of collecting the data. Their role is unique in the entire workflow which has the right to edit/modify the data, so the responsibility is clear and well determined. As an important tool for quality of the collected data, the enumerators have the possibility to add comments to the questions to offer more valuable explanations to the Supervisors or Headquarters for reviewing process of the interviews. Also, the Enumerators should correct the errors which are discovered by the Supervisors/Headquarters in the Rejected Interviews.

Supervisors (staff from UJIRs) check the correctness of the data. The main activities related to the quality carried out by the supervisors:

- request to Headquarters for creation of survey assignments during the PHC2021 as request of the Interviewers based on field activities
- track the overall progress of their teams (Reports)
- check the completed questionnaires,
- Approve/Reject the questionnaires (Interviews), only the finalized questionnaires by the Enumerators. The Rejected questionnaires goes back to the Enumerators' tablets for correction or complete the unanswered questions.

Observers (Chief enumerators, ATU Coordinators of county/Bucharest and its districts/city/town/communes) accesses information through impersonation, looking through the eyes of a different user, like Headquarter or Supervisor without being able to approve/reject questionnaires (read only). In actual census, the observer will read individual questionnaires including any comments left; see the reports; see lists of users and their profiles. The main objective of one Observer is to monitor the activities and to request changes and corrective actions.

P2.0.1. DOP Add GPS

The same services we have already used in the preparation of the geo-database will be used here. We are aware, that there may be occasions when we will not be able to appropriately geo-reference a response, which may either result in misallocation of the census unit(s) or double contact. Respondents confronted with the latter are kindly asked for their understanding.

P2.0.2. DOP Send Email

Given the complexity and dimension of the Self - Registration Process, and its integration into the Survey Solutions CASS, is strongly recommended/propose and INS decided to use Amazon SES, as the email platform. For more details see on the Amazon SES website (https://aws.amazon.com/ses/). People who, for several reasons, did not received the email containing the link to the individual electronic questionnaires, make a call in Call Center and operators check if the system send it or not. They also give advice for possible other reasons for non-receiving the emails to the people.

P2.0.3. DOP Non-Response

Non-response can come in various forms, ranging from complete refusal to an item level non-response. Several measures had been taken to avoid non-response in the first place:

i. The questionnaire has been designed to only contain fully relevant questions, as to keep the response burden as low as possible.

ii. The questionnaire has been designed to be of as little complexity as necessary, again keeping the response burden low.
iii. The timing of the self-enumeration phase has been chosen long enough, so that everybody receives the possibility to fill it.
iv. A census Call Center, as well as local facilitator, should help to reduce any technical response burdens.

With these measures is expected to maintain a high response level already during the self-enumeration phase, and subsequently to the tablet-based enumeration phase. Nevertheless, for cases in which all of the above measures fail and a non-response at any level is the result, the following two sections explain when and how a follow up will take place. Finally, for cases where the result remains a non-response, imputation measures as described in the corresponding section will be applied.

P2.0.3.1. DOP Non-Response CAWI

Partially incomplete questionnaires, for which people did not repeat the self-registration, will be routed to the CAPI data collection.

P2.0.3.2. DOP Non-Response CAPI

Non-responses during this phase will - after careful probing by the Enumerator and accepted by Supervisor and Headquarter - move into the imputation cycle. The enumerator should try to obtain the answers from the people enumerated.

P2.0.3.3. DOP Imputation

Details on the imputation procedure will be applied as per INS decision regarding this task.

