ROMANIA

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

OUTPUT No. 13

Report on advisory services provided to Recipient on the Recommendations and guidance on documented plan for the integrated system for the PHC 2021 implementation (including user management, ID encryption, disaster recovery, maintain system availability in case of breakdown)

January 2023

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 January 2023 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 corresponds to Output 13 under the above-mentioned agreement.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ARA	Assistant for self-enumeration
ATU	Administrative Territorial Units
CAPI	Computer-assisted personal interview
CAWI	Computer-assisted web interview
CNP	Personal Identification Number
CONRENA	Project the Consolidation of the National Statistical System and modernization of the statistical production processes for carrying out national censuses
CRS	Coordinate Reference System
DB	Database
DTS	Territorial Statistics Directorates
EPSG	European Petroleum Survey Group
ETRS	European Terrestrial Reference System
EU	European Union
GIS	Geographical Information System
GPS	Global Positioning System
INS	National Institute of Statistics
INSPIRE	Infrastructure for Spatial Information in the European Community
IP	Internet Protocol
IT	Information Technology
LAEA	Lambert Azimuthal Equal-Area
PIN	Personal Identification Number
PHC2021	Population and Housing Census, round 2021
RAS	Reimbursable Advisory Services
SICCA	Solution for the Architecture and Computation of the Costs of Products and Activities of INS
SIRUTA	Information System of Administrative Territorial Units Register
SDC	Statistical Disclosure Control
STS	Special Telecommunication Service
VPN	Virtual Private Network
UI	User Interface
WB	World Bank

Introduction

The purpose of this report is to present the ***recommendations and guidance on documented plan for the integrated system for the PHC 2021 implementation (including user management, ID encryption, disaster recovery, maintain system availability in case of breakdown).*** 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 Output is the result of support provided to the National Institute of Statistics (INS) in implementing the Population and Housing Census (PHC), round 2021, and for documenting the data collection and production system for this major statistical works of social statistics.

Romania has implemented for the first time a fully digitalized (100% paperless) PHC (March 14 – July 31, 2022). That has enabled the efficient data collection for 95.4% (18.15 million) of the estimated population of Romania through a dedicated online environment and has ensured the avoidance of any loss of the collected data due to manual or machine reading faults inherent to the traditional paper and pencil census.

The knowledge and experience gained by the INS during the current PHC lays the foundation for future surveys implementation in social statistics and the transition to intercensal statistics for population and housing in annual registry-based processes. The logistics related to data collection according to the optimal combination identified (CAWI-CAPI) and implemented during PHC is reflected in this report consisting of seven sections.

The overview of the flow of the PHC integrated system, the first section, presents the main phases of the data collection processes organization, using the physical and online/internet environments (INS and STS IT environments, solutions and applications, procedures) and the main tasks of the institutions responsible, which allowed to exploit the data sources for the census purposes.

The second section presents a synthesis of the data flow for organizing the data collection at all stages of the PHC, including details regarding the processes of gathering data from administrative sources, and the data collection based on logical sequences within the CAWI and CAPI phases.

The PHC data collection system, presented in the third section, exposes its overall architecture (hardware, operation systems, software applications, databases and electronic services), the high availability structure, performance design, the back-up, and the security and testing process for data collection system. The detailed configuration and installation of the entire system is described in the as-built report as an annex (including the code for commands) dedicated to the IT specialists and system administrators of the INS.

The reports for monitoring the data collection of the PHC are presented in the fourth section, which provides information on the process and samples of the reports for the daily monitoring of data collection during the CAWI and CAPI phases, overall and by counties, and the reports regarding the activity of the census staff. The templates of the monitoring reports and the scripts used for calculating the volumes of questionnaires and the payment amounts are provided in the annexes and archived in electronic format.

The PHC data collected was processed during enumeration or post-enumeration for validation, error correction, deduplication, imputation of missing or incomplete data, and for the calculation of statistical indicators included in the production of the census results. The scripts applied for the

validation of the data collected during the CAWI phase and the ones for the validation and approval of the data collected during the CAPI phase, part of the technical assistance provided, are delivered as annexes to this report and archived in electronic format for further installation and running by the INS staff. However, the scripts are installed on the INS IT environment (*'Data Processing server'*). The processing steps are described in the fifth section.

The INS disseminates the PHC results and has considered as useful for further users of the PHC data to automatize access through an application which is presented in section six. The PHC data dissemination comprises three specific solutions: data hypercubes, mandatory to be published on the Eurostat portal; the grid distribution of the PHC data, which allows the dissemination of 21 indicators (13+8) on grids (INSPIRE) of 1sqkm for the entire Romania's territory; and the tabular dissemination of the PHC results. While the INS team will take care of the data hypercubes by preparing the IT environment and installing the necessary applications provided by the Eurostat to retrieved data from the repository of the PHC data organized on the INS IT environment, this section presents the grid distribution and the dynamic tables developed under the project. The structure, the functions, the data inputs, outputs and flows describe the applications developed. The applications are installed on the INS IT environment (*'Data Processing server'*) for further running by the INS staff.

The seventh section of the report presents a series of conclusions on the PHC, round 2021. The PHC has represented a major statistical event for Romania and the first paperless census, which was prepared and implemented in a joint effort of the INS and other public authorities during the past five years. This in turn has required substantial technical assistance under the ongoing RAS delivered through fifteen (15) outputs, out of which ten (10) represented direct assistance dedicated to the PHC and five (5) were cross-supporting other components of the CONRENA project (the GAC, the intercensus, and the SICCA).

The supporting tools for data production and for the sustainability of the processes and applications deployed for INS's use are presented in section eight of the report. Additional information can be found in other reports delivered under the RAS and referenced in the sections of this report.

1. Overview of PHC integrated system

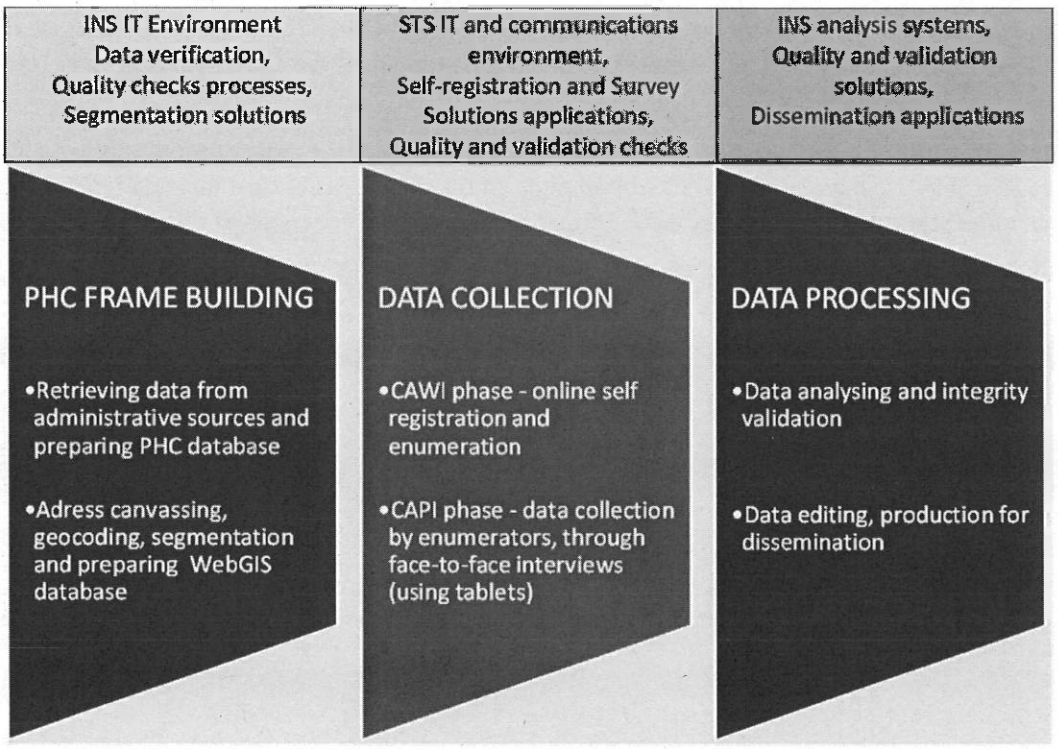
The PHC round 2021 has been the largest statistical operation conducted by the INS in the past decade for collecting, processing and disseminating data on the structure of the Romanian resident population and housing. The purpose of the PHC has been the full enumeration of the population, dwellings, housing units and households which existed/lived in Romania on December 1st, 2021. The PHC2021 required and applied optimized instruments, respectively for:

i. preparing and building the census framework
ii. full electronic data collection in two different modes, CAPI and CAWI including complete geo-referencing of all the collected data and a continuous quality management applied at each stage of the data collection process
iii. processing the collected data and producing statistics for dissemination

Corresponding to the assumed purpose of the PHC, the INS and other responsible counterparts have put in place administrative procedures and instructions, physical and logical infrastructure and communications, and controlling tools for the preparation and implementation of data collection.

The integrated system designed and implemented (see Figure 1) fulfilled the expectations of the data collection process and ensured a very high rate of data collected from 18.15 million persons (95.4% of Romania's residential target population), with a consistent list of variables and with qualitative and timely results.

Figure 1 - PHC integrated system process



The data collection process has been organized physically and in online/internet environments supported by INS and STS to allow the use of data sources for the census purposes. The

administrative data sources gathered from public institutions were organized within the INS environment. The collected data from the residential population was organized through self-registration and Survey Solutions applications in CAWI and CAPI phases using an online/internet environment. The census data was hosted and secured using the STS data collection systems and the statistical production was and is ensured, until the results publication, through analysis systems organized within the INS environment.

The phases related to data collection were and are ensured by data quality control, safety and security physical and logical processes described in the reports of several outputs delivered under the same RAS (please see Output 4.1b, Output 7a, Output 10a, Output 10b, Output 10c).

The running of the actual PHC was based on the set up defined in Output 4.1c: *Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out)*.

