UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

POCA
Programul Operațional Capacitate Administrativă
Competența face diferența

Instrumente Structurale
2014-2020

ROMANIA

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

OUTPUT No. 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

October 2022

Revised November 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.

Copyright Statement

This report was delivered in October and revised in November 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 corresponds to Output 10 under the above-mentioned agreement.

Table of Contents

Table of figures

List of Acronyms

API	Application Programming Interface
CAPI	Computer-assisted personal interview
CAWI	Internet-computer-assisted web interview
DMZ	De-Militarized Zone
HTTPS	Hipertext Transfer Protocol Secure
INS	National Institute of Statistics
IT	Information Technology
PIN	Personal Identification Number
PHC	Population and Housing Census
RAS	Reimbursable Advisory Services
SDC	Statistical Disclosure Control
SHA-1	Secure Hash Algorithm 1
STS	Special Telecommunications Services
SuSo	Survey Solutions
VPN	Virtual Private Network
WB	World Bank

Introduction

The purpose of this report is to present the ***set of draft statistical census tools using multi-modal methods to promote data protection and 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 the applied and used methods to avoid security risks in the CAWI and CAPI stages and processes of data collection (e.g. pre-enumeration, questionnaire fulfillment, data transfer to self-enumerated persons/enumerators, data transfer between enumerators, supervisors, headquarters), and on dedicated environments (e.g. data collection environments and data processing environment) that supported the Population and Housing Census (PHC), round 2021.

The report has three sections. The first section provides an overview of the PHC data collection system's physical and logical protection and security measures that were applied during the census at cross-organizational level within the participant organizations (INS, STS, service providers).

The second section refers to the methods applied for data security during the census through the set-up for security on the STS and INS IT environments through the Survey Solutions features - the data collection solution used during the CAWI and CAPI stages of the census and the specific methods used for retrieving data from the administrative sources in preparing and implementing the census.

The third section presents several recommendations derived from the census implementation, regarding actions needed for data protection and security in view of the future surveys to be implemented by INS.

Complementary to this report, additional information can be found in other reports submitted to INS as part of this project, in particular the following:

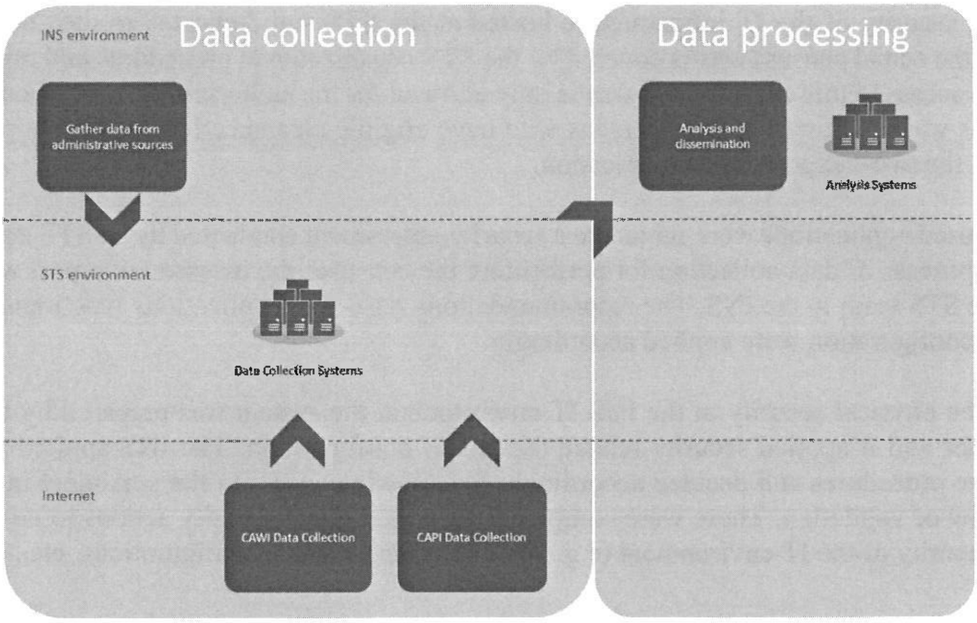
- Output no. 4.1c: *Report on advisory services provided to Recipient on the Documented plan for the integrated system for PHC2021 implementation (details how the IT infrastructure implementation for PHC2021 will be carried out);*
- Output no. 10a: *Report on Recommendations to the Recipient on how to perform the PHC2021 piloting process;* and
- Output no. 7a*: Methodology to assess and promote continuously and in real time the quality and coverage of the collected data PHC2021 and data protection/security.*

1. Overview of the PHC data protection and security

The PHC data protection and security relies on a variety of measures applied at cross-organizational level within the participant organizations (INS, STS, service providers) and targeted IT systems configurations carrying sensitive data from the end points data entry (respondents and interviewers) to secure storage and data processing.