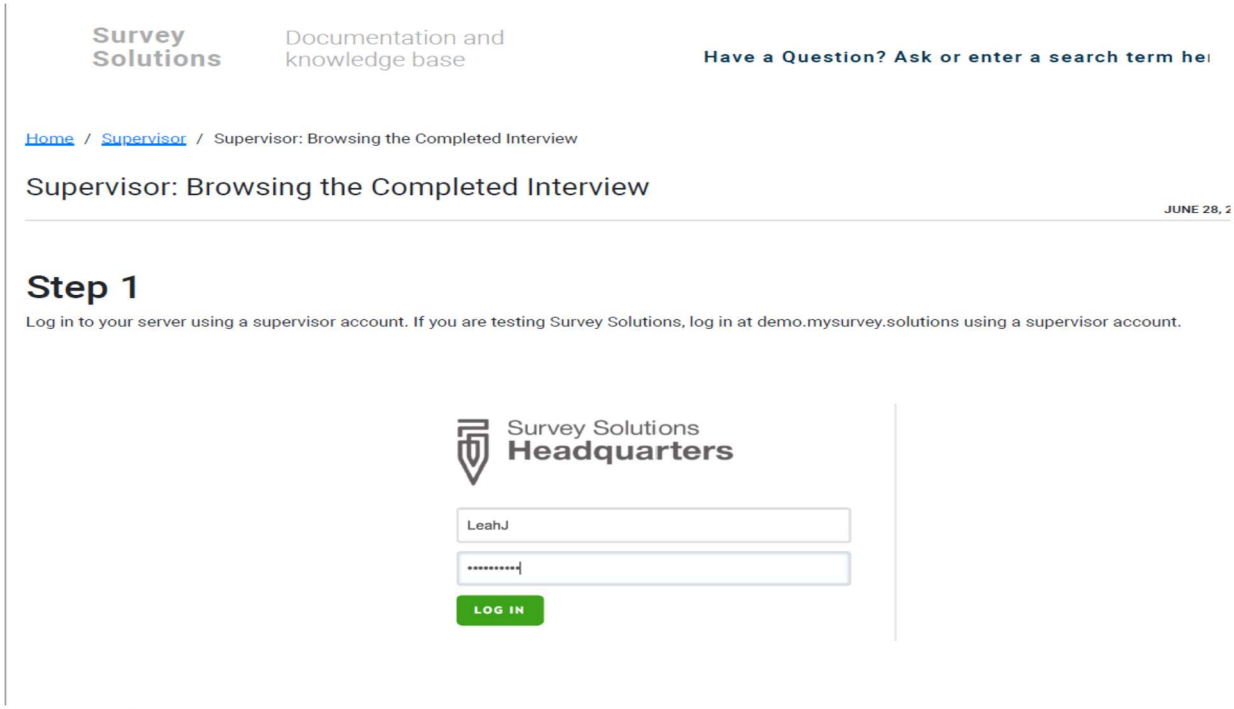
P2.0.4. Identify Dwelling

Identification of the right dwelling inside the structure is a challenging task. This section outlines the procedures to correctly identify the dwellings who responded and who didn't respond, so we do not miss anybody. In some cases, this may not always be possible, and some respondents, already enumerated, may be contacted a second time during the tablet interview collection (CAPI data collection). In such case the respondent should point this out to the enumerator, and after provision of the PIN, the enumeration will be closed.

P2.0.5. Manual Quality Inspection

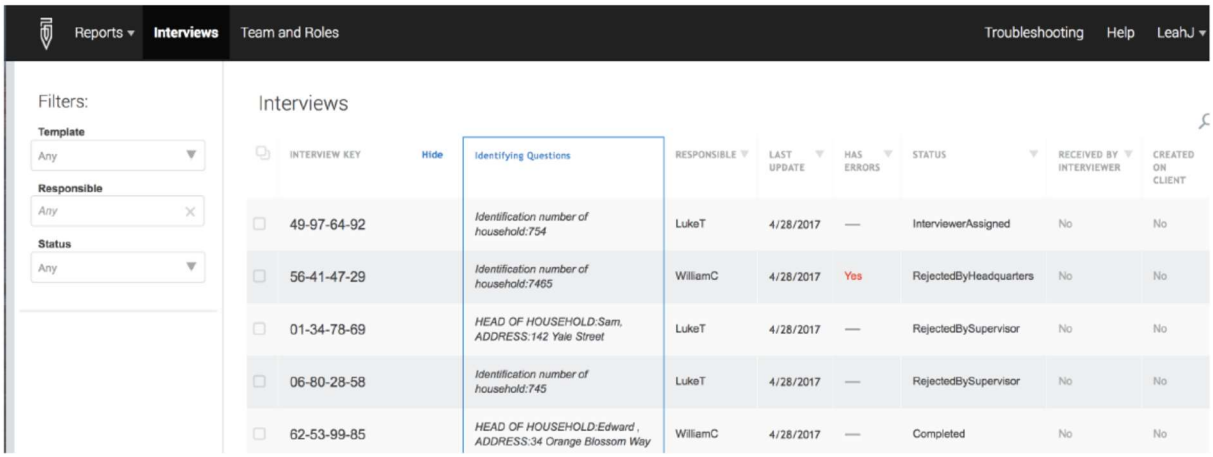
Manual quality inspection will take place when either Internal or External validation checks are not passed as they are established by the INS. The Internal tools provided by the Su So which can be applied for this purpose are described bellow (see the steps Figure 6) by the section included from the support website: (https://docs.mysurvey.solutions/). Internal validation refers to the validations build into the questionnaire, whereas External validation refers to the ones after the individual questionnaire is received, like area related validations, comparison with other databases, response timings in comparison to other questionnaires etc.