The institutional roles and primary responsibilities for the PHC integrated system set up were distributed as follows:

INS
- Ensure resources for hosting, contracting, installation, configuration and acceptance of INS IT integrated system
- Ensure resources for exploitation of integrated IT system previous, during and after PHC2021
- Provide network connectivity via Optical Fiber Channel
- Provide acceptance criteria and administer acceptance tests
- Provide support to the performance testing process

STS
- Provide hosting and Virtual Machines for software application, Database and Load Balancing servers for Data Collection Activities
- Provide network connectivity via Optical Fiber Channel and VPN between STS and INS
- Configure access from the application load balancer to the Internet
- Configure FQDN (Fully Qualified Domain Name) on the insse.ro domain
- Ensures the interconnection with the population database for Personal Identification Number (CNP) verification

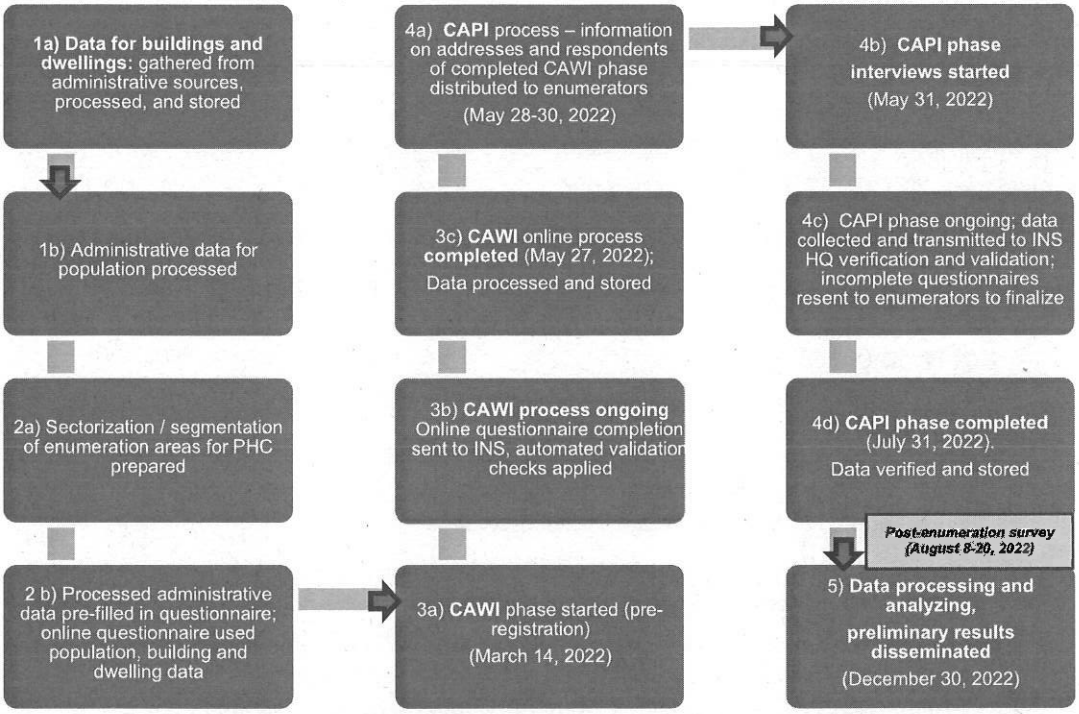
WB
- Act as overall system integrator and offer recommendations and guidance
- Manage process of installation, configuration and testing of the applications
- Installation, Configuration and Optimization of PostgreSQL Database server
- Installation and configuration of Survey Solutions application servers, the application load balancer, the monitoring software, the testing toolset, the Self-Registration Portal
- Conduct performance testing process, analyze results and provide recommendations
- Provide as built documentation for the applications installation

Thus, the INS, STS and WB teams were involved continuously in cooperating, monitoring and running the actual census phases, respectively on providing and ensuring the complete functioning of PHC integrated system, 24/7 during March 14 to July 31, 2022.

2. Data flow on all stages of PHC

The process and data flow were implemented and worked as planned in the report of Output 4.1c: *Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out)* see Figure 2 below.

Figure 2 - Process Diagram for actual PHC



The process is explained further in this section and is complemented with the quality assurance framework applied for data collected and statistical tools to promote data protection and security (see detailed information in the reports of Output 7a: *Methodology to assess and promote continuously and in real time the quality and coverage of the collected data PHC2021 and data protection/security;* and Output 10c: *Set of draft statistical census tools using multi-modal methods to promote data protection and security).*

The data collection process has been coordinated and implemented by the INS and the central and local authorities. The process considered the use of the available data for population and buildings, from the collection process per se (CAWI, CAPI) but also to maximize the information provided by data from administrative sources.

2.1. Data flow of data gathered from administrative sources

This was a preliminary data collection for population, buildings and dwellings. The activity was done by the INS team in the preparatory period of the census. The characteristics for most of the buildings and dwellings, including the GIS coordinates and the addresses of the buildings, were collected, processed and stored in an initial database (see details of the WebGIS preparedness in the report of Output 7a). The buildings and dwellings, especially the addresses and geo-data, were needed in their own right and were used as a starting reference for the enumeration of population.

The sectorization process for creating segments (enumeration areas) of the census population/area was realized and verified by the INS and DTS staff and was meant to facilitate census logistics and sampling activities.

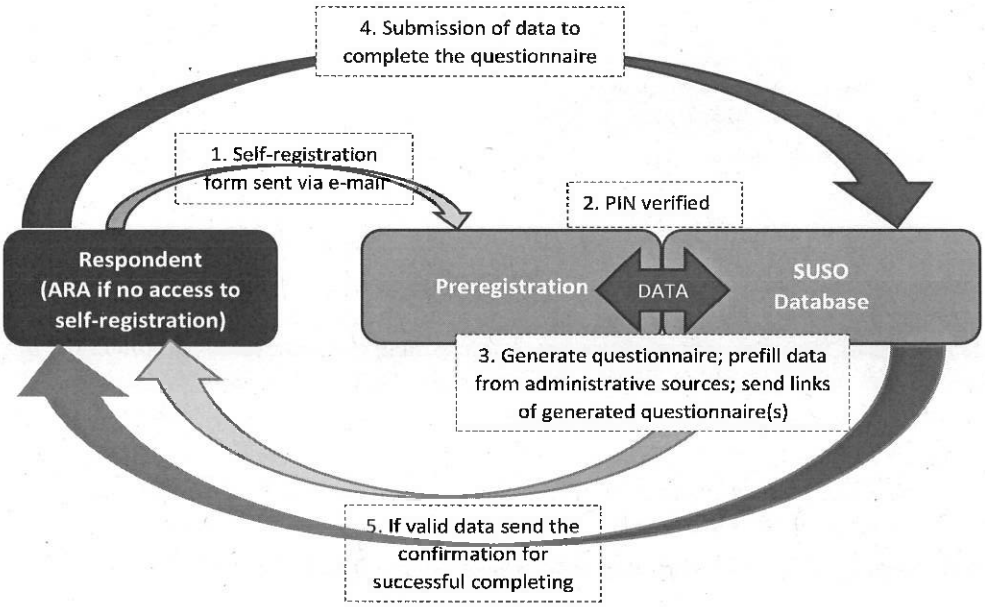
Data for population from administrative sources was processed for the pre-filled process of personal questionnaires into the Survey Solutions data collection system, and also for the validation of the collected data in the post-census processing phase. Part of the data for buildings and dwellings characteristics was collected from one member of the household. The CNP of the persons was transformed into an Unique Statistical Identifier by an irreversible HASH encryption function.

2.2. Data flow on CAWI

Pre-registration - the respondents accessed the pre-registration application hosted on internet and the head of the household fills in his CNP and address of household/households, the e-mail and phone number and the number of household members, and by case their CNP and e-mail. The head of the household also has the possibility to declare the address for the secondary dwellings. A confirmation e-mail of pre-registration sent by insse.ro domain, containing the link to the individual questionnaire/form and instructions for completing the questionnaire was received by respondent. The CNP is encrypted once pre-registered, and the Unique Statistical Identifier was allocated to each respondent. The household heads received in their emails also the links to the secondary dwelling's questionnaire/form.

Self-enumeration – the respondents accessed the link of questionnaire received by e-mail and complete the form online and transmitted the information to the main application server. The submission and data validation on CAWI data flow is presented in figure below (source: report of Output 7a).

Figure 3 - PHC Questionnaire validation workflow in Survey Solutions CAWI



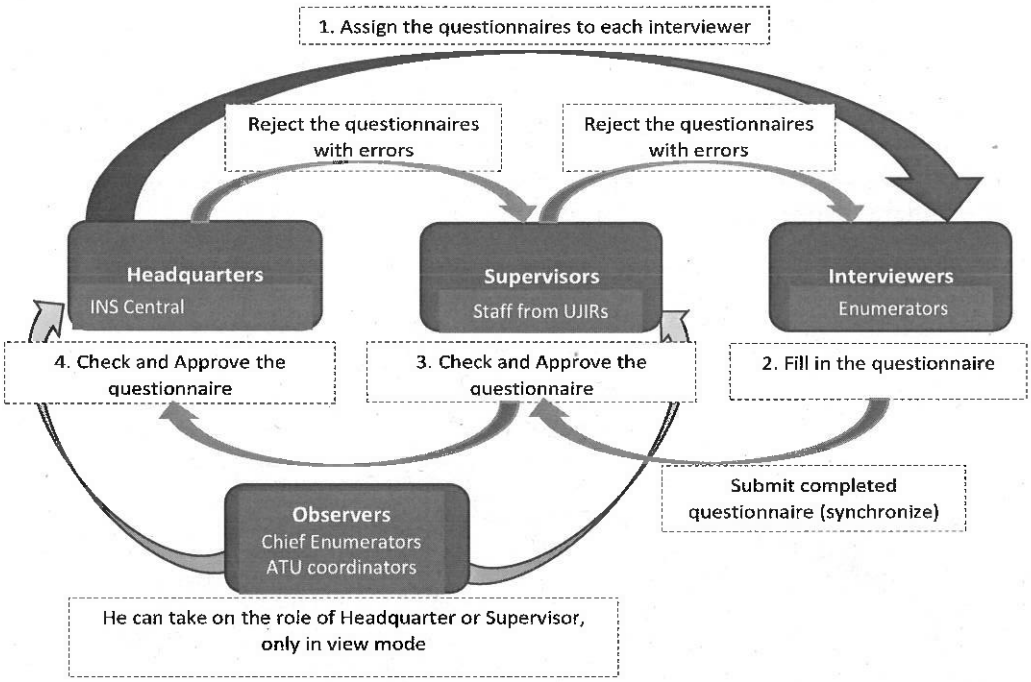
On the main server an automated validation procedure was executed, checking the provided information for acceptability. If in line with requirements, the self-enumeration phase of the process

was completed/over and the online questionnaire was sent to INS for final data storage. If not, on the CAPI phase, the enumerator had to contact respondents and follow up by face-to-face CAPI data collection phase. For validated questionnaires the respondents have could obtain an online confirmation of the success of the self-enumeration process by accessing a dedicated web platform.

2.3. Data flow on CAPI

The INS conducted face-to-face CAPI interviews with people who have not self-enumerated or have provided incomplete information during self-enumeration, using enumerators who get responses on their hand-held devices. The submission and data validation on CAPI data flow is presented in the figure below (source: report of Output 7a).

Figure 4 – PHC Questionnaire validation workflow in Survey Solutions CAPI



Based on automated validation, the CAPI-completed questionnaires were approved as Headquarters and transmitted to the INS central database. The supervisors (from DTS) verified the incomplete questionnaires or the ones with errors and reassigned them to enumerators to collect the missing or incorrect data. After validation of the questionnaires, they were transmitted to the INS central database.

The INS again checks this information for acceptability, as headquarters. If acceptable, the enumeration phase of the process is completed/over. If not, the supervisors of the enumerators were contacted for changes, and the changes were done by the enumerators at request of the supervisors.

The information, once acceptable, was then processed and stored.