An overview of the data flows between different stages of the PHC (see Figure 1) exposes the processes, environments and areas associated with potential risks for data security during collection and processing.

Figure 1 - Data flows during PHC stages



The risks of affecting the physical security and the logical security are the two main categories of risks for the components and processes of the entire system used during data collection and processing and, consequently, these will have effect on the data security and protection (see figure 2).

Figure 2 - The components potentially affected by the risks

Environment	The components of census data collection and processing exposed to Physical & Logical security risks
INS	➢ Applications and solutions used for gathering and transferring data from administrative sources; ➢ Applications and solutions used for data analysis during collection and daily reporting, investigating the information provided during self-enumeration in cases of malfunction of the system; ➢ Validation of data collected by the enumerators; ➢ Tablets and data stored by the enumerators during CAPI data collection (assuming the enumerators as part of INS data collection environment);
STS	➢ Core IT hardware and software solutions that compose and ensure the running of the data collection system (in primary and secondary data centers);
Internet	➢ Survey Solution Web interface for CAWI data collection (including preregistration); ➢ Survey Solution Web interface for CAPI data collection and validation

During the census, the system's physical security relied on security controls implemented on the STS datacenters and on the INS and the WB internal procedures and regulations for the users in their accessing the system from terminals hosted by either the INS and/or the WB.

The system logical security had different levels of protection:

- the STS Network firewalls were protecting access to the systems hosted in STS datacenters;
- all machines have implemented hosts firewalls which allow only the minimal traffic required for system functionality;
- the Windows Servers were protected by Windows Defender native anti-virus;
- the administrative access to the system was protected by restricted VPN point-to-site solution, available only to authorized personnel.

The physical security of the IT infrastructure hosted at the STS and dedicated to data collection (and used in the actual census) was/is covered by the STS organizational procedures and protocols. The physical access to this environment was/is only allowed for the authorized STS personnel. The logical access was/is allowed only to persons who have eligible clearance level, and it is possible only through the two-factor VPN authentication.

The web exposed applications were subject to a security assessment conducted by an STS dedicated team. In the process of data collection for performing the censuses the assessment report was presented by the STS team to the INS. The recommendations regarding applications functionality and applications configuration were applied accordingly.

In terms of the physical security of the INS IT environment, the system was prepared by the INS and the vendor and it applied security related aspects of configuration. The INS applied its own administrative procedures and decided accordingly for allowing access to the personnel in charge with data view or validation. These were completed with the logical security actions to ensure the robustness security of the IT environment (e.g. VPN solution, systems configurations, etc., already in place).

The data collection process included sensitive information regarding the PIN of the enumerated persons. This was secured through pseudonymization by a HASH function with the SHA-1 algorithm.

2. Methods applied for data security during the census

2.1. Survey Solutions data security features

Survey Solutions is conceived with the understanding that it will be used in applications involving handling personally-identifying data (PID) such as names, addresses, phone numbers, GPS locations and other sensitive information about the respondents such as income, consumption/spending, savings, etc. Unauthorized access to this data may result in reputational risks for the organizations conducting the survey and negative consequences for the persons whose data has been misused.

The attempts to misuse the data can be coming from:

- persons involved in the survey (insider attacks);
- persons not involved in the survey (hacker attacks).

Survey Solutions reduces the risks of **insider attacks** by limiting the users' access to certain information. One of the effective mechanisms of doing this is partitioning the work into multiple facets implemented by different teams. This split of work is implemented hierarchically, with each team responsible for only a dedicated segment, and thus not having access to other segments. While access to all data is maintained for some users coordinating the work of the teams, this compartmentalization reduces the possibility of searching for a specific person's data (such as data of a famous politician, athlete, or other person of interest).

Data manipulation risks are further reduced through the concept of **user roles**, which are issued to every user that is authorized in the system. The roles are prescribed to every account and only once (during the account creation). The role determines which functionality of the program is available to the user holding this account. For example, the possibility to approve the interviews is restricted to accounts in roles *Supervisor* and *Headquarters*, but access to answers to individual questions is structured in such a way that the *Supervisors* and *Headquarters* users may read the answers, but not modify them, while the users with the account type *Interviewer* are empowered to both read and edit operations.

The assignment of roles is the privilege of the *Administrator* user, who is also empowered to create all the other users' accounts, reset their passwords and perform other account-management tasks. Four persons have administrators' rights.