Figure 6 - Manual Inspection process



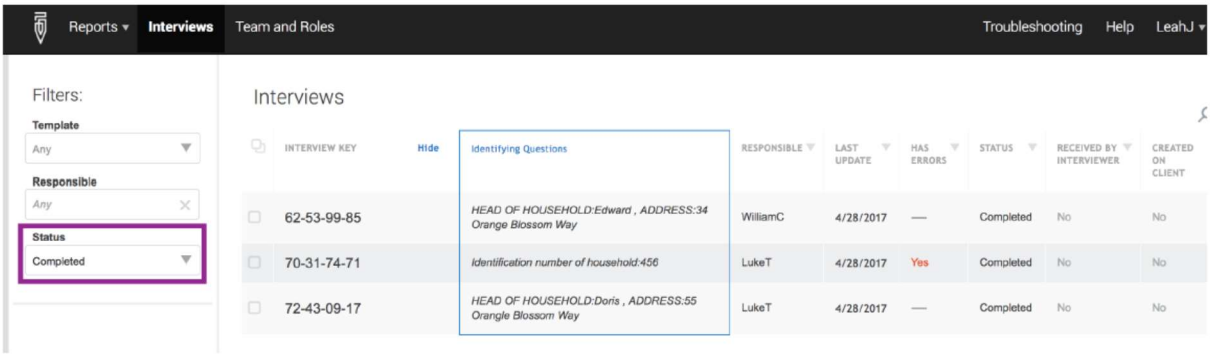
Step 2

Go to the *Interviews* tab



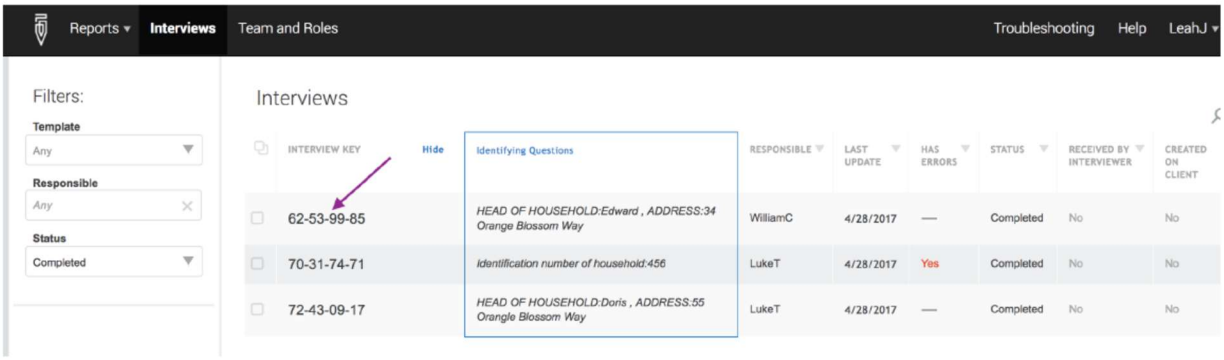
Step 3

From the filter on the left side of the screen, select *Completed*



Step 4

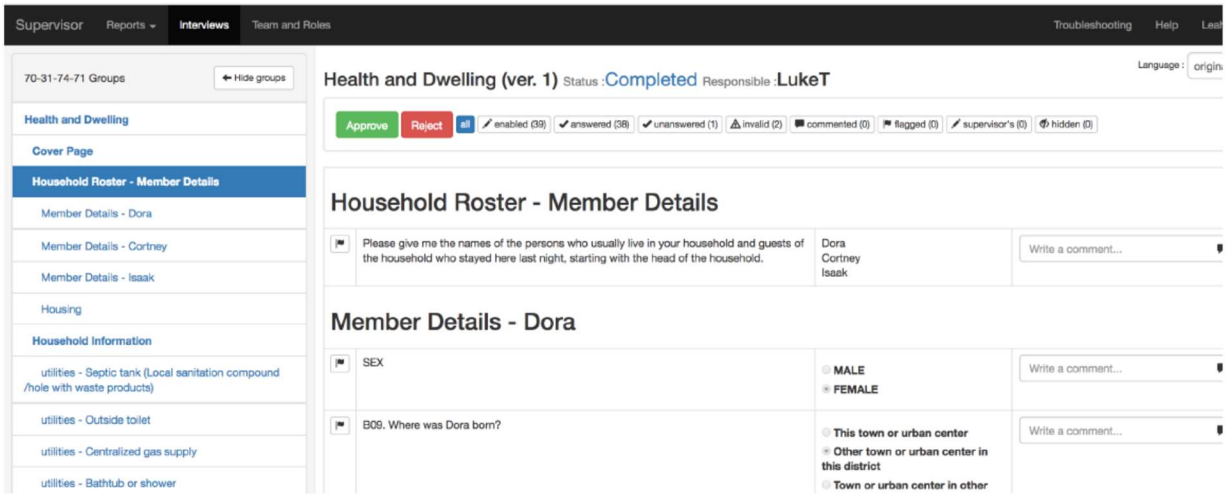
Click on the *Interview Key* for the interview that you would like to browse.



Now you can see all the answers given by the respondents—the answers marked in red are invalid according to the validation rules.

Step 5

If the interview should be approved, click on the *Approve* button. To return it to the interviewer, click on the *Reject* button.



P2.1. DOP PIN

This is the 'standard' approach for the census enumeration process, given data requirements for the transition to a census system based on administrative sources. As estimated, this process will be available to 99% of the population given the widespread availability of the PIN in Romania, even in the foreign resident community.

P2.2. DOP NO-PIN

For members of the household who are foreign nationals and do not own a PIN (persons without PIN), the respondent can directly create a temporary artificial PIN inside the pre-registration form, starting from declaring the sex and the date of birth. The temporary PIN is created adding to the first 7 characters determined on sex and date of birth and 6 characters of figure 9.