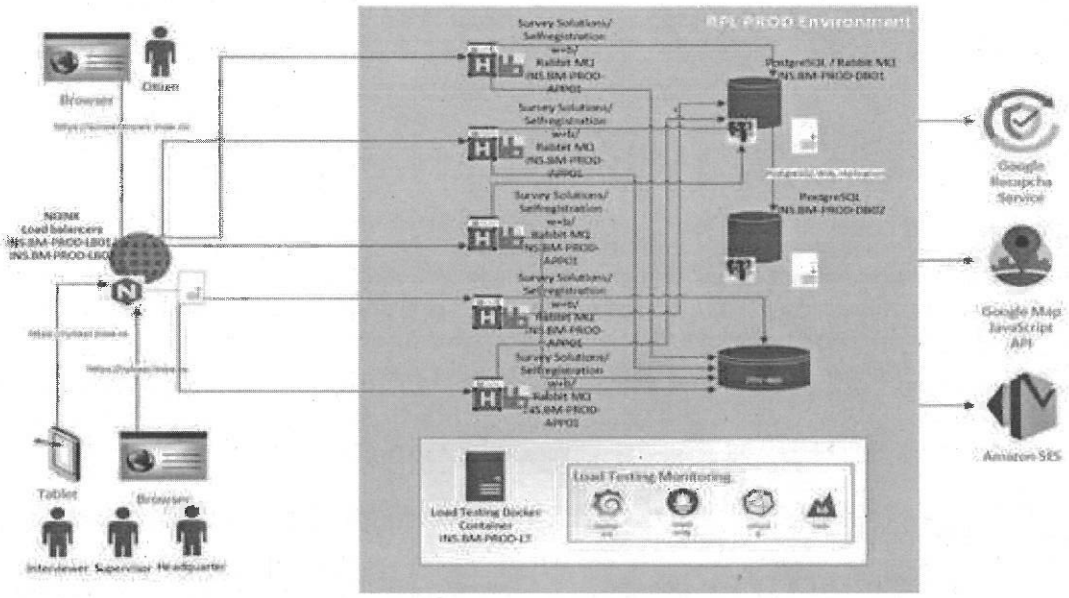
3. Actual PHC infrastructure for data collection

The reports related to PHC development and implementation provided detailed information on infrastructure (hardware, applications and communications) put in place for data collection process (see the reports of Outputs 4.1b, 4.1c, 4.2a, 7a, 10a, 10b, 10c under the same RAS[1]). Further are presented the steps and content of the actual PHC as built report.

3.1. Overall architecture

Based on the observations obtained from pilot and estimations of workload during the performance testing (see *Annex 1: PHC Test planning, execution and results*, Output 4.2a) in order to ensure that data collection process planned capacity and agreed responses times are met, and corroborated with the security testing for Internet exposed components of the system (subject of STS experts' activities covered up on existing STS – INS protocol and security reports), the actual PHC data collection system was built on a three-tier architecture, consisting in a pair of 2 load balancers serving the web requests, 5 application server nodes and a pair of 2 database nodes. Significant metrics are captured from all nodes and sent to a monitoring node.

Figure 5 - PHC Production environment



[1] Output 4.1b: *Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA)*
Output 4.1c: *Report on advisory services provided to Recipient on the Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out);*
Output 4.2a: *Recommendations and guidance provided to the Recipient in running the integrated IT system with its components, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA);*
Output 7a: *Report on advisory services provided to Recipient on Methodology to assess and promote continuously and in real time the quality and coverage of the collected data PHC2021 and data protection/security;*
Output 10a: *Report on Recommendations to the Recipient on how to perform the PHC2021 piloting process;*
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 ;* and
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*

The nodes are virtual machines hosted in STS premises. Following resources were dedicated to each node:

Table 1 - PHC IT resources assigned

Name	Provisioned Space	Guest OS	Memory Size	CPUs	IP Address
INS.BM-PROD-APP01	1000 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP02	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP03	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP04	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-APP05	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-PROD-DB01	1050 GB	Ubuntu Linux 20.04 (64-bit)	256 GB	48	
INS.BM-PROD-DB02	1050 GB	Ubuntu Linux 20.04 (64-bit)	256 GB	48	
INS.BM-PROD-LB01	58.08 GB	Ubuntu Linux 20.04 (64-bit)	32 GB	16	
INS.BM-PROD-LB02	58.08 GB	Ubuntu Linux 20.04 (64-bit)	32 GB	16	
INS.VM-PROD-SL	98.09 GB	Ubuntu Linux 20.04 (64-bit)	48 GB	16	

An additional 2T sized storage mounted via NFS on first database server node was intended to be used for online backup, by case.

Load balancers

Load balancers are based on NGINX community edition version 1.18. NGINX software is compiled from source in order to enable future customizations. Additional sticky modules were added to NGINX in order to enable session persistence using session cookies.

The role of the two load balancers is to balance web requests coming from Internet (routed via STS network equipment) to the 5 application nodes in a consistent manner, providing session persistence. They are configured in an active-passive high availability configuration, so when one node is down, the other can take over the requests.

Database servers

Database servers are based on Postgresql 12 community edition database management software. Postgresql 12 is installed from specialized repository dedicated to current operating system version.

Database nodes serves as information storage for keeping assignments, interviews, and system and user management information. The database nodes uses active-passive replication, so one node (MASTER) is on read-write mode while the second node (BACKUP) is on read-only. Automatic failover is implemented so in case MASTER is down, the BACKUP node is promoted as master. Also, database was configured for online backup, which make possible point in time recovery in case of information inconsistency.

Application servers

Application servers consists in instances of Survey Solution version 21.1.7.30747 which is the latest stable version at the moment of installation. Each node communicates with load balancers on port 9700 via http. A dedicated component called ExportService is enabled only on the first node. This

component carries out data export and communicates with all other nodes on port 5001. The service can be called only on server side, and it is not exposed to Internet. Data from Survey Solutions can be either exported via specialized archives format, using web interface or by connecting to another Survey Solutions instance via API.

Monitoring server

Monitoring server relies on a docker set of containers, with Prometheus as central monitoring collector, using also Grafana for metrics visualisations and InfluxDB for metrics recording. Additional functionality is added by a customized k6 deployment which can be used for performance testing.

3.2. High availability

High availability was considered in order to avoid single point of failure in PHC system design. Also, the recovery point objective of one hour (maximum time in which interviewers can stay disconnected from the system) was considered. High availability is implemented in each of the 3 tiers as follows:

- On load balancers, there is an active-passive configuration which bring up a second node in case current node NGINX service is down
- On application servers, since the requests are routed from load balancers, a health check can indicate to load balancers if the current application server node is down, and the request can be re-routed to another node
- On database servers, there is an active MASTER node which server all the requests. Data is replicated to a second node (BACKUP) and there is automatic failover. The failback mechanism is manual for better control of the failure source.

On load balancers, the high availability mechanism was implemented using keepalive software package. Keepalive was configured to bring up a virtual IP on the MASTER node while the NGINX service on this node is healthy, and to bring down the virtual IP on MASTER and bring it up on the BACKUP node in case NGINX is failing on the MASTER node.

Figure 6 – PHC load balancers high availability mechanism

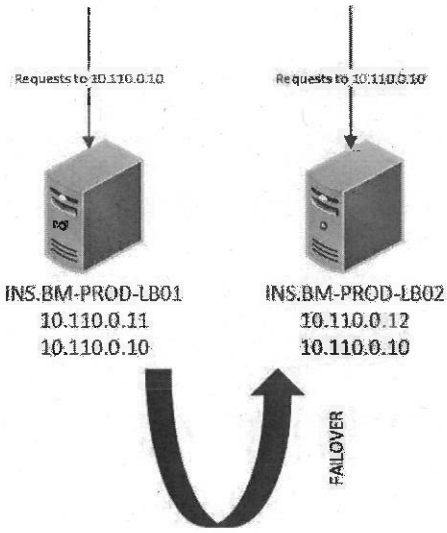
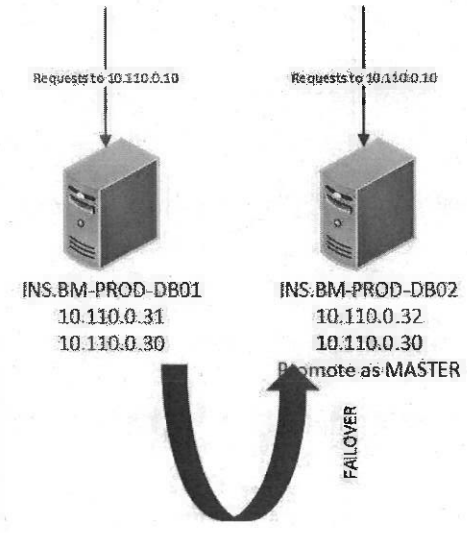


Figure 7 - PHC Database high availability mechanism



Similar high availability mechanism was implemented for databases. In database nodes case, additional action is needed in order to promote BACKUP node as MASTER in order to be able to server also write requests to SQL engine.

In order to check whether first node is acting as a master so keepalive to be informed about state changes, a custom mechanism was implemented. On the master node, a check script updates information in a table by using current date and time on each 5 seconds. If the information from the table (which is replicated to second node) differs from current date and time with more than 100 seconds, then it means that, either the Postgresql server is down, or the server is in read-only mode, so in this case automatic failover can be triggered.

In case of database servers, the failback mechanism is manual and is described in the chapter regarding operations. Since the failure causes can be multiple, from physical failures to software configurations issues, it is recommended that failback to be operated manually in order to recover initial state.

Database servers were configured for replication. The replication mechanism chosen was logical replication (see https://www.postgresql.org/docs/12/different-replication-solutions.html). This type of replication has the advantage of being a build in mechanism, providing a high level of consistency and having a minimal footprint on master server operations.

3.3. Performance

The system was sized in order to enable 25000 interviewers to work concurrently. Since the recommended configuration for a single node was proven to carry out at least 6000 interviewers, then 5 application nodes should be more than enough to handle PHC.

3.4. Backup

Due to high availability architecture, backup of server configurations is not needed since each node configuration is backed up in another node. So the backup of the system reduces to backup of the data.

Postgresql was configured for online backup, so point in time consistent recovery is possible in order to recover from logical failures. Archive logging is configured on each MASTER node in both datacenters.

3.5. 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 or WB.

The system logical security has different levels of protection:

- STS Network firewalls are protecting access to the systems hosted in STS datacenters. Access to public web interface of the system is restricted to national IPs.
- 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
- Administrative access to the system is protected by restricted VPN point-to-site solution, available only to authorized personnel
- Access to web application is possible only to authenticated users. Username and password are mandatory for accessing system functionalities, and privileged users are additionally

protected by 2 factor authentication. Also, a captcha mechanism (Google Recaptcha service) is used in order to deny programmatic access.

The overall security of the system was assessed by STS according to best practices. In order to continuously protect against common vulnerabilities, security patches for system components are applied on regular basis.

3.6. Log management

All the systems logs are centralized by STS. This allows setting alerts in order to notify responsible persons for particular events, related to system security, availability and reliability.

3.7. Disaster recovery

Disaster recovery mechanism relies on DNS switching and database replication. First node on the secondary site is continuously replicating from first node of the primary site using streaming replication, so the database on the secondary site is up-to-date in a consistent manner.

When a disaster situation is declared by responsible persons, the disaster recovery procedure is triggered and will have those main phases:

1. Switch network connectivity from primary to secondary site. This will use DNS record updating, so the changes will be applied according to agreed TTL. In this case, this will be one hour. The second site load balancers will publish initially a maintenance notification HTML page.
2. Meanwhile, a manual procedure will promote the first database node on the second datacenter to become MASTER. Replication with secondary node in second datacenter will be set up, and when everything is up, the other services will be started.
3. The maintenance notification page will be replaced with a redirect to SuSo services, so the users can access new environment.