The product of data export from Survey Solutions is an archive of the database containing all information obtained as part of data collection. This product is of the highest value, as it contains all responses to all the questions (including the PID). Potential leak of this file is one of the highest severity threats. Only users with the roles *Headquarters* and *Administrator* have a possibility to initiate the data export process and acquire its product (this action can also be performed using robotic API accounts, which are optional in Survey Solutions). This **downloadable archive can thus be additionally protected with a master password** that the *Administrator* of the Survey Solutions data server can assign (separately for each workspace). The use of the password on the downloadable archive further reduces the number of the users who have access to data to only the Administrator himself, and the users to whom he entrusts the master password (typically a backup admin).

Protecting the account password is the user's responsibility. To ensure that the password is not known to any other user in the system Survey Solutions enforces **changing the password when it**

is used for the first time to a new password, that is known only to that user. This is done before any work with that account is carried out and the action of changing the password is logged.

To further increase the security of passwords a second mechanism of authentication is possible, where the users can selectively activate **2-factor authentication (2FA)** for their accounts. In that case to log in to the server the account holder must provide the correct password and also a correct token which is issued by an authenticator app running on the user's phone or PC. A number of open source alternatives and implementations of such authenticator apps using TOTP (*time-based one time-password*) algorithm are available from major software vendors (Microsoft, Google, etc.) (2FA is not implemented on the tablets).

Users that need their password to be reset or 2FA deactivated must contact the server *Administrator*.

The **password policy** that applies to all passwords in Survey Solutions (minimal password length and set of characters) is configurable outside of the software in the configuration file. By default, it requires passwords of length 10-12 characters with at least one small, one capital letter and one digit present. The same policy applies to passwords that the users specify as their own when they change them.

Survey Solutions system also has built-in defense against unauthorized access (**hacker attacks**). The primary entry point to any password-protected system is the login page, which screens the users into authenticated and invalid. This page theoretically may be subjected to brute-force attack (probing large number of combinations of user names and passwords). To discourage such attacks the software reacts on invalid login attempts: Using counters of login attempts (specific to each account) when a certain number of invalid login attempts is detected a **CAPTCHA challenge** is presented to the user. If the account is being attacked by a robot, this is an effective way to prevent it from continuing brute force attacks. If the combinations are being tried by a person, which can solve the CAPTCHA challenge, then that person will have an additional (configurable) number of attempts before the account is locked for any login attempt for a fixed amount of time (configurable). A combination of CAPTCHA and account locks allow effective prevention of denial-of-service (DOS) attacks, where a malicious user knowing a valid login name cannot effectively disable that account by entering invalid login + password combinations.

There is no CAPTCHA or account locking on the tablet (Interviewer and Supervisor's applications).

During the CAWI sessions of data entry the communication of data from the user's browser to the data server is protected by an **SSL-certificate**, which enables secure exchange of the information using HTTPS. While Survey Solutions can technically work over regular HTTP such use is discouraged as it leaves the exchange traffic unprotected to interception attacks.

The data communicated in the CAPI sessions is protected by the same technology (secure communication over HTTPS).

The **data stored in the tablet** (during interviewing sessions and also for postponed interviews) is secured with the following two mechanisms:

1. answers to any questions are encrypted;
2. data is stored in a 'private storage', which is an application-specific isolated storage in the Android OS.

The debug packages of data that can be sent to the server by the staff using tablet applications also contain data in encrypted form where that data reflects the answers of the respondents. These packages can be seamlessly decrypted on the server to which they are sent by the server administrator. This happens automatically and without any need of special actions by the user.

To prevent dumping the whole tablet data the **manifest of the Android** app contains flags indicating that the application should not be installable to a removable media or backed up using Android's built-in application backup and transfer facilities.

To prevent pushes of data from other devices each interviewer account is linked to a single physical device tracked by its internal unique number. While the interviewers may switch to a different device (re-link) in case one gets lost, stolen, or physically damaged, that device will not be allowed to push the data to or receive the server from the server in case someone who finds it attempts to do so.

Survey Solutions software makes each server appear unique by generating a random signature during the installation process. While this is not a security feature *per se*, it helps prevent accidental mistakes of the interviewers of sending the data to a wrong server.

The source code of the software (both the mobile version and the server-side application) is available in its entirety on a public site which allows experts to freely investigate possible attack vectors and provide comments to the developers and users alike.

Additional measures that increase the security of the whole system when implemented are:

1) Making sure that the login is activated on the interviewer devices (Android OS lock screen is protected with a PIN or password or other similar method). This will further protect confidential data by creating yet another level of protection.
2) Encryption of the DB on the server, for example, by placing it on the disk volumes protected with OS built-in or 3rd party data encryption tools.
3) Since the database itself is not encrypted in the server room, physical security and maintenance personnel screening must be in place to prevent any unauthorized access at the physical level.