P2.3. Mail Back process

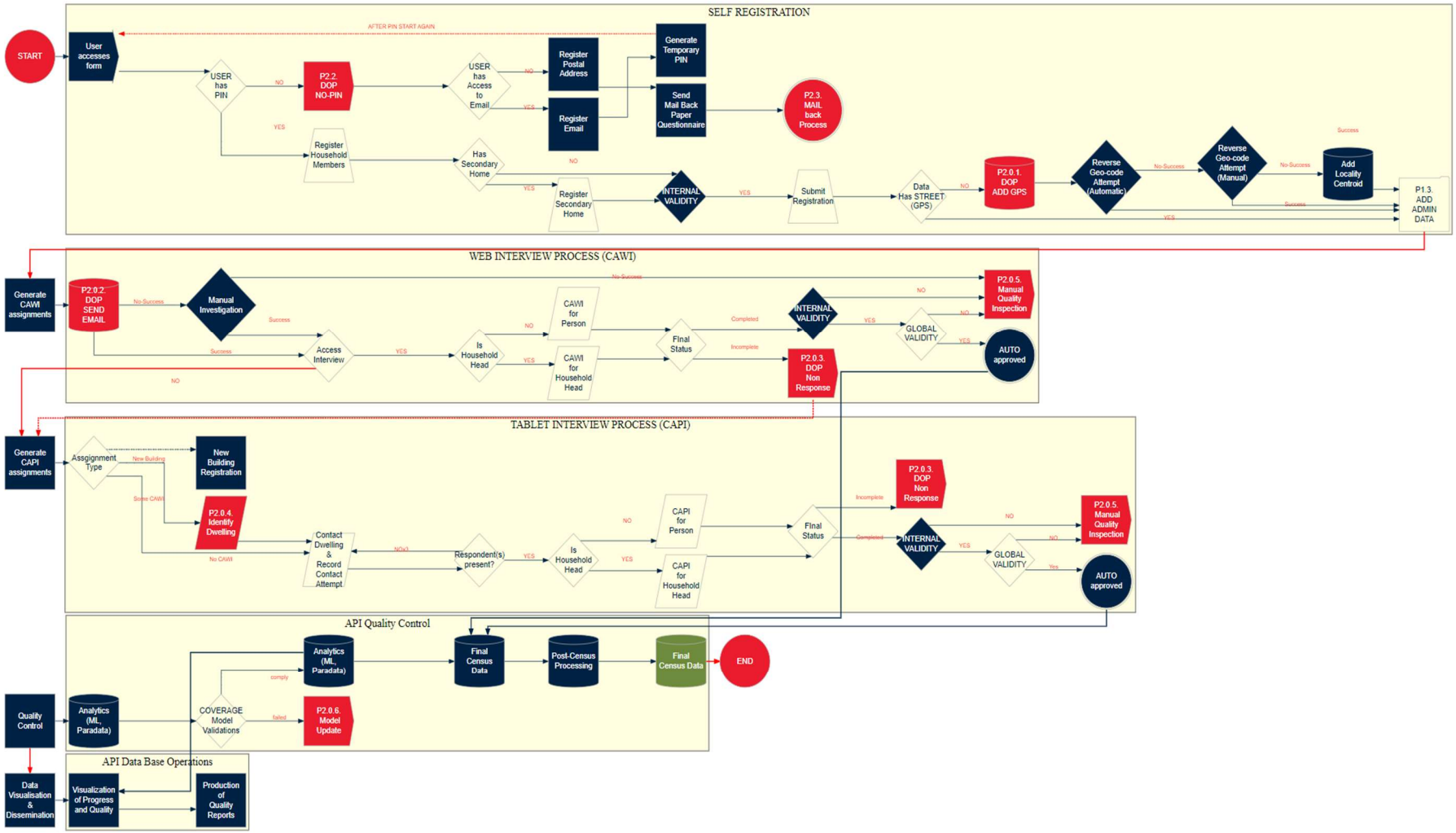
The mail back process was recommended for self-registration, to deliver a maximal inclusive approach for self-registration, however INS decided to capture any individuals who cannot self-register during the CAPI process.

Based on the data retrieved from the 2021 pilot census, an initial model was defined. To further increase the precision of this model, it will also use data throughout the census, to update the model. This is the process group covering the actual data collection process. It can again be separated into 4 major sub-groups:

1. Self - Registration
2. Web Interview Process (CAWI)
3. Tablet Interview process (CAPI)
4. API Quality Control (including DB operations)

The completely integrated workflow for this process group is depicted in the following diagram.

Figure 7 - Integrated workflow



3. Special Purpose Applications for coverage check

The PHC2021 use Special Purpose Applications (SPA) in support of data collection, monitoring, evaluation, and statistical production. The SPAs are tools developed and tested initially for pilot paradata available on Survey Solutions and now are installed on '*Data Production Server'* at xx.xx.xxx.xx/phc_paradata/ (available for INS staff; no credentials are required). The SPAs are designed to analyze the census paradata (from both phases CAWI and CAPI) exported from Survey Solutions and hosted on '*Data Production Server'* path **/arhiva1/rpl2022/exports/** and are envisaging different layers, respectively:

- National - summary for questionnaires & interviewers & segments -xx.xx.xxx.xx/phc_paradata/
- Counties - summary county by segment/interviewer xx.xx.xxx.xxphc_country/)
- Segments - summary segment by building location, xx.xx.xxx.xx/phc_segments/

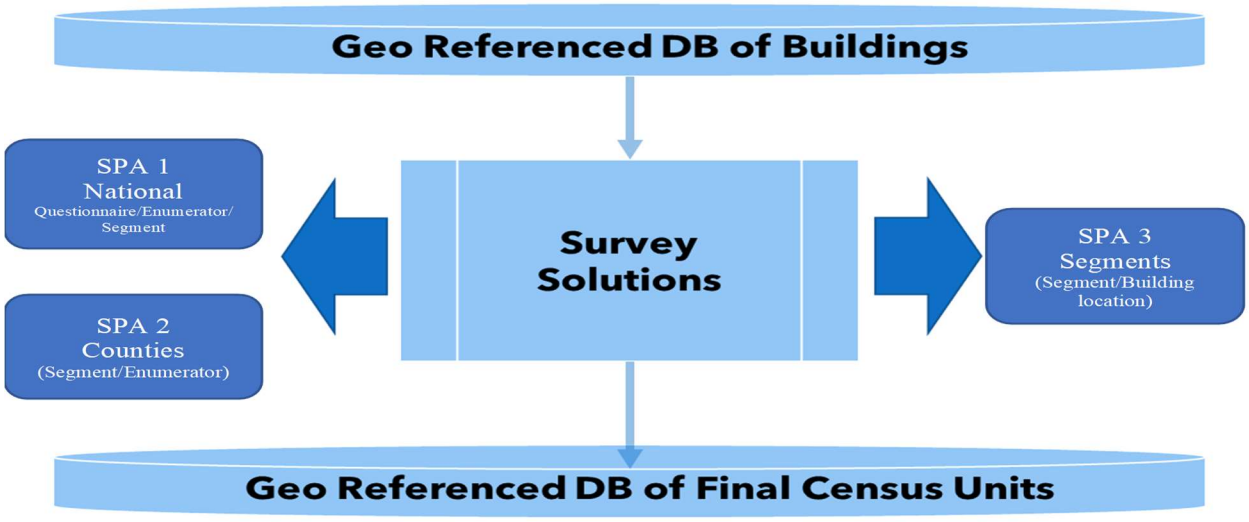
The dimension of census exported paradata is considerably large (above 120Gb unzipped/phase):