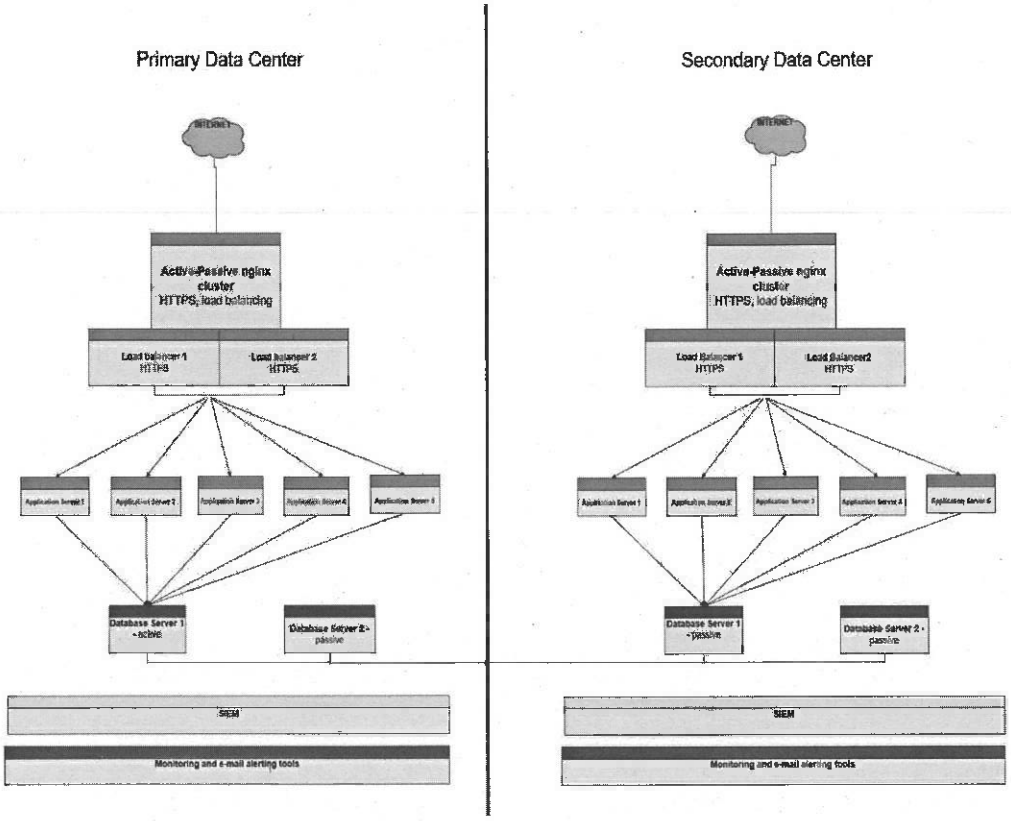
It is very important to maintain configuration on secondary datacenter to be the same as in primary datacenter, so every time when configuration changes in primary datacenter, it should be replicated on secondary datacenter. Following resources were used in secondary datacenter in order to cover the capacity of the primary datacenter:

Table 2 - PHC Secondary datacenter resources

Name	Provisioned Space	Guest OS	Memory Size	CPUs	IP Address
INS.BM-DR-APP01	1000 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-DR-APP02	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-DR-APP03	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-DR-APP04	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-DR-APP05	82.09 GB	Microsoft Windows Server 2019 (64-bit)	96 GB	24	
INS.BM-DR-DB01	1050 GB	Ubuntu Linux 20.04 (64-bit)	256 GB	48	
INS.BM-DR-DB02	1050 GB	Ubuntu Linux 20.04 (64-bit)	256 GB	48	
INS.BM-DR-LB01	58.08 GB	Ubuntu Linux 20.04 (64-bit)	32 GB	16	
INS.BM-DR-LB02	58.08 GB	Ubuntu Linux 20.04 (64-bit)	32 GB	16	
INS.VM-DR-SL	98.09 GB	Ubuntu Linux 20.04 (64-bit)	48 GB	16	

Also, a dedicated NFS storage is presenting 2T space to the first node of INS.BM-DR-DB01.

Figure 8 - STS Survey Solutions initial configuration



Virtual Machines (Testing Target)

- Load Balancer
  - 2x - 15vcpu, 32GB RAM
  - Linux Server (Ubuntu 20.04 LTS)
  - NGINX Load Balancer
- Application Servers
  - 5x - 24vcpu, 96GB RAM
  - Configured with Windows Server 2019
  - Survey Solutions HQ
- Database Server
  - 2x - 46vcpu, 256GB RAM
  - Windows Server or Linux (Ubuntu 20.04 LTS)
  - PostgreSQL Server (version 12.4)

Load Testing Machine (Source of Load)

- 3x 16vcpu, 48GB RAM
- Linux OS (Ubuntu 20.04 LTS)
- Docker

Network(s):
DMZ – 2 x Load Balancers
Private – 5 x Applications Servers, 1 x Database Server, 1 x Load Testing

Ports between networks:
Internet - Load Balancers 80, 443 (port 80 will be re-directed to 443 after SSL cert installation)
DMZ – Internet 80, 443

Load Testing – Load Balancers – 443, 9100, 8080
Load Balancers – Application Servers – 9700

Private – Internet 80, 443

The detailed configuration of the system is presented in **Annex 1 - PHC data collection information system – As-built report.**

Note: This is STS configuration - the information in Annex 1 to be used for the configuration of the INS's future structural surveys on its new integrated IT system.

4. Monitoring and reporting during the PHC data collection

4.1. Daily monitoring and reporting

The monitoring and reporting process for data collection represented a daily task realized for the entire period of the PHC (March 15 – July 31, 2022) as a joint effort of INS and WB teams. The solutions provided consisted scripts to generate reports for monitoring of data collection during CAWI phase (March 15 - May 27) and CAPI phase (May 31-July 31); and by counties, ATU, ARA and enumerators.

The monitoring and reporting process used the data gathered from Survey Solutions, namely the data export of the activities recorded during CAWI phase related for generated, filled and sent/ validated questionnaires and the activities performed by ARA (assistants for self-enumeration) and during the CAPI phase related to the activities performed by the enumerators for filling the questionnaires the day before, about the opened questionnaires and the completed ones by each census staff with roles in Survey Solutions and by ATU and counties.

The CAWI reports generated by the scripts available on ***'Data Processing server'*** path **/home/calex/RPL_CAWI_stat/daily_stats_v3.*R*** were structured as:

- A daily report for monitoring data collection by county comprising total number of questionnaires completed in urban and rural area with or without ARA support;
- A daily report for monitoring data collection by locality (Administrative Territorial Unit) comprising total number of questionnaires completed;
- A daily report for monitoring data collection by ARA comprising total number of questionnaires validated or unvalidated completed, the county and ATU where the place of assistant's activity was registered; and
- A daily report for monitoring the type of errors and messages sent to respondents during self-registration process;

The daily monitoring process provided the data for PHC daily data collection communication releases during CAWI provided by INS team in aggregations by weeks about residence (urban/rural) and self-enumeration method (with or without ARA).

Figure 9 - Sample of daily report of CAWI data collected (by counties)

g_jud	g_jud_code	chestionare	urban	rural	ara	non_ara
ALBA	10	177877	109211	68666	107800	70077
ARAD	29	249046	147412	101634	153441	95605
BACAU	47	257419	128585	128834	156150	101269
BIHOR	56	373203	189769	183434	205204	167999

Figure 10 - Sample of daily report of CAWI data collected (by ATU)

g_jud_code	g_muni	g_muni_code	g_sat	g_sat_code	chestionare
10	BLANDIANA	3397	ACMARIU	3413	222
10	BLANDIANA	3397	BLANDIANA	3404	340
10	BLANDIANA	3397	POIENI	3431	2
10	BLANDIANA	3397	RACATAU	3440	22
10	BUCIUM	3459	BUCIUM	3468	75
10	BUCIUM	3459	BUCIUM-SAT	3495	96

Figure 11 -Sample of daily report of CAWI data collected (by ARA)

jud	uat	codr	nr_chest_valide	nr_chest_invalide	cod_ara_corect
AB	MUNICIPIUL_ALBA_IULIA	001017 0001	395	8	1
AB	MUNICIPIUL_ALBA_IULIA	001017 0002	549	4	1
AB	MUNICIPIUL_ALBA_IULIA	001017 0003	1021	4	1

Figure 12 - Sample of daily report with type of errors/messages recorded during CAWI

interview__key	interview__id	roster	id1	variable	type	message__number	message
41-74-13-67	0872abc356014e8	NA	NA	AA_SLR	2	2	AtenÅ£ie! Anul Ã®nregistrat nu poate fi mai mare decÃ¢t 2021 !
38-43-83-92	bc3baa6a643b456	NA	NA	AA_SLR	2	1	AtenÅ£ie! Anul Ã®nregistrat nu poate fi mai mic decÃ¢t anul de naÅ¥te
83-59-52-06	ac723fe6f5044ccf9	NA	NA	AA_SLR	2	1	AtenÅ£ie! Anul Ã®nregistrat nu poate fi mai mic decÃ¢t anul de naÅ¥te
04-64-09-99	0fee9b1a588f4a00	NA	NA	AA_SLR	2	1	AtenÅ£ie! Anul Ã®nregistrat nu poate fi mai mic decÃ¢t anul de naÅ¥te

The CAPI reports generated by the scripts available on '***Data Processing server***' path **/home/calex/RPL_CAPI_stat/monitorizare_date_colectate_Judete.Rmd** were structured as:

- A comprehensive daily report for monitoring data collection by country and by county comprising total number of questionnaires completed in each county and by each enumerator/supervisor and the stages of questionnaires (completed, (re)allocated, rejected/approved by Supervisor; rejected/approved by Headquarter;

Figure 13 -Sample of daily report of CAPI data collected (country, county, enumerator)

County	ID Supervizor (SUP)	ID enumerator (INT)	Total no. of questionnaires assigned to enumerator	Total no. of questionnaires on server	Total no. of enumerated persons	out of which, by stages					
						Completed	Questionnaires re(alocated)	Rejected by supervizor	Aproved by supervizor	Rejected by Headquarter	Aproved by Headquarter
AB	SUP_AB_01	INT_10170001	283	450	336	0	0	0	316	0	134
AB	SUP_AB_01	INT_10170002	343	506	346	0	0	0	237	0	269
VS	SUP_VS_N1	P_167320_0005	2	0	0	0	0	0	0	0	0
		Total COUNTRY	**6026618**	**7809721**	**8812645**	**73191**	**82**	**7076**	**3258433**	**68**	**4470871**

Figure 14 - Sample of daily report of CAPI data collected (county, enumerator)

County	ID Supervizor (SUP)	ID enumerator (INT)	Total no. of questionnaires assigned to enumerator	Total no. of questionnaires on server	Total no. of enumerated persons	out of which, by stages					
						Completed	Questionnaires re(alocated)	Rejected by supervizor	Aproved by supervizor	Rejected by Headquarter	Aproved by Headquarter
AB	SUP_AB_01	INT_10170001	283	450	336	0	0	0	316	0	134
AB	SUP_AB_01	INT_10170002	343	506	346	0	0	0	237	0	269
AB	SUP_AB_01	INT_10170003	324	383	515	0	0	0	187	0	196
AB	SUP_AB_01	INT_10170004	322	498	238	0	0	0	210	0	288

In the same process of monitoring the PHC activities and specific results by enumerator or time for completion of questionnaire, the analysis of the paradata can offer valuable information. For this is recommended to use a special purpose application available for INS staff (no credentials are required) on ***Data Processing Server*** path **phc__censustools.**

Following the monitoring process and the distribution of tasks, the information generated was consolidated into daily reports that allowed the PHC coordination team from INS to inform and advise DTSs and enumerators, supervisors, local coordinators and observers about the status of data collection and the level of coverage.