The main aspects that should be remembered are:

- To protect data in transit survey solutions uses HTTPS protocol which is industry standard for such tasks. It is in used in both CAWI and CAPI data collection phases.
- In order to access any collected data users always need to provide their login and password.
- Passwords are stored in survey solutions database using salted hashes so it cannot be reversed even in case of receiving a hashed value of password.
- Supervisor and Headquarters logins also have an ability to enable two factor authentication to add additional layer of protection.
- Login page has protection against brute force attack by use of locking mechanism and requiring completing captcha in case of multiple failed login attempts.
- Data at rest on tablets is stored inside isolated android application storage and has additional encryption that is applied to all collected data until it is sent to Headquarters module of the Survey solutions.

2.2. Data security during the census

The methods applied for securing the data during retrieval from the administrative sources are:

- Information from the administrative sources is stored inside the internal network of the INS and not exposed publicly.
- PINs are stored as hashes so the self-registration portal can only retrieve a single record from source after the user provided his actual PIN.
- After the completion of form, the user receives an email with a unique link. The link is valid until the related interview is completed. After completion only INS has access to the collected data by using the Administrator or Headquarters credentials.
- Another security configuration applied to the self-registration phase was the masking of the PIN in the self-registration e-mail notifications. This configuration was applied at the template level of the self-registration application. The mask of the PIN consists in the hiding of 5 digits, from the number 6th to the 10th, as follow 12345xxxxx123.

Data collected during the CAWI stage is available to respondents until the interview is completed. After the completion only INS has access to the collected data by using Administrator or Headquarters credentials.

Data collected during the CAPI stage, data on tablets, as described in 2.1, is stored in an encrypted form until it is sent to Headquarters. After being sent the interview answers are deleted from the tablets and stored only in the central database of Survey Solutions.

Regarding the micro-data security, only the Headquarter account has access to the full collected data file. Supervisors only have access to interviews that were completed by their team. After the interview is approved by the Supervisor, it becomes visible only to users with Headquarters role.

For the aggregated data security, the INS uses internal or partnership protocols (e.g. STS) for transferring the data from the data collection systems to the own IT environment (by case) and applies the rules accordingly.

The data collection process was monitored with the Survey Solutions tool available as a component of the platform. In the CAWI phase of the PHC, the validation scripts were run to determine the completeness of the collected data from self-enumeration. Daily, based on these validations, a data set was created for presenting the information on self-enumeration. During this process the PINs of the persons were and are pseudonymized by a HASH function with the SHA-1 algorithm.

In the CAPI phase, reports were produced by R scripts running on an analytics server, based on exported data from Survey Solution for the daily monitoring of the Interviewers' activities and the aggregation of the data at counties level. The validation scripts run daily on the collected data through the CAPI method in order to determine the errors and approve the validated questionnaires.

The approval of the questionnaires identified as correct by the scripts was done by API calls. The validation rules were established by the INS team and implemented in RStudio on the analytics server. During the monitoring activities no PIN analysis was done, and the algorithms checked the minimum number of characteristics of the statistical units collected in the PHC.

For the anonymization of the collected data and to prepare the public use dataset for dissemination, an SDC procedure as outlined by Eurostat is applied[1]. The SDC is part of the tools considered for data protection and security and will be applied by INS following the concepts described by the WB team during training in December 2021 (five (5)-day training on SDC for the INS staff under Output 12), and the methods chosen by the INS as part of the technical assistance received from the WB during June 2022 under Output 10b: *Report on advisory services provided to the Recipient on the Technical assistance and best practice recommendations for SDC for training INS staff, data confidentiality, and ways to secure micro-data and aggregate data.*

2.3. The IT environment set-up for security

2.3.1. STS hosted IT environment (data collection)

The STS environment used for data collection was a dedicated environment for the data collection stage of the PHC. It consisted of a DMZ hosting two load balancers which forward web requests from citizens and interviewers to the applications servers. A pool of five application servers carried those requests and distributed them across applications (SuSo, self-registration). The data was stored in a Postgresql 12 database consisting of a cluster of two nodes on active-passive configuration. The same structure was replicated in a dedicated secondary STS datacenter hosted in another location.

Apart from the server infrastructure, the dedicated IT environment was protected from external access using the STS standard firewall configuration (the STS internal standards and procedures do not allow third parties access to those configurations and are referred to as "standard") and 2-Factor VPN access for the authorized personnel.