- cawi_rpl2021_04032022_v9_1_Paradata_All_Final.zip (file dimension 25.824 Gb)
- rpl2021_capi_12052022_v1_1_Paradata_All_Final.zip (file dimension 22.179 Gb)

and the run of SPA's could take very long time to calculate the indicators and produce the reports. Alternative options for faster results could be the use of scripts prepared for paradata (available on '*Data Production Server'* path home/calex/RPL_CAPI) or the breaking of paradata by counties and producing 42 separate files that can be analyzed much easier but referring only to counties' parameters. Nevertheless, the results will support future developments of nationwide surveys.

The SPAs places into the applications' schema are presented in figure below.

Figure 8 - Main Data Stream and SPAs



Source: Authors

SPAs (Special Purpose Applications) are built on top, and follow all conventions defined by the main data stream:

- connect through the API
- are located on different server
- require the main SURVEY SOLUTIONS flow, since they are calibrated to it

- Survey Solutions does not require them.

The data stream consists of:

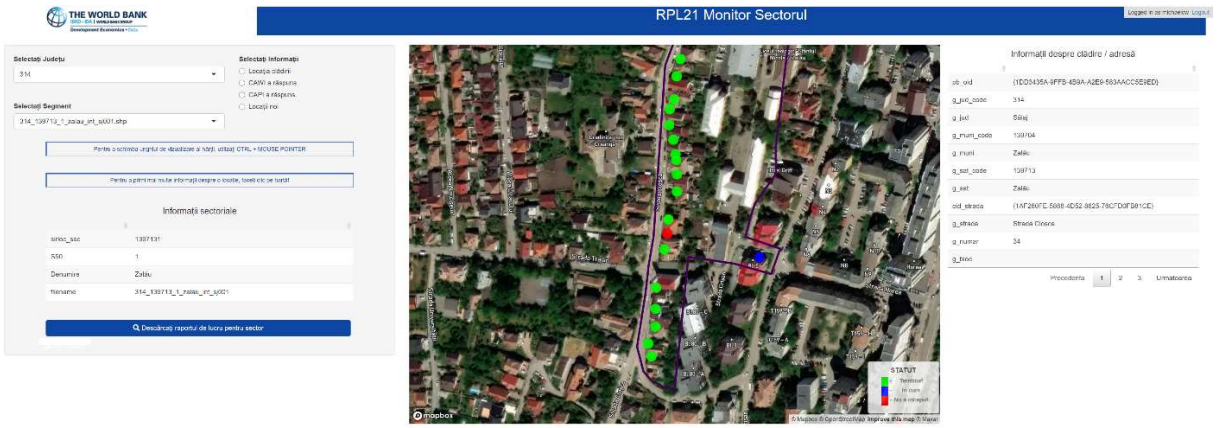
- Geo-referenced DB of buildings constitutes the starting point.
- all location-based standards and classifications need to be maintained (i.e., SIRUTA Codes)
- all spatial ID variable names ('schema') need to be harmonized and are frozen (i.e., no spontaneous changes) until after the census - maintaining this standard in subsequent years is recommended.
- a fully geo-referenced DB of final census units is the result

3.1. SPA Segments - summary segment by building location

Tools Directory **../phc_segments -** support the reports for a selected segment regarding interviews collected via web-mode, via tablet, count the addresses and dwellings in the database - see figure below.

The application is already available in Romanian, and allows to select the administrative unit, and subsequently the required sector. After selection it will display the summary statistics for the corresponding sector, as well as produce a more detailed report as a MS Word document, with the variables described below. The main user group for this app are the group of supervisors, directly communicating with the enumerators.

Figure 9 - Sector tool



Report from **../phc_segment**

.	N_adresa	N_locuinte	CAPI	CAWI	TOTAL
TOTAL	51	51	42	1	43

srn	adresa	N_locuinte	CAPI	CAWI	TOTAL	STATUT
1	Strada Sintandrei 375	1	1	0	1	Terminat

Explanation:

N_adresa	Count of addresses in the database for the selected segment
N_locuinte	Count of dwellings in the database for the selected segment
CAWI	Count of INTERVIEWS collected via web-mode
CAPI	Count of INTERVIEWS collected via tablet
STATUT	Status of the building, one of not-started/in progress/completed

3.2. SPA Counties - summary County by segment/interviewer

Tools Directory **../phc_country -** support the reports on sum of count of addresses and dwellings in segment files, count the interviews collected via web-mode and via tablet (sums), and average duration in the segment – see the figure below.

This application allows the selection of whole counties, and provide summary statistics for it. Its main user group is the group of Survey Solutions headquarter user or any other high-level observer. The application can be run with central access and should only have a small number of users. Subsequently it also provides progress reports for each of the districts downloadable in word format.