The templates of the reports are available in electronic format and are delivered archived with this report as **Annex 2 - Sample reports of daily monitoring (CAWI, CAPI).**

4.2. Census staff's activity monitoring and reporting for payment

The reporting process on the activity of the census staff had two dimensions: one regarding the reporting on activity and the other one used for the calculation of the gross amounts paid to census staff.

The monitoring of data collection (using the daily exports from Survey Solutions) was used for the coordination of activities at the level of enumerators, supervisors, coordinators and for administrative processes regarding the payment of the work done by them. The information regarding the number of the questionnaires filled, the completion status by sections of dwelling, person, building of the questionnaires provided the base information for the monitoring and coordination of the enumerators and later for the payment for their effort (see the sample of a reports by enumerator and by county/ATU in the figures below).

Figure 15 - Sample of payment report for enumerator's activity

Payrol statement - enumerator period May 31– July 31, 2022)								
Situaţia de plată - pentru recenzori (perioada 31 mai – 31 iulie 2022)								
ATU name		ATU SIRUTA						
Denumire UAT SLATINA		**SIRUTA UAT**	**125347**					
ATU cathegory (locality type 1 - usual; type 2 - isolated or with territorial dispersion)						Enumerator ID		
Tip localitate (Localitate1- obişnuită; 2-izolată sau dispersată teritorial)				1		**Cod recenzor (user recenzor)**		INT_123456789
(Date) **Data**	(Password) **Parola**	(From valid questionnaires eligible for payment:) **Din chestionarele validate eligibile pentru plată:** (TOTAL NO. OF SECTIONS FOR DWELING (LC) ENUMERATION COMPLETED) **TOTAL SECŢIUNI PENTRU RECENZAREA LOCUINŢEI (LC) COMPLETATE**	(Rate LC) **Tarif LC** **(LEI)**	(From valid questionnaires eligible for payment:) **Din chestionarele validate eligibile pentru plată:** (TOTAL NO. OF SECTIONS FOR PERSON (P) ENUMERATION COMPLETED) **TOTAL SECŢIUNI PENTRU RECENZAREA PERSOANEI (P) COMPLETATE**	Rate (P) **Tarif P** **(LEI)**	(From valid questionnaires eligible for payment:) **Din chestionarele validate eligibile pentru plată:** (TOTAL NO. OF SECTIONS FOR BUILDING (SC) ENUMERATION COMPLETED) **TOTAL SECŢIUNI PENTRU RECENZAREA SPAŢIULUI COLECTIV DE LOCUIT (SC) COMPLETATE**	Rate (SC) **Tarif SC** **(LEI)**	Total ammount for payment **Sumă totală de plată** **(LEI)**
A	B	1	2 = 1 x 3.5Lei	3	4 = 3 x 7Lei	5	6 = 5 * 3.5Lei	7

Figure 16 - Sample of payment report for enumerator's activity by county and ATU

Payrol statement - enumerators period May 31– July 31, 2022)									
Situaţia de plată - pentru recenzori (perioada 31 mai – 31 iulie 2022)									
County									
Judet	**JUDET**								
(ATU name) **Denumire UAT**	(From valid questionnaires eligible for payment:) **Din chestionarele validate eligibile pentru plată:** (TOTAL NO. OF SECTIONS FOR DWELING (LC) ENUMERATION COMPLETED) **TOTAL SECŢIUNI PENTRU RECENZAREA LOCUINŢEI (LC) COMPLETATE**	(Rate LC) **Tarif LC** **(LEI)**	(From valid questionnaires eligible for payment:) **Din chestionarele validate eligibile pentru plată:** (TOTAL NO. OF SECTIONS FOR PERSON (P) ENUMERATION COMPLETED) **TOTAL SECŢIUNI PENTRU RECENZAREA PERSOANEI (P) COMPLETATE**	Rate (P) **Tarif P** **(LEI)**	(From valid questionnaires eligible for payment:) **Din chestionarele validate eligibile pentru plată:** (TOTAL NO. OF SECTIONS FOR BUILDING (SC) ENUMERATION COMPLETED) **TOTAL SECŢIUNI PENTRU RECENZAREA SPAŢIULUI COLECTIV DE LOCUIT (SC) COMPLETATE**	Rate (SC) **Tarif SC** **(LEI)**	Total ammount for payment **Sumă totală de plată** **(LEI)**	ATU cathegory **Tip localitate**	ATU **Siruta UAT**
0	1	2 = 1 x Tarif_LC Lei	3	4 = 3 x Tarif_P Lei	5	6 = 5 * Tarif_SC Lei	7	8	9

The calculation of total amount paid to an enumerator and the total amounts distributed to ATUs and counties through State budget allocations. The scripts are available on '***Data Processing server***' paths**/home/calex/RPL_CAPI_stat/situatii_plata_CAPI_v5.R** ; **/home/calex/RPL_CAPI_stat/situatii_plata_CAPI_zile_v2.R**

The samples of the reports are presented in **Annex 3 - Sample payment reports CAPI** and are provided archived in an electronic format.

5. Validation processes of the collected data

The purpose of the validation process was to support the preparedness of census data for releasing the preliminary results and, also, to send them to Eurostat, and the dissemination of the results to other users (national / international): publications (predefined tables) and tables produced following ad-hoc requests. This section presents information on the process and tools for the validation of the data collected.

Processing the census data consisted of data validation, error correction, deduplication, imputation of missing or incomplete data, calculation of statistical indicators included in production of census results. Also, all these activities envisaged to processes personal data in such a way as to ensure their adequate security. The census data collection process involves several steps to ensure the integrity and accuracy of the data: Data Collection, Data Processing, Data Editing, Data Imputation, Data Disclosure Review (SDC), Data Quality Control, Data Release (Dissemination).

Using the CAWI and CAPI methods for data collection for census provided advantages in the correctness of the data, mainly because of the validation rules included in the questionnaire design and the real-time validation during the data collection phase. Also, data enumerators were trained to follow strict guidelines to ensure that the information collected is accurate and complete.

5.1. Data processing during CAWI, CAPI data collection

During and after data collection, the INS processed the data to ensure that it is accurate and complete. This includes checking for errors and inconsistencies and making sure that all data is in the correct format.

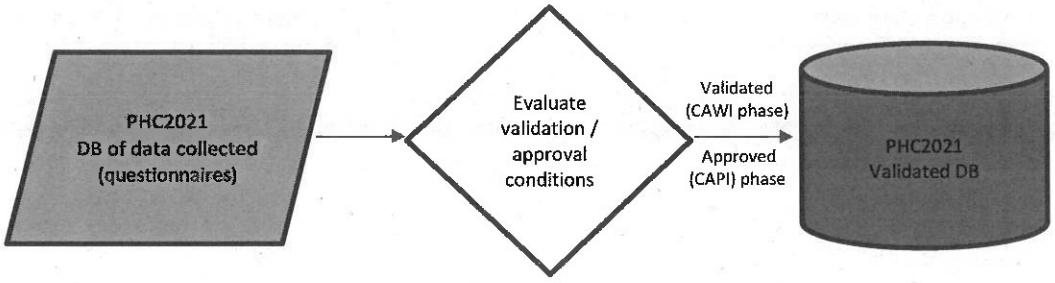
The reduction of the risk of input errors (typos) during filling out the questionnaires was achieved through advanced navigation and validation rules within the design phase of the questionnaire. The capabilities of the data collection platform offered the possibility of keeping the validation conditions from the questionnaire, both for the CAWI method and for the CAPI.

The difference between the design of the CAWI and CAPI questionnaire consisted in the display of information clarifying the questions and answers, being adapted to people who self-enumerate in the case of CAWI and enumerators in the case of CAPI.

In the CAWI and CAPI questionnaire the validation rules were implemented for typos, for intercorrelation between answers and logical conditions related to the workflow in the questionnaire. A method of reducing typos in CAWI was to use lists instead of free text boxes as much as possible.

The data processing flows were similar, but CAWI process ended with the validation of questionnaire since the CAPI process had validation and ended with the approval of questionnaire (see figure below).

Figure 17 – Automatic flows for validation (CAWI phase) / approval (CAPI phase)



In the CAWI phase of the census the automatic validation rules were implemented for completeness of the collected data. Daily, based on the available collected data, the automatic algorithm for validation was applied on all finished/completed questionnaires (available on '*Data Processing server*' path **/home/calex/RPL_CAWI_stat/validari_v3.R** and provided in electronic format with this report, **Annex 4: PHC data collection validation and approval scripts (CAWI, CAPI).**

After this validation, based on questionnaires for persons a list of the persons which finalized the self-enumeration with success was generated and sent to the web platform which issued the prove for self-enumeration. The automatic algorithm was elaborated by INS census team and implemented as R scripts. A method of identifying multiple questionnaires filled out by one person was implemented, and the most recently completed correctly and completely questionnaire was kept in the database.

In CAPI phase the automatic validation rules were implemented for identification of errors and for completeness of the collected data. The validation rules are stricter than the ones from CAWI phase because the enumerators are trained to complete the questionnaire and they must be aware of the correctness of the data collected. An automatic process for approvals of the correct questionnaires is implemented as R script in analytic server and have run daily on all data collected (available on '*Data Processing server*' path **/home/calex/RPL_CAPI_stat/validari_CAPI_v2.R** and **/home/calex/RPL_CAPI_stat/aprobari_automate_RPL_v1.10.R** provided in electronic format with this report, **Annex 4 - PHC data collection validations and approvals scripts).**

Based on the results of the application of the algorithms for validation and approval of the questionnaires, the status reports for the questionnaires, for the enumerators and for the supervisors, were made daily.

5.2. Data processing post data collection

The phases described below are components of the after-census data processing.

Data editing - the INS census team used a variety of editing procedures to identify and correct errors in the data, that included checking for inconsistencies, missing values, and out-of-range values.

Data imputation - for missing data, the INS census team used administrative sources for data completion and statistical imputation methods to estimate values. This approach helped to ensure that the final data is as complete and accurate as possible.

Data disclosure review - the census team reviewed the data to ensure that it does not disclose any information about individual respondents. This is done to protect the privacy of respondents on dissemination data. Statistical Disclosure Controls methods are applied on raw microdata to prepare the final microdata for dissemination. The proposed method for SDC, the record swapping, derived also from the technical assistance provided and trainings in this field attended by INS experts.

Quality control of data - the census team conducted extensive quality control checks on the data to ensure its accuracy and completeness. They also conducted independent evaluations to assess the quality of the data.

Data release (dissemination) - as a last phase, the INS census team releases the data to the public. They also make sure to provide the necessary documentation and metadata to help users understand and correctly use the data.

6. Data editing and production for dissemination

The PHC data dissemination comprises three specific solutions: data hypercubes, mandatory to be published on Eurostat portal, grid distribution of PHC data allows the dissemination of 21 indicators (13+8) on grids (INSPIRE) of 1sqkm for entire Romania's territory and the dynamic tables dissemination with PHC results. While the INS team will take care of data hypercubes (preparing the IT environment and install the necessary applications provided by Eurostat to retrieved data from the repository of PHC data organized on INS IT environment), this section presents the grid distribution and dynamic tables developed under the project.