The access to the applications through the load balancers was made using the SSL connections. The SSL certificates were installed on both load balancers and the access from the Internet was possible only by using HTTPS.

During the CAWI and CAPI phases of the PHC, different security configurations were applied at load balancers levels (they are available in **Annex 1** – the PHC data collection system - As Build-report).

In the self-registration (CAWI) phase, due to the INS requirements that the self-registration portal (autorecenzare.insse.ro) had to be accessible also from Internet addresses outside Romania, additional security conditions were applied at the load balancers configuration, restricting the access to SuSo only for interview completions. The restriction was applied to the NGINX engine using a regular expression which forbids access to other functionalities of the SuSo from external (Internet) locations:

```
((\/[Ww]eb[Ii]nterview\/|\/js\/|\/fonts\/|\/CompanyLogo\/|\/[Ii]nterview\/|\/interview$|\/WebInterviewResources\/|\/css\/|\/img\/|\/locale\/|\/pdf$))
```

This regular expression is used for filtering requests only to allow SuSo paths relative to interviews completion.

[1] [1] https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

In the CAPI phase, the IP-geo-blocking access restrictions were applied at national level and the self-registration application was disabled.

All the servers of the STS-hosted infrastructure were configured to use also host firewalls (UFW on Ubuntu 20.04 machines and Windows Defender Firewall on Windows 2016 machines). Only the necessary ports needed for the applications' normal function were open. For details, please refer to **Annex 1** – the PHC data collection system - As Build-report.

At the beginning of the PHC, all servers operating in the systems were updated to the latest security updates.

The applications hosted on the application servers nodes were subject to a security assessment performed by the STS and the assessment results were applied accordingly - for reference see the files distributed by the STS to INS – *Raport de evaluare a vulnerabilităților aplicatia web rplcapi.insse.ro, 2022,02.14 (En: Assessment report of the vulnerabilities of the CAPI web application)* and *Raport de evaluare a vulnerabilităților aplicația web autorecenzare.insse.ro, 2022.02.20 (En:Assesment report of the vulnerabilities of self-registration web application).*

The STS-hosted environment used for the data collection was connected through a secure VPN tunnel with the INS-hosted infrastructure – a task implemented by INS through own procedures. The only permitted connection between the INS IT environment and the STS IT environment is a https connection (over encrypted VPN), respectively between the analysis server and the load balancers, in order for the data processing operators to receive data from SuSo (API access for exports and data validations).

2.3.2. INS-hosted IT environment (data processing)

The newly acquired INS's infrastructure consists of a VMWare cluster environment allowing for the creation of virtual machines. The only component communicating with the data collection infrastructure hosted by the STS is the analysis server which relies on a high-computing Ubuntu machine with the necessary tools (R-Studio, Shiny server) needed for data manipulation after and during the data collection process.

The newly acquired INS infrastructure is protected by a VPN, allowing only authorized personnel to access its components. The INS manages the VPN access accordingly to internal procedures.

The analysis server is configured with individual accounts in order to track each person's access and there are shared spaces used to keep the collected and processed data with the necessary permissions in order to allow only responsible persons to read or manipulate it. The analysis server's individual accounts are managed by INS according to internal procedures.

3. Recommendations for future surveys' data protection and security

For future population censuses, a particular focus should be on the testing of the security of the SuSo application and self-registration component. All the selected releases of SuSo and self-registration should be subject to security assessments carried on by independent entities.

The scheduling of the security testing for next censuses should be planned carefully with enough time in advance in order to allow responsible teams (SuSo, self-registration) to apply the recommendations from the security assessment into new releases or branches and to properly close the potential security issues.

As for the PHC, round 2021, the data collection components were hosted by the STS. The same approach is recommended for future censuses. In case that the INS intend to use its own infrastructure for data collection, it needs to take into consideration the challenges that come with such intention which consist, but are not limited to:

- existence of a secondary datacenter for disaster recovery scenarios;
- implementation of procedures for enabling 24x7 support for incident response;
- developing the technical capacities of a team with proper knowledge in order to support high availability configurations, security configurations (including but not limited to intrusion prevention and intrusion detection systems); and
- performance challenges related to systems, applications and databases.

Finally, each security and data protection concern should be centered on the agreed census flow. The flow itself should be analyzed from security and data protection perspectives in order to avoid any breaches that might compromise the overall system security and data protection.

4. Annexes

Annex 1- PHC data collection system - As Build-report

UNIUNEA EUROPEANĂ

POCA
Programul Operațional Capacitate Administrativă
Competența face diferența!

Instrumente Structurale
2014-2020

Competence makes a difference!

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