Figure 10 - Country and county level detail

Reports from **../phc_country**

tempgr	N_adresa	N_locuinte	CAPI	CAWI	TOTAL
TOTAL	1201	2051	1133	43	1176

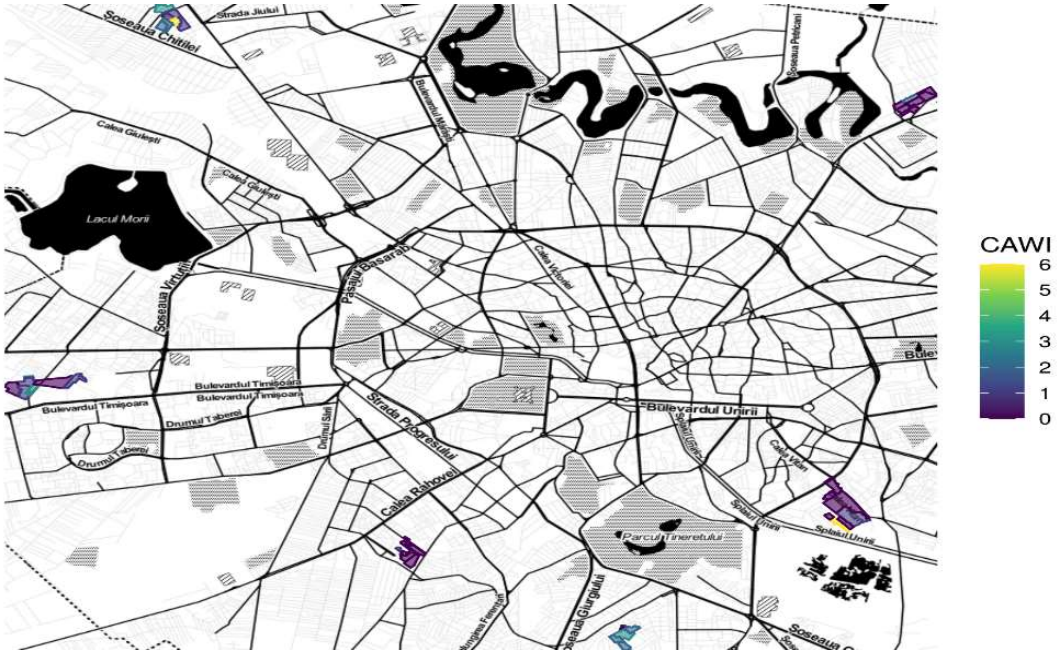
g_jud_code	sector	g_sat_code	N_adresa	N_locuinte	CAPI	CAWI	TOTAL
403	1	179141	1201	54	40	3	43

_responsible	g_jud_code	sector	g_sat_code	CAWI_dur	CAPI_dur	Terminat	In_curs	Nu_inceput
int_s1001	403	1	179141	24.00	3.57	30	10	0

Explanations:

N_adresa	SUM of Count of addresses in segment files
N_locuinte	SUM of Count of dwellings in segment files
CAWI	Count of INTERVIEWS collected via web-mode (sums)
CAPI	Count of INTERVIEWS collected via tablet (sums)
T../In../Nu	Count of the building, one of not-started/in progress/completed

Figure 11 – CAWI average duration in the segment



CAWI_dur - Average duration in the segment

CAPI_dur - Average duration in the segment

Explanations: the colors show only the duration for CAWI, since in this specific area only CAWI was collected.

3.3. SPA National, Paradata for questionnaires, interviewers & segments

Tools Directory **../phc_paradata -** is supporting the paradata reports which contain the exact calculation of duration etc. based on the individual response timings in seconds (from the questionnaire paradata). See figures below.

The main purpose of this application is to provide a direct live view of all generated paradata, for both CAWI and CAPI phases. It identifies outlying units, which are particularly useful during CAPI data collection, as well as produces a report containing a detailed overview of the data collection process. In addition, it also allows to transform and download the paradata into a more convenient format, and as a .csv file, such that the data can be analysed with other software packages, like SPSS.

Figure 12 - Load data into paradata SPA

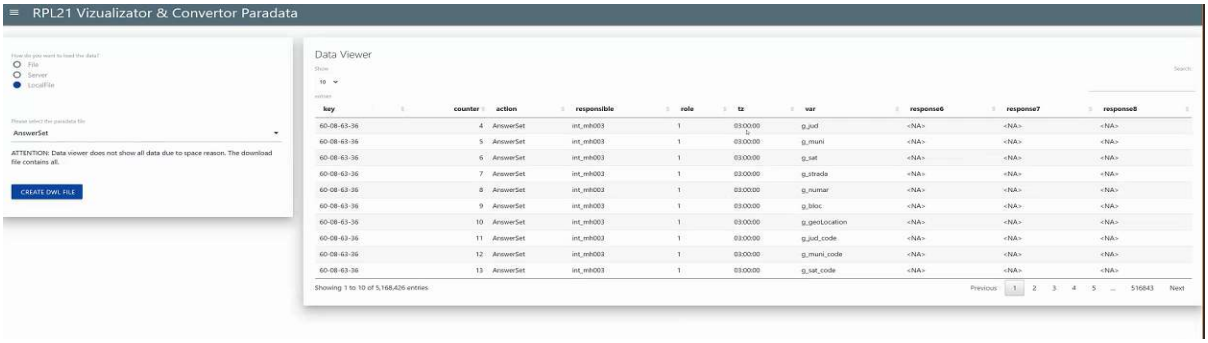


Figure 13 - Paradata Questionnaire/Question summary and identification of outliers

Figure 13 shows (from left to right): Average timing by question (calculated over all completed interviews), total interview duration for the 10 fastest and 10 slowest interviews, Number of invalid responses for the 10 interviews with the highest number and for the 10 interviews with the lowest number.