6.1. Grid distribution of PHC data

Population grids are a powerful tool for understanding and analyzing the characteristics and patterns of human populations. They allow researchers to study the interactions between human activities and the environment, and to understand the underlying causes of phenomena that span administrative boundaries[2]. For example, population grids can be used to study the effects of flooding, commuting, and urban sprawl on communities. By using population grids to analyze census data, researchers can gain a more detailed and nuanced understanding of the demographics, behaviors, and needs of different populations, and can use this information to inform policies and decision-making. Additionally, population grids can be a useful tool for assessing the distribution and availability of resources, such as housing, healthcare, and education, and for identifying areas that may need greater support or intervention. In short, transforming census data into population grids can provide a wealth of insights into the dynamics and complexities of human society, and can help us to better understand and address the challenges facing our communities.

Eurostat disseminates population grids as part of its efforts to provide detailed and accurate data on the demographics, behaviors, and needs of the European population. These population grids are produced using data from the European Union's population and housing censuses and are designed to provide a detailed and consistent picture of the population at the level of individual cells, typically covering an area of 1 square km.

The population grids disseminated by Eurostat use a variety of variables to describe the characteristics of the population within each cell, including demographic characteristics such as age, gender, and nationality, as well as information on housing, education, employment, and other socio-economic factors. These variables are used to create detailed profiles of the population within each cell, allowing researchers to understand the characteristics and needs of different populations at a granular level.

The CRS used by Eurostat's population grids, is the ETRS89-LAEA projection (EPSG:3035), which is a variant of the European Terrestrial Reference System (ETRS89) that is optimized for use in Europe. This projection is widely used in Europe for mapping and spatial analysis and is particularly well-suited for applications that involve large-scale mapping and the visualization of data across national or regional boundaries. See info-box for details.

The toolbox created for INS to support the transformation of the Romanian census data into population grid data consists of two components:

i. An R script transforming the data into grid like statistical units and storing it in a PostGIS-PostgreSQL database.
ii. An R based Shiny application for the visualization of the different population grids.

[2] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_grids

The ETRS89-LAEA projection is a method used to map locations on the surface of the Earth onto a flat plane. It is commonly used in Europe, and it is particularly well-suited for mapping large areas at small to medium scales.
In simple terms, a projection is a way of taking the curved surface of the Earth and flattening it out onto a 2-dimensional plane, like a map. The ETRS89-LAEA projection is a specific type of projection that is known for its ability to maintain the correct shapes of large areas, while also minimizing distortion of size. The LAEA projection is also known for its ability to maintain accurate direction and angle measurements.

R-script for the transformation

To make the data available in the form of population grids as required by Eurostat[3] as well as to minimize processing times for the visualization in the shiny application, the is pre-processed into on population grid containing 21 variables, of which 13 variables are the ones mandatory by Eurostat, and 8 which INS considered as relevant in the context of Romania.

The **13 variables** as required by Eurostat are:

Table 3 - The census topic categories Eurostat

	Census topic categories to be broken down on the 1 km² reference grid	STAT. G.
0.	SEX.0.: Total population	0.
1.	SEX.1.: Male	1.
2.	SEX.2.: Female	2.
3.	AGE.G.1.: Under 15 years	3.
4.	AGE.G.2.: 15 to 64 years	4.
5.	AGE.G.3.: 65 years and over	5.
6.	CAS.L.1.1.: Employed persons[1]	6.
7.	POB.L.1.: Place of birth in reporting country	7.
8.	POB.L.2.1.: Place of birth in other EU Member State	8.
9.	POB.L.2.2.: Place of birth elsewhere	9.
10.	ROY.1.: Place of usual residence one year prior to the census unchanged	10.
11.	ROY.2.1.: Place of usual residence one year prior to the census: move within reporting country	11.
12.	ROY.2.2.: Place of usual residence one year prior to the census: move from outside of the reporting country	12.

[3] Eurostat. EU legislation on the 2021 population and housing censuses Explanatory Notes. Luxembourg: Publications Office of the European Union; 2019, https://ec.europa.eu/eurostat/documents/3859598/9670557/KS-GQ-18-010-EN-N.pdf/c3df7fcb-f134-4398-94c8-4be0b7ec0494 ; Data on the category 'employed persons' shall be transmitted as far as possible, subject to availability in the transmitting Member State

The **8 additional variables** as required by INS are:

Table 4 – The census topic categories (8) additional INS

	Census topic categories to be broken down on the 1 km² reference grid	STAT. G.
13.	EDU.1: Population with low level of education	0.
14.	EDU.2: Population with medium level of education	1.
15.	EDU.3: Population with high level of education	2.
16.	EDU.1.1.: Illiteracy	3.
17.	INA.1: Economically inactive population	4.
18.	INA.1.1.: Economically inactive population - pensioners	5.
19.	INA.1.2.: Economically inactive population – pupils, students	6.
20.	INA.1.3.: Economically inactive population – other inactive	7.

Important for the seamless processing of these variables is the quality of the incoming data, which is outlined in the following:

- one single csv file,
- one single id for each unit,
- with all 21 variables,
- plus the latitude and longitude present,
- resulting in a file with exactly 23 columns,
- without any missing values in the variables of interest,
- without any missing coordinates,
- without any coordinates outside the country,
- with all variables numeric except the ID,
- and with SDC (spatial) applied.

All variables used in this process need to be transformed to represent count data, which means the variable contains 1 if the person belongs to a particular population group and 0 otherwise.

Formally this is written as an indicator function:

$$\mathbf{1}_{group}(x) = [x \in group]$$

Subsequently the cleaned and transformed data will be transformed into a spatial object, and its geographic coordinates in latitude-longitude notation (EPSG:4326) will be transformed into the above described ETRS89-LAEA projection (EPSG:3035) for the creation of the grid codes.

Shiny application

For the visualization of the data produced in the previous step a shiny application has been developed and made available to INS on their analytics server.

Figure 18 - R Shiny based application for the visualization of the population grids.

The application has the following functionalities:

- Selection of the county (province).
- Selection of the layer (variable) of interest.
- Visualization of the selected population grid.
- Mouse-over on a single grid cell displays a pop-up with the corresponding grid cell information.
- Download of the selected population grid as pdf file.

It can be accessed under the following URL: 10.2.106.10/phc_grid/[4]. For the creation and visualization of the interactive maps, the R mapdeck package[5] is used, see details in the corresponding info box.

Why the R mapdeck package?

Visualizing spatial polygons datasets can be a challenge, as the polygons may overlap or be too numerous to display effectively. This can lead to cluttered and unreadable visualizations. Additionally, large datasets may be too resource-intensive to load and render in a web browser or mapping application.

The R mapdeck package can help with the visualization of large datasets consisting of spatial polygons by providing a set of tools for creating interactive and web-based maps. It uses the Mapbox GL JavaScript library and Deck.gl to render maps, which allows for fast and efficient rendering of large datasets. The package also provides a number of features that can help to reduce clutter and improve the readability of the visualization, such as the ability to filter data, color-code polygons based on different attributes, and adjust the transparency of polygons. Additionally, the package allows users to customize the look and feel of the map through a wide range of styling options.

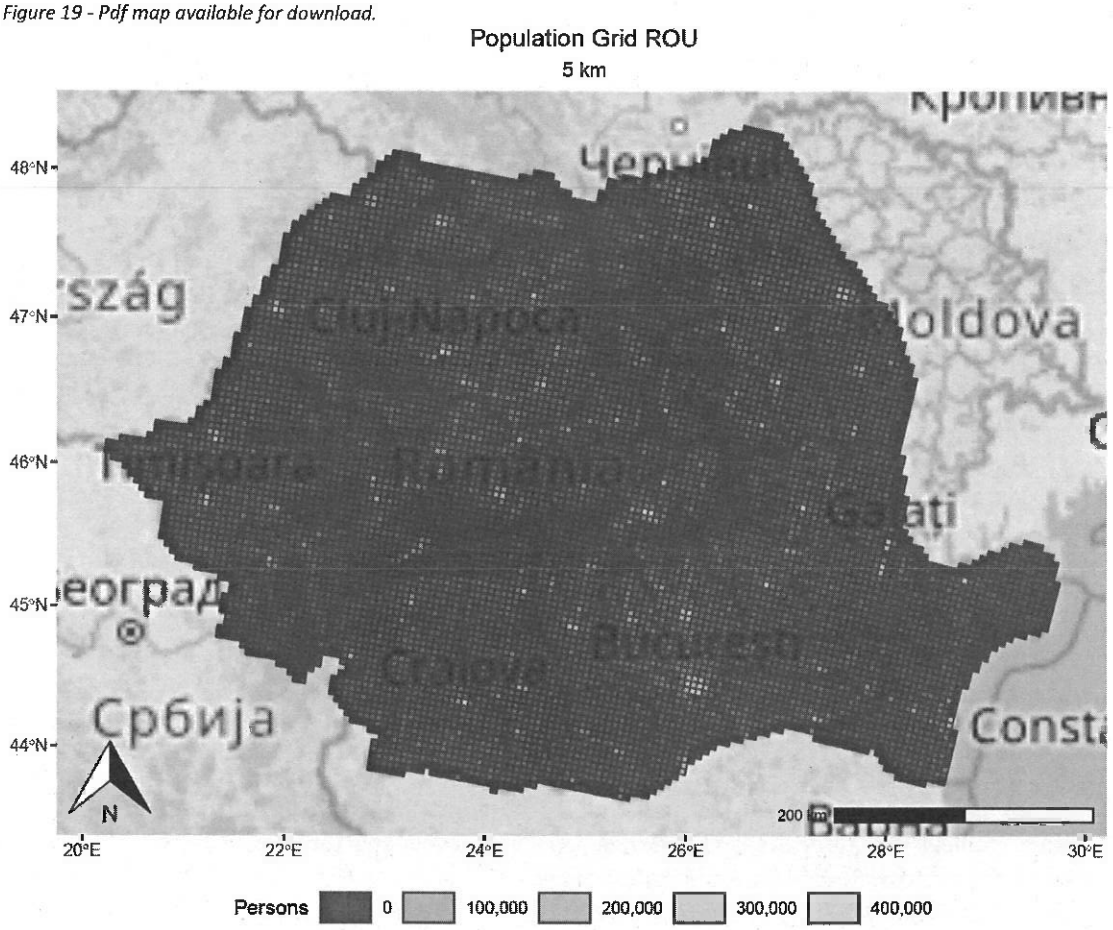
To use the R mapdeck package a free Mapbox subscription is required. A valid subscription can be obtained here: https://www.mapbox.com/. For details on Deck.gl see here: https://deck.gl/#/.

[4] This URL is based on INS's internal network configuration at the times of writing the report. Any future change in this configuration may also affect the host IP address.

[5] https://symbolixau.github.io/mapdeck/index.html

After selecting the county, and the corresponding population layer, the map will be displayed as shown in **Error! Reference source not found.**. After the map is loaded, it is possible to download a pdf file of the map displayed on screen. This is shown in **Error! Reference source not found.**2.

Figure 19 - Pdf map available for download.



The source code is distributed over the standard R shiny application files, consisting of:

- **ui.R** file wich contains the UI code,
- **server.R** which contains the server side code,
- **style.css**: the css styles for the UI
- **helpers**: a directory of helper functions
- **www**: for logos etc.

These files need to be copied into a folder (currently named ***phc-grid***) inside the ***/srv/shiny-server*** location to run the application. All files are provided as is, however since INS has access to the source code as well as a team of competent R programmers, they are free to modify it in whatever way they prefer.

6.2. Dissemination (tabular dissemination) of PHC data

Tabular dissemination is the key component that INS will use to make the census data widely available to the public. To a certain extent, the dissemination interface will replace the tabular dissemination used in the 2011 census, where all tables were generated beforehand. The approach proposed ensures a level of flexibility in terms of exporting the data and providing the appropriate tools for third-party data requests. In addition, the interface will provide a minimal mapping visualization of the data, in line with certain technical limitations (described below). The dissemination interface is built using the R programming language and the {shiny} framework and has been tested for internal use only.

General framework and logic

The dissemination interface uses the Eurostat interface as a starting point and structures the data based on topics, subtopics and disaggregation levels. The data navigation tree starts from 8 basic topics. The list of topics and subtopics was drafted by INS and considered national dissemination needs and data categories, Eurostat requirements, and technical requirements. The list includes the following topics:

1. DEMOGRAPHIC CHARACTERISTICS
2. ETHNO-CULTURAL CHARACTERISTICS
3. EDUCATIONAL CHARACTERISTICS
4. MOBILITY OF THE RESIDENT POPULATION
5. ECONOMIC CHARACTERISTICS
6. HOUSEHOLDS
7. RESIDENTIAL BUILDINGS
8. HOUSING AND LIVING CONDITIONS

Each subtopic is divided into subtopics, summing a total of 145 subtopics, including several territorial and indicator-level disaggregations. For example, the subtopic "The resident population by gender and age group" is included in the topic "Demographic characteristics". In addition, the subtopic includes disaggregations at different territorial levels and for each indicator, as well as their respective crosstabulations, as described below:

1. DEMOGRAPHIC CHARACTERISTICS
    a. The resident population by gender and age group
        i. Territorial levels available
            1. National
            2. Macroregions
            3. Development regions
            4. Counties
            5. Localities
        ii. Indicator disaggregation
            1. Gender (Total / Male / Female)
            2. Age groups (Total, 0-4, 5-9, 10-14, 15-19, 29-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+)

The resulting database is significantly complex and includes a large number of variables. As a result, the general framework is built on a hierarchical selection process that first selects the main data source, filters and picks the grouping variables, filters the indicators needed and then returns a pre-processed data frame used for dissemination purposes and visualization.

Technical requirements and software used

Similar to the PHC grid distribution grid application described above, the dissemination platform will use the same infrastructure and setup described in Output 4.1b, and no additional setup is necessary. The platform requires running installations of R and {shiny} (open-source version). RStudio Server is not necessarily required for deployment, but it is recommended to allow for fast modifications and updates.

Following the installation, the shiny application must be moved to the folder served on the web. On a normal installation, these application folders containing all files and data would reside in ***/srv/shiny-server/***. However, due to the large files in use by the INS team and the configuration of resources for this particular installation, the server folder has been moved to ***/arhiva1/srv/shiny-server/***[6]. Copying and pasting the folders to *the* ***/arhiva1/srv/shiny-server/phc_diseminare/*** will make the application available on ***Data Processing Server*** the local network at the address: ***http://xx.xxx.xx/phc_diseminare/***.

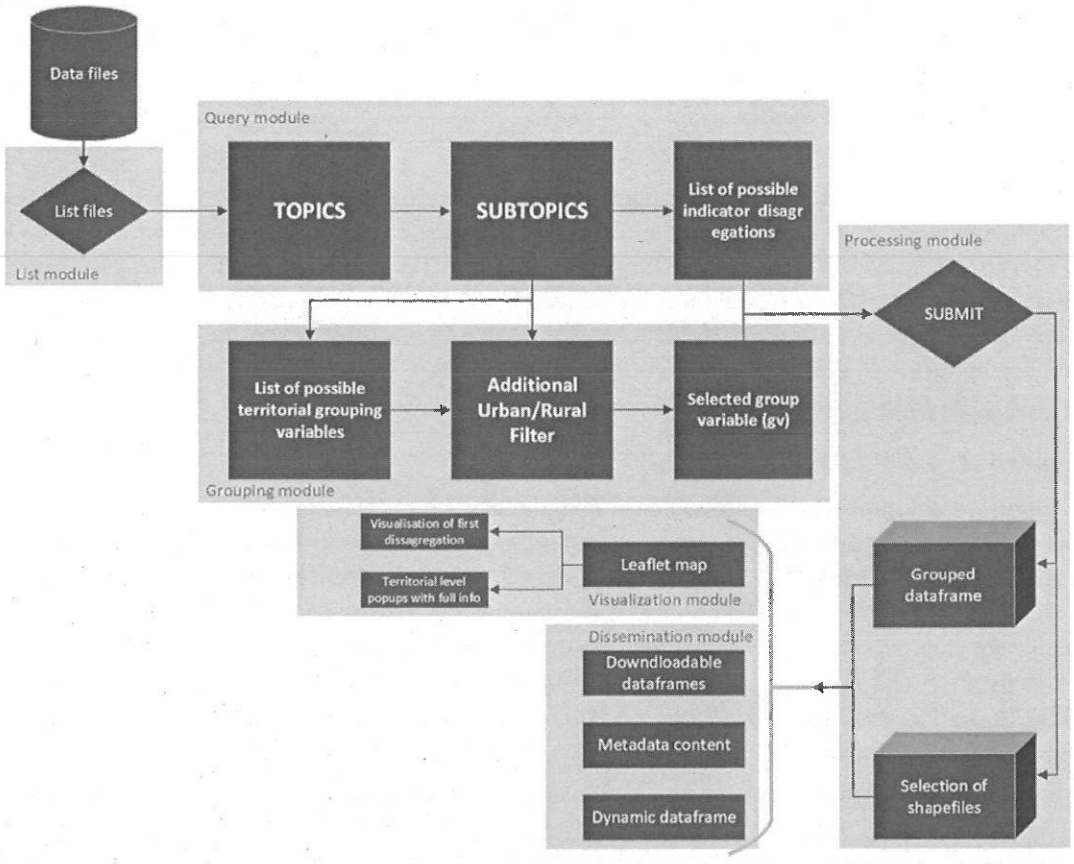
Solution design and process flow

The proposed solution incorporates several components (defined as modules) that interlink and creates a file query system that can return several indicators based on a set of standardized pointers. The dissemination platform's main data component is the pool of tables generated in advance by INS with all calculations included and SDC applied at the lowest administrative-territorial unit available for each subtopic. The data is fed into the platform, passing through each module to reach the visualization and dissemination stage. The main modules include the listing module, query module, grouping module, processing module, visualization and dissemination modules. The general process is described in the diagram below.

The dissemination platform lists the data files in the source folder on the first load. Next, it creates a list of names correlated with a preset set of topics and subtopics, returning the topics and subtopics available in the data. Then, based on the subtopic selection, the query module returns the available indicators and disaggregation levels and the available grouping variables. Finally, based on the combination of indicators, disaggregation and grouping variables, the user submits the query to the processing module. This module runs the query and returns two datasets: a grouped data frame based on the selections of the user and the required shapefiles for rendering the data.

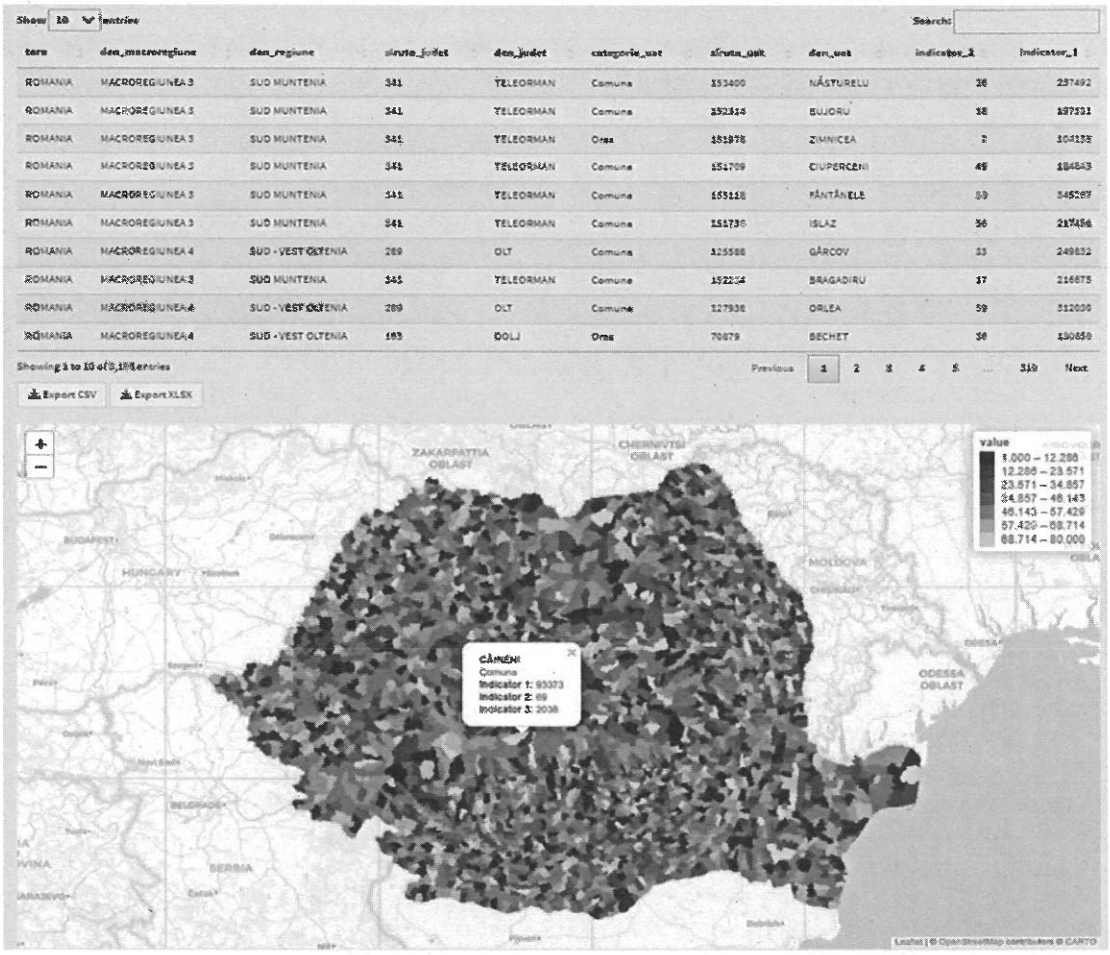
[6] We do recommend that INS uses the default installation as recommended by Rstudio (now Posit) and the R documentation.

Figure 20 - The dissemination process flow



Starting from the generated data, the visualization module renders the selected indicator on the map using the leaflet framework. Since multiple indicators cannot be rendered simultaneously, the map only renders the first indicator in the selection. The map uses OpenStreet Map as the background drop, as this is the most lightweight solution and requires no additional setup. In addition, for each ATU/shape, the map will render a popup summarizing the selection of indicators. The map allows the selection of each shape, zooming and panning. Simultaneously with the map, the platform returns a table consisting of the data generated by the user selection and download links for CSV and XLSX formats of the data, as well as an option to save the map in .png or .pdf format (see image example below).