Figure 14 - Paradata Enumerator summary

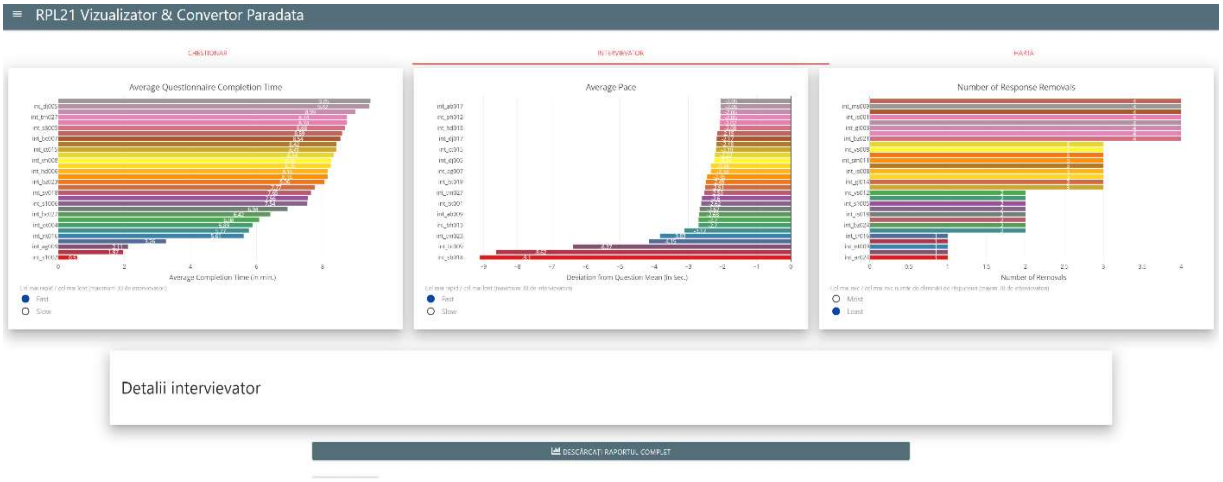


Figure 14 shows (from left to right) the average questionnaire completing time by interviewer, the average pace of each interviewer (calculated as deviations from overall question mean), and the total number of response removals.

Reports from **../ phc_paradata**

.	**mean_duration**	**mean_durationNOBREAK**	**mean_startHour**	**mean**_RespTime	**N_obs**
Toată România	59.82	11.32	10.68158	5.58	961

Explanations:

mean_duration (min)	- mean duration of interview, calculated from 1st ANSWER SET TO LAST
mean_durationNOBREAK (min)	- mean duration of interview, calculated from 1st ANSWER SET TO LAST but without BREAKS (breaks as signalled by tablet, or when response time >3min), where both long questionnaire and short questionnaires are included.
mean_startHour (24h)	- mean of the hour when most of the interviews did their first question
mean_RespTime (sec)	- mean of the average response time to all questions in seconds
N_obs	- total number of enumerators

responsible	mean_duration	mean_durationNOBREAK	mean_startHour	mean_RespTime	N_obs
int_bt008	46.84	7.50	13	3.78	56

Explanations:

responsible - enumerators's username

N_obs - total number of interviews completed by the enumerators

.	Av_ResponseTime	**Av_Duration**	**N_questions**
Toată România	8.190066	8.190066	151

Explanations:

Av_ResponseTime (sec) - average response time by question in the segment (excludes breaks)

Av_Duration (sec) - average duration for interview time in the segment

N_questions - count of represented questions

counterMedian	var	Av_ResponseTime	Av_Duration	N_obs	tot
152.5	AA_ALM	13.28	11.96	1075	Toată România

Explanations:

counterMedian - position in the interview process for the median of interviews. (explanation of calculation: not all questions are asked 1, 2, 3, 4. in every interview, for various reason. To get a common ranking, was taken the position it represented in 50% of the interviews. All data is updated continuously with every new data read.)

var - the name of the variable

Av_ResponseTime (sec) - average response time in the segment (excludes breaks)

Av_Duration (sec) - average duration time in the segment

N_obs - how many answers received for this variable in the survey process (only the final answers)

4. Statistical support for data quality

The report on PHC pilot (Output 10a) presented several recommendations regarding preparedness for actual census and ones related to data quality and coverage and data security (see section 3.5 - Main working areas for improvement; section 4 - Recommendations for actual PHC go-live production (2021)). Beside these recommendations an important aspect regarding data quality and coverage consists of statistical support for data quality, methodological handbook and guide, training materials for statistical staff and census staff, prepared for self-enumeration and for CAPI method delivered to enumerators and statistical staff.

The quality data starts at very beginning of the census, from the preparation of the data collection process, and an important role comes to the training of the involved personnel. The Enumerators, Supervisors, Observers and Headquarters have different role, accordingly the training materials and handbooks developed and delivered in training sessions.

The training materials, handbooks, and video-presentations by roles developed under Survey Solutions platform and by INS, are the following and are available on public sources:

- Enumerator:
  - Questionnaire - generated from Survey Solutions
  - Enumerator's handbook
  - Question types - video presentation
  - Basic elements in using Survey Solutions, for Enumerators - video presentation https://drive.google.com/file/d/1-6y8sMu6878LPx0Yuo_yyzmwxv7muOVb/view?usp=sharing
  - Error messages
  - Interviewer/Enumerator Training Series - Youtube channel https://www.youtube.com/playlist?list=PLIjqNDszKtS7MWAzTjaFDG7c8WNFDSpAE
- Supervisor:
  - Survey Solutions Platform User Guide - Supervisor https://drive.google.com/file/d/1gXur5CWh0qb4BVdxx7wS1BtHu7ZXNHAs/view?usp=sharing
  - Supervisor Training Series - Youtube channel https://www.youtube.com/playlist?list=PLIjqNDszKtS5Mv4YCiJPDV3OOvjB8YX8I
- Observer:
  - Survey Solutions Platform User Guide - Observer https://drive.google.com/file/d/1uVl7WBTYfi8ZDOsYA1SV_Cxq8QMsI6oy/view?usp=sharing
- Headquarter:
  - Survey Solutions Platform User Guide - Headquarter https://drive.google.com/file/d/1wkI5pDTLgEqBleFMp0CX_x26ZZLyn871/view?usp=sharing
  - 2FA Autentication mode (Activare_autentificare_in_doua_etape_2FA_Survey Solutions.pdf)
- Data control and validation on the tablets: https://drive.google.com/file/d/1-FWjpBXhOd8jlH3W-pKJrWYm3XA8RYd5/view?usp=sharing

The most important training materials and handbooks are published on a dedicated page of the INS website.

5. Data protection and security

The information from this section is detailed in the report of *Output 10c: Report on advisory services provided to Recipient on the Set of draft statistical census tools using multi-modal methods to promote data protection and security* and complemented with Output 10b: *Technical assistance and best practice recommendations for SDC, data confidentiality, ways to secure micro-data and aggregate data.*

Data protection relies on implemented controls for both data in motion and data at rest[5]. Information circulates from database servers to enumerators' tablets and back. Also, for self-registration component, data is posted via browsers on the web servers, then passed to application servers. Data at rest on tablets is totally encrypted (outside of application), and by that the data is protected in the case of potential incidents regarding the tablet (destroyed, lost or theft) and only the application can decrypt data, once authorized users are accessing it. Also, in case of lost or theft the user of the application can be locked. The communication between tablets or browsers and application servers is also encrypted using SSL. From application servers to database servers everything is within a physically protected zone, so no encryption needed (potential cross-site scripting and SQL injection were subject of the STS security assessment – see the report attached to Output 10c). In addition, the whole Survey Solutions database is password protected, and any data export can only take place with corresponding qualified credentials.

Personally identifiable information (PII) is handled in the information cycle for PHC in both main phases: CAWI and CAPI. For CAWI, the person will use the self-registration portal to register, then data is transmitted to Survey Solution component in order to generate interviews. The person will receive a registration code to be used for support services and interview authentication. In the case of PHC was used only the Google ReCaptcha but still, even without a code, the STS security assessment was not considered this behaviour as a vulnerability. Interviews are accessed using a unique generated link which is transmitted by e-mail to registered person. The e-mail service is hosted by Amazon Simple E-mail Services on a Central Europe zone in order to comply with EU data export regulations. The e-mail messages are designed so no PII information is transmitted through e-mail service, by using masking techniques (for example only first 5 digit and last 3 digits of personally identifiable code are shown in the message). In order to access the interview (questionnaire), the respondents are asked for a captcha mechanism (see the PHC actual implementation above). Once interview (questionnaire) is completed, the interview link became unusable.

For CAPI phase, data is collected only by enumerators on tablets. Interviewers are instructed in order to capture PII according to national and EU regulations. Information is stored on encrypted container within tablets, then transmitted to application servers using encrypted channel. Then interview data is subject to approval flow within Survey Solution component. For Survey Solutions component, for system users generic name is used (for example, INT_AB_1). Other users accessing the system (supervisors, observers, headquarters) are also using generic account names, and their actions are tracked within audit capabilities of Survey Solutions component.

[5] https://en.wikipedia.org/wiki/Data_at_rest

The data processing and dissemination take place in INS hosted environment and all the security and data protection relies on INS' organization procedures (included VPN solution), and also the overall INS hosted infrastructure security.

The system physical security relies on security controls implemented on STS datacenters and INS and WB internal procedures and regulations for users accessing the system from terminals hosted by either INS and WB (for details INS-STS collaboration protocols should be consulted; being internal, are not available for public).

The system logical security has different levels of protection:

- STS Network firewalls are protecting access to the systems hosted in STS datacenters.
- All machines have implemented hosts firewalls which allows only minimal traffic required for system functionality
- Windows Severs are protected by Windows Defender native anti-virus. Servers are connected to Internet for outbound communication and security updates are applied automatically.
- Administrative access to the system is protected by restricted VPN point-to-site solution, available only to authorized personnel (INS staff has access).
- Access to self-registration component in CAWI phase is publicly available. Information is validated in the registration page and captcha mechanism (Google Re-captcha service) is used in order to deny programmatic access.
- Access to web application is possible for the population for Survey Solution component in CAWI phase by using the registration custom link received at the end of self-registration, so access to interview data is protected with a unique identifiable information. The link is valid only until interview is completed.
- Access to web application is possible only to authenticated users for Survey Solution component in CAPI phase. Username and password are mandatory for accessing system functionalities, and privileged users are additionally protected by 2 factors authentication using TOTP. Also, a captcha mechanism (Google Re-captcha service) is used in order to deny programmatic access.

The overall security of the production system was subject to STS security assessment according to best practices and national regulations and compliance. The PHC was conducted when STS has reviewed the setup and confirmed its compliance with all applicable restrictions and regulations. Based on the assessment results, the possible vulnerabilities and overall security related findings were addressed either via security updates or system configuration changes. The report of IT data collection system audit performed by STS was provided and is available at INS (see also the Output 10c, also).

Competence makes a difference!

Project selected under the Administrative Capacity Operational Program, co-financed by European Union from the European Social Fund