Figure 21 - Data disseminated format saved



Data requirements and conditions

The data table files are the base for the dissemination platform. Due to the large variation and complexity of the topics and subtopics required, the files generated by INS (145 files, one for each subtopic) as the basis for the dissemination platform must be standardized and follow the same schema every time. In addition, INS needs to continuously provide and update the required shapefiles, including names and codes for all ATU levels, regions and macroregions.
The minimal list of columns to be included in one data table file is:

- country
- macroregion_code (NUTS)
- macroregion
- region_code (NUTS)
- region (NA if not applicable)
- county_code (SIRUTA, NA if not applicable)
- county_name (NA if not applicable)
- locality_code (SIRUTA_SUP, NA if not applicable)
- locality_name (NA if not applicable)
- locality_category (NA if not applicable)

Additionally, all relevant columns required for grouping/disaggregation will be included for each subtopic, and clear instructions about missing data and 0 levels should be provided.
As an example, for the example subtopic used above, "The resident population by gender and age group", the provided columns should include:

- country
- macroregion_code
- macroregion
- region_code
- region
- county_code
- county_name
- locality_code
- locality_category
- gender_total
- gender_male
- gender_female
- age_group_total
- age_group_0-4
- age_group_female_0-4
- age_group_male_0-4
- age_group_ 5-9
- age_group_female_ 5-9
- age_group_male_ 5-9
- age_group_ 10-14
- age_group_female_ 10-14
- age_group_male_ 10-14
- age_group_ 15-19
- age_group_female_ 15-19
- age_group_male_ 15-19
- age_group_ 25-29
- age_group_female_ 25-29
- age_group_male_ 25-29
- age_group_ 29-24
- age_group_female_ 29-24
- age_group_male_ 29-24
- age_group_ 30-34
- age_group_female_ 30-34
- age_group_male_ 30-34
- age_group_ 35-39
- age_group_female_ 35-39
- age_group_male_ 35-39
- age_group_ 40-44
- age_group_female_ 40-44
- age_group_male_ 40-44
- age_group_ 45-49
- age_group_female_ 45-49
- age_group_male_ 45-49
- age_group_ 50-54
- age_group_female_ 50-54
- age_group_male_ 50-54
- age_group_ 55-59
- age_group_female_ 55-59
- age_group_male_ 55-59
- age_group_ 60-64
- age_group_female_ 60-64
- age_group_male_ 60-64
- age_group_ 65-69
- age_group_female_ 65-69
- age_group_male_ 65-69
- age_group_ 70-74
- age_group_female_ 70-74
- age_group_male_ 70-74
- age_group_ 75-79
- age_group_female_ 75-79
- age_group_male_ 75-79
- age_group_ 80-84
- age_group_female_ 80-84
- age_group_male_ 80-84
- age_group_ 85+
- age_group_female_ 85+
- age_group_male_ 85+

The solution presented here is based on a list of static tables stored in the same folder as the shiny application files. These can be updated manually, and the application will include new files on reload. Another option is to store the data tables inside a PostgreSQL database. The only changes in this respect include modifying the listing module of the platform for it to connect to the database and list the available tables and modifying the processing module to import geometries from the spatial database.

6.3. Technical requirements for PHC dissemination applications

The standard set-up as recommended in the Output 4.1b: "*Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA)*" is sufficient. No extra hardware requirements are necessary.

For running the applications, a valid R installation[7] is required as described here:
https://www.digitalocean.com/community/tutorials/how-to-install-r-on-ubuntu-18-04
https://www.digitalocean.com/community/tutorials/how-to-install-nginx-on-ubuntu-20-04
https://www.digitalocean.com/community/tutorials/how-to-set-up-shiny-server-on-ubuntu-20-04 .

As well as a valid shiny open sources installation as described here:
https://www.rstudio.com/products/shiny/download-server/ubuntu/

For all of them default configuration is sufficient. The applications have been developed and tested for internal use only, however INS may use it publicly if they wish. Nevertheless, any required changes to the IT configuration are the full and sole responsibility of INS.

Preinstalled requirements at a glance

- Server running Ubuntu OS
- R software environment
- Shiny Server Open Source
- Additional R packages: tidyverse, data.table, shiny, shinydashboard, shinyjs, rmarkdown, knitr, tinytex, sf, scales, leaflet, DT, shinyFilters, mapdeck

A new installation of the same configuration should be possible, provided the user has a basic understanding of the Ubuntu operating system, the command line, and the R programming language. The basic[8] steps for replicating this installation are:

- Installing the Ubuntu on the new server and configuring the system to work inside the INS network
- Installation of R from CRAN
- Installation of Shiny Server

Other required resources

PHC data processing was recommended to be done and prepared using applications based on the statistical programming language R. It is important and becoming mandatory that INS has to enlarge and develop a department that aimed at building capacity in relation to the R environment. The statistical offices that use these tools have a small department within the IT dedicated to R.

Inside INS should be mixed teams between personnel with R programming & development skills and statisticians on certain fields of interest (social, economy, etc.) to further develop and implement R solutions and use them for statistical production. The RStudio Server and Shiny Server could be installed for each user to use them regularly in their activity. The regular users of applications can take courses (as the ones proposed through Multiannual Training Plan[9]) and together with the IT department could expand or improve the applications proposed.

---

[7] https://cran.r-project.org/doc/manuals/r-release/R-admin.html

[8] Full documentation and description of this process can be found here:
How to get your very own RStudio Server and Shiny Server with DigitalOcean (deanattali.com)

[9] Multiannual Training Plan for INS staff was prepared by WB team under the Output no. 12 of same RAS Agreement and contains mandatory or specific IT courses for statisticians, including the one on R programming language.

7. PHC2021 - concluding about the actual implementation

The overall census participation figures (data collected) have confirmed that "everybody counts" for the census and that efforts are rewarded when solving the challenges of a first-time paperless census implemented in Romania, a country of 19.02 million inhabitants, out of which 51% of the population is located in rural areas. A strong and reliable integrated system for PHC data collection was put in place and operated to address the challenges and ensure a smooth process of data collection both in the CAWI and in CAPI phases for all respondents, whether they were located in urban or rural areas, digitally educated or not, whatever ethnicity they had or being native speakers in Romanian or other mother tongues registered in country (e.g. the electronic questionnaires were made available for all 53 registered ethnic groups and in 17 languages used around the country).

The integrated IT system and data infrastructure support system for the census met its objectives: first, to support all aspects of the census process from start to finish, from data collection to data dissemination; second, given the tight calendar which the census followed, and the time limits on various parts of the census process, it provided reliability and stability; third, it provided security and confidentiality of the data collected; and, fourth, the system demonstrated that it was cost effective keeping in perspective that the IT infrastructure required for a large project such as the census may not be needed on an ongoing basis. The time, the effort, and resources for data availability and digitization, validation and imputation were considerably reduced from 24 months, as in 2011 for the previous PHC in Romania implemented with Pen-and-paper Interview (PAPI) data collection, to up to 4 months nowadays. Nevertheless, the PHC IT infrastructure was, is and will remain a component of the integrated IT system of the INS, at least from the perspective of the intercensal surveys.

From the perspective of the CONRENA project, the PHC2021 concentrated substantial technical assistance delivered through fifteen (15) outputs, out of which:

- Ten (10) outputs represented direct assistance and dedicated to the PHC:
  - Output 3a: *Report on advisory services provided to Recipient on the Note with recommendations on terms of reference for the PHC2021 promotion campaign*
  - Output 3b: *Report on advisory services provided to Recipient on the Note on planning, management and implementation of the PHC2021*
  - Output 3c: *Report on advisory services provided to the Recipient on the Notes on reviewed legislation for PHC2021 (four (4) draft notes with recommendations, one (1) implementation plan) including the report on two (2) workshops on PHC2021 legislation*
  - Output 4.1c: *Report on advisory services provided to Recipient on the Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out);*
  - Output 7a: *Report on advisory services provided to Recipient on Methodology to assess and promote continuously and in real time the quality and coverage of the collected data PHC2021 and data protection/security;*
  - 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;*
  - Output 10a: *Report on Recommendations to the Recipient on how to perform the PHC2021 piloting process;*
  - 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 ;* and

- 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 the actual report
- Output 13: *Report on advisory services provided to Recipient on preparing the Recommendations and guidance on documented plan for the integrated system for the PHC2021 implementation (including user management, ID encryption, disaster recovery, maintain system availability in case of breakdown);*

and

- Five (5) output were cross supporting PHC and other components of the CONRENA project (GAC, intercensus, SICCA).
  - Output 1: *Report on advisory services provided to Recipient on the Analysis of NSS operating capability for producing Official Statistics and the achievement of the general agricultural census and the population and housing reconciliation*
  - Output no. 4.1b*: Report on advisory services provided to Recipient on the Recommendations to develop the technical documentation in view of organizing the procurement of an integrated IT system including, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA)*
  - Output no. 4.2a: *Recommendations and guidance provided to the Recipient in running the integrated IT system with its components, hardware (including tablets) and software (licenses), including data storage, protection and security of data for running the activities developed under the RAS (GIS, PHC2021, GAC2020, inter-census periods, SICCA);*
  - Output 6e: *Report on advisory services provided to Recipient on the Report on two (2) five-day workshops for NIS statisticians in methodologies and tools for transforming questionnaires in intelligent statistical e-questionnaires for inter-census*
  - Output 12: *Report on advisory services provided to Recipient on the Development of a multi-annual training needs plan (2021- 2024) for NIS including STDs based on training needs; carrying out of: one (1) two (2)day training for NIS specialists and relevant NSS stakeholders, three (3) three (3)-day trainings on the working method for developing European Regulations and one (1) five (5)-day training on SDC for NIS staff; recommendations to identify appropriate institutions where study tours on statistics could take place; two (2) study tours; and conference on good practices at EU level*

The PHC, round 2021, was prepared and implemented in a joint effort of the INS headquarters and DTSs and other public authorities during the past five years, and demonstrated the importance of continuous collaboration with stakeholders, strong and coherent planning, and the benefits of innovation of methods, tools and applications for statistics production.

8. Annexes

Annex 1 - PHC data collection information system – As-built report.
Annex 2 - Sample reports of daily monitoring (CAWI, CAPI).
Annex 3 - Sample payment reports CAPI
Annex 4 - PHC data collection validation and approval scripts (CAWI, CAPI).

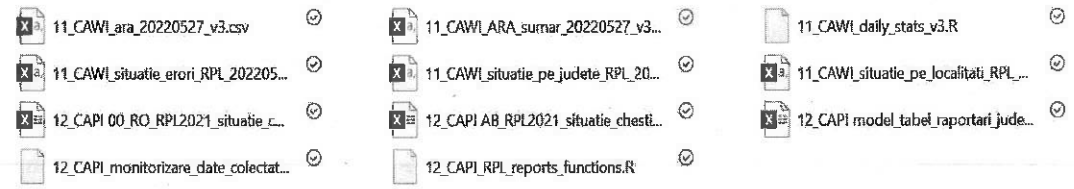
All annexes are provided archived in an electronic format to this report.

Annex 1 - PHC data collection information system – As-built report.

The annex is provided archived in an electronic format and contains the files listed below:

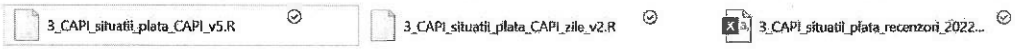
Annex 2 - Sample reports of daily monitoring (CAWI, CAPI).

The annex is provided archived in an electronic format and contains the files listed below:



Annex 3 - Sample payment reports CAPI

The annex is provided archived in an electronic format and contains the files listed below:



Annex 4 - PHC data collection validation and approval scripts (CAWI, CAPI).

The annex is provided archived in an electronic format and contains the files listed below: