CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Session 6

29-31 October 2019, the Hague, the Netherlands 22 October 2019

Ensuring data confidentiality of All-Russian Population Census 2020

(Rosstat, Russian Federation)

ATroitskaya@gks.ru

Abstract and Abstract

During the preparation of Census Round 2020, Rosstat will continue the practice of protecting the microdata confidentiality, that began in the 2002 and 2010 Censuses. Rosstat provides users with access to the Census microdata database. At the same time, much attention is paid to ensure the confidentiality of personal data. The primary information in the database is impersonal, but there is still the possibility to identify a specific respondent by unique characteristics. Such identifiers can be a date of birth or a nationality that isn't common among the residents of a settlement, etc. First of all, the risk of disclosure of personal data is high for residents of settlements with a small population. To protect personal data, Rosstat uses a method of data distorting (perturbation) in the summary tables. This makes impossible to obtain the information about a particular respondent or household. This approach allows to provide access to the microdata base of All-Russian Population Censuses of 2002 and 2010 to any Internet user.

JOINT UNECE/EUROSTAT WORK SESSION ON STATISTICAL DATA CONFIDENTIALITY
29-31 October 2019, The Hague, Netherlands

Rosstat, Oksana Dmitrieva DmitrievaOV@gks.ru
Rosstat, Anna Troitskaya ATroitskaya@gks.ru

Ensuring data confidentiality of All-Russian Population Census 2020

One of the main issues that is continuously linked to the use of any microdata is their confidentiality. Any use of microdata, even in anonymized form, may pose a threat to privacy.

Statistical data confidentiality protects the government or governmental data:

• from transferring the individual information collected during statistical surveys;

• from the disclosure of information that was transmitted to official statistics by third parties solely for the purpose of producing statistical data.

Thus, provided:

• guarantees of confidentiality of the personal lives of individuals.

Protection of confidentiality is one of the main responsibilities of the Federal State Statistics Service (Rosstat).

1. Legal Data Protection Issues

Confidentiality protection mechanisms are based on legal acts and regulatory standards applicable in the Russian Federation.

Information collected for statistical purposes, regardless of whether they are collected as part of mandatory reporting or are provided on a voluntary basis, can only be used for statistical purposes.

The main law governing the protection of personal data during the 2020 All-Russian Population Census is the Federal Law "On the All-Russian Population Census" from January 25, 2002 № 8-FZ.

The protection of information at the state level is regulated by the following basic regulatory acts:

1. Federal Law from July 27, 2006 № 149-FZ “On Information, Information Technologies and Information Protection”.

2. Federal Law of the Russian Federation from July 27, 2006 № 152-FZ “On Personal Data”.

3. “Requirements for the protection of personal data during its processing in personal data information systems” (approved by the Decree of the Government of the Russian Federation from November 1, 2012. № 1119).

Decree of the Government of the Russian Federation from March 21, 2012 № 211 also approved a list of measures aimed to ensure the fulfillment of obligations described in the Federal Law "On Personal Data" and regulatory legal acts adopted in accordance with it and by operators of personal data that are state or municipal bodies (hereinafter – the List).

In accordance with the List, one of the measures aimed primarily at minimizing the risks of harm to specific citizens in the occasion of leakage of their personal data from personal data information systems is the depersonalization of personal data in accordance with the requirements and methods established by the authorized body for protection rights of personal data subjects.

Roskomnadzor (the Federal service for supervision of communications, information technology and mass media) established requirements and methods for anonymizing personal data processed in personal data information systems in the Russian Federation. The established requirements and methods for anonymizing personal data will be applied during the 2020 All-Russian Population Census.

2. Ensuring data confidentiality in the automated system for the preparation of the All-Russian Population Census 2020

Focusing on the positive experience of conducting Censuses in many countries of the world, and also taking into account the positive dynamics in the development of information technologies in Russia, Rosstat plans to conduct All-Russian Population Census - 2020 using the experience gained in ensuring data confidentiality and three methods of collecting information about the population:

1) self-completion of electronic questionnaires in the Internet by respondents.

2) the completion of electronic questionnaires on tablet computers with specialized software installed by interviewer.

3) the traditional completion of paper questionnaires (two-sided machine-readable Census forms).

In each entity of the Russian Federation, all three methods of primary data collection All-Russian Population Census -2020 will be used.

Ensuring data confidentiality will be achieved by organizational and technical measures.

The main stages of All-Russian Population Census -2020 to be automated are:

• preparation for All-Russian Population Census -2020;

• running All-Russian Population Census -2020;

• automated processing of materials;

• results of All-Russian Population Census -2020;

• publication of the results of All-Russian Population Census -2020;

• post-census use of All-Russian Population Census -2020 materials;

• storage of the results of All-Russian Population Census -2020.

The main task is to develop algorithms for using data confidentiality tools in an automated system for preparing, conducting and summarizing All-Russian Population Census -2020 using laptop computers and Internet, to maximize the protection of confidential data from unauthorized access in the field, district, regional and federal level of AS VPN (Automated system of All-Russian Population Census).

To ensure the collection of information about the population via the Internet, data collection will be organized by the Federal State Information System "Unified Portal of State and Municipal Services (Functions)" (EPGU), the Federal State Information System "Unified System of Interdepartmental Electronic Interaction" (SMEV), designed to organize information interaction between the information systems of SMEV participants in order to provide state and municipal services in electronic form, the information system of Executive power (State organizations) (IS OIV) and an automated system (AS) of All-Russian Population Census (VPN).

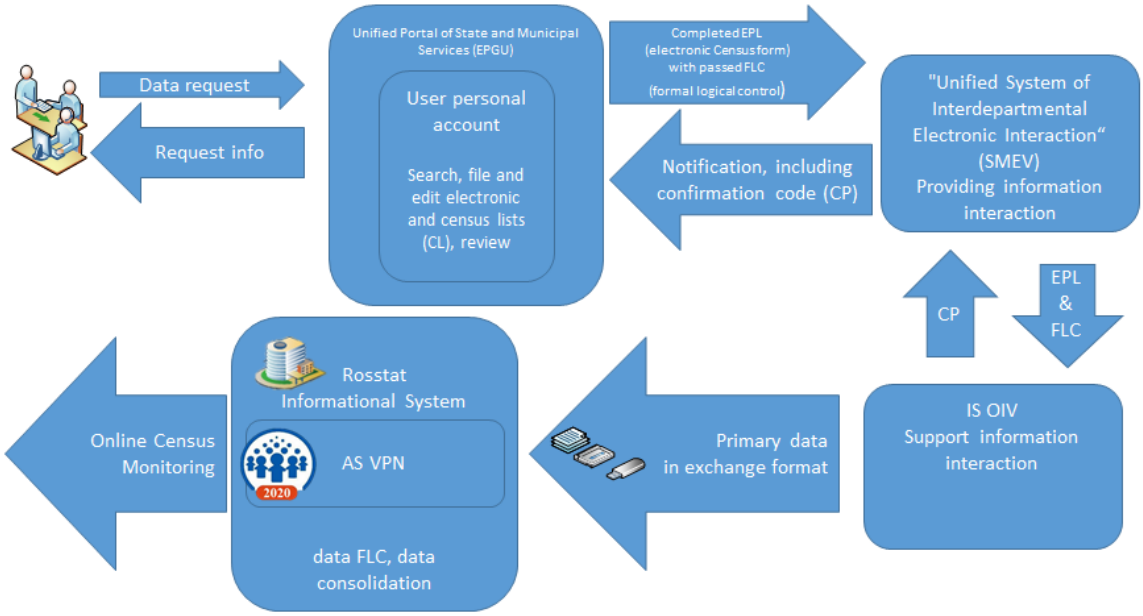

Fig. 1 - Scheme of the interaction of EPGU, SMEV, IS OIV and AS VPN

At the stage of collecting information about the population by self-filling respondents with electronic questionnaires (EPLs) in the Internet, the following algorithm and methodology for using data confidentiality tools will include:

- secure data transfer methods;

- secure data transfer protocols;

- protection options for the collected confidential data when they are stored on servers.

The EPGU website already has means of ensuring data confidentiality, as well as means of protection against DoS attacks.

The SSL secure data transfer protocol, in addition to providing secure data transfer, also allows client authorization using client SSL certificates.

To transfer a client certificate, a PKCS № 12 format file is used.

The PKCS № 12 format is a format intended for encrypted storage of a key pair (private key and certificate), which is recognized and used by many browsers and mail agents.

The IT complex is a combination of hardware and software designed to perform the basic functions of protection against unauthorized access to personal computers based on:

- the use of personal users identifiers;

- password mechanism;

- blocking the OS loading from removable data storage (like USB flash drives);

- integrity control of hardware and software (general, application software and data files) PC (AS);

- ensuring the trusted boot mode of operating systems installed on a personal computer (AS) that use any of the file system supported by the complex.

3. Anonymization methods:

Area	Data class	Data type	Final recommendation
Database of the All-Russian Population Census	Census	Calculated data	Perturbation method

4. Organization of anonymized data processing during All-Russian Population Census 2020

The basis of ensuring data confidentiality during All-Russian Population Census 2020 is the experience of the 2010 All-Russian Population Census and the 2018 Pilot Census.

When the Operator (user of open data base) uses the depersonalization procedure, shared storage of personal data and depersonalized data is not allowed.

During processing anonymized data by the Operator, if necessary, de-depersonalization can be carried out. After processing, personal data (obtained as a result of such de-depersonalization) is destroyed.

The processing of personal data before the implementation of depersonalization procedures and after performing de-depersonalization operations is carried out in accordance with the current legislation of the Russian Federation with the use of measures to ensure the security of personal data.

5. General solution architecture

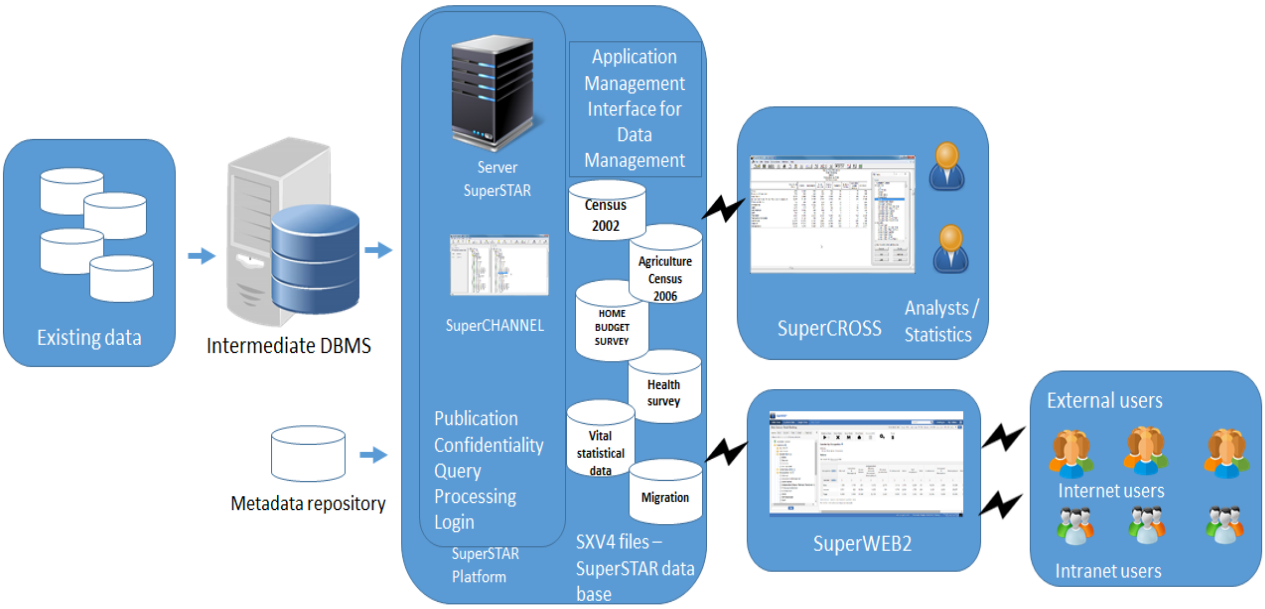

Fig. 2 – Solution architecture

6. Protection of confidential data in the databases of the All-Russian Population Census

The ability to apply data confidentiality is one of the key characteristics of the database of the All-Russian Population Census.

The open database applies privacy rules before displaying the resulting dataset. The rules are identified during the design of the system as a result of examination of the data provided for publication. Confidentiality methods are applied in cases where the aggregated data sets contain cells obtained from a small number of records. Since the number of records is limited, and if there is any information about the sources, the user can use the provided data to identify a person or household.

The database of the All-Russian Population Census contains several reliable industrial methods for adjusting and hiding values in cells to prevent personal identification.

The data perturbation (noise method) algorithm is specifically designed to conceal microdata when published on the Internet.

The specified algorithm is used in cases where the user views the tables containing indicators related to the data for a person / household, which can hypothetically identify a specific person. Using the algorithm, minor distortions are introduced into such data randomly (data noise). The method allows you to publish in the public domain a large amount of detailed information, leaving the total numbers (the sum of distorted values) close to real values.

If the built-in method (data perturbation) does not satisfy all the requirements, then the system has means to add its own data protection rules. These rules can be added using the software configuration.

The general approach of applying the perturbation method to the calculated and absolute values is as follows. Using the algorithm, minor distortions are introduced into the calculated data randomly (data noise), and small values change. The limitation of the values to be changed is determined during module configuration. In absolute terms, changes are made to the number of objects that contributed to each specific cell, and the value of the absolute value is adjusted in accordance with the perturbed number. For example, if a cell with a sum of 500 has 2 objects that contributed to the value, then if the number is changed by 3 using the perturbation method, then the amount changes to 750.

7. The architecture of the perturbation method implementation in the database of the All-Russian Population Census

The perturbation method is used in the database as an addition to the existing functions for hiding classified information (such as, for example, random rounding). The method provides more levels of protection than random rounding. A similar mechanism will be used to ensure the confidentiality of the 2020 All-Russian Population Census.

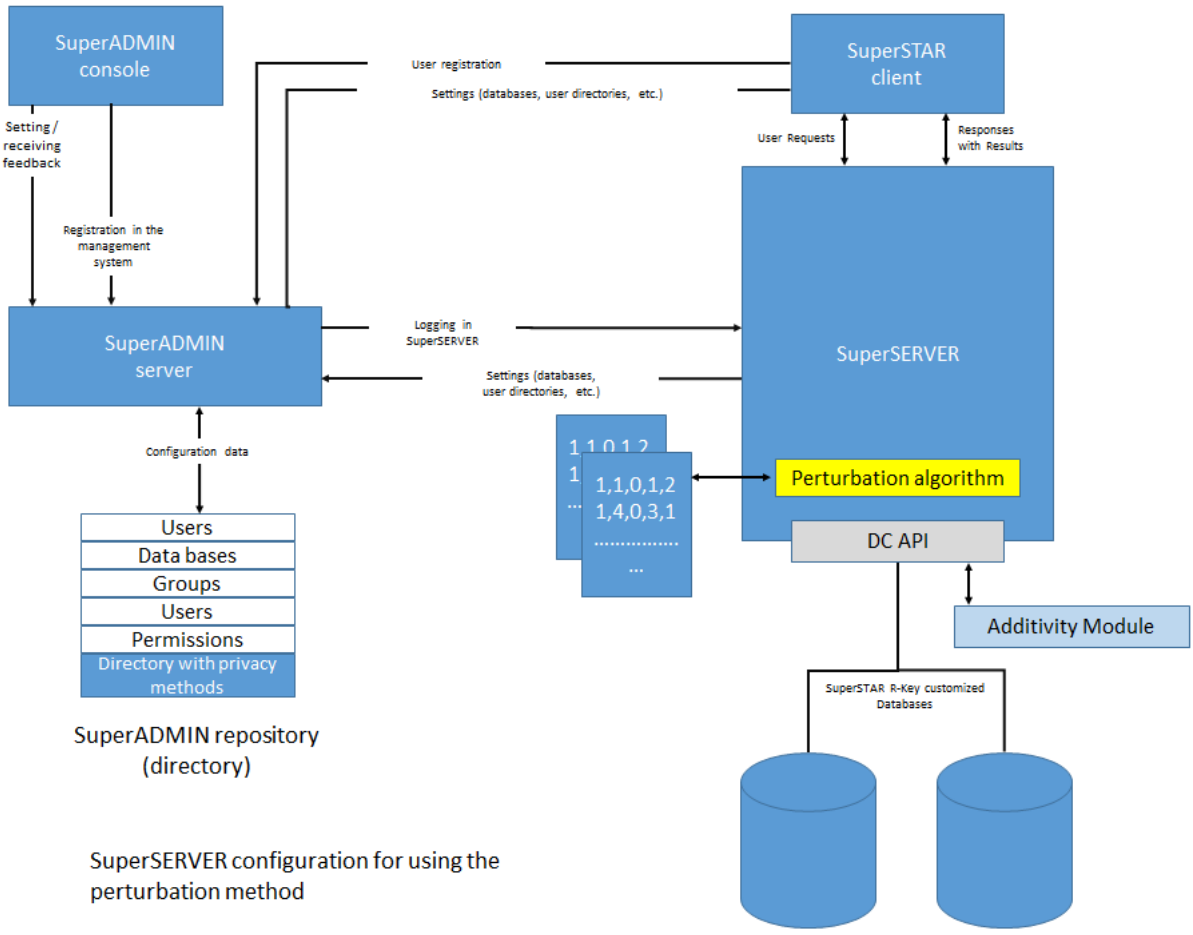

Fig. 3 - Configuration for using the perturbation method

8. Functional mechanism

The process begins with preparing the source data. An additional column is added to each data table in the initial data block. This column contains some values (R-keys) for each record in the table. The source data is then converted to All-Russian Population Census databases.

When a user makes a request through a database to output a table from a database, the request is transmitted to a server that applies a perturbation module and other data settings and returns data or a message corresponding to the request. As part of this process, the request searches for a table to obtain the correction values that will be applied to the values in the cells.

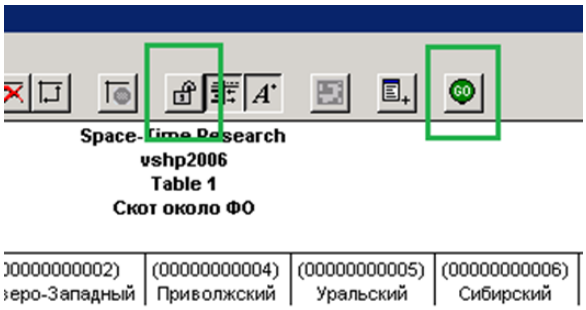

Fig. 4 - Toolbar with the image of the lock and the “Cross-tab” icon

To enable the option of hiding confidential data, its necessary to click on the icon with the image of the lock so that the lock is closed.

It will be possible to enable or disable the mode of hiding confidential data both before the construction of the table, and for an already generated table. In the latter case, to recalculate the values in the table after turning the algorithm on or off, you must click on the “Cross-tab” icon.

9. Used modules

Perturbation module

A key (called a cell key) is generated for the resulting table, and for each cell in the table. This is done using an algorithmic operation using an R-key from each unit of the record, which contributes to the value of the cell.

Additivity Module

The operation with perturbations affects the additivity of the table, so the table values are then passed to the additivity module. After that, the final table is returned to the user, containing values that have been noisy, where necessary, and for which additivity has been adjusted accordingly.

The code with the additivity module is external to the database server and is accessible through the application program interface for data control.

The following table illustrates the use of confidential data protection using an example of a sample from the database of the All-Russian Population Census:

	All population	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 and older	no age indicated
Bronnizy city	191	6	12	67	38	48	10	6	4	-
Dzerjynsky city	81	4	10	18	22	10	13	-	1	3
Dolgoprydny city	796	23	47	256	260	141	31	6	6	26
Dubna city	71	3	10	9	6	17	9	5	7	5
Zheleznodorojny city	86	4	12	16	13	16	8	4	3	10
Zhukovskiy city	3712	35	105	985	1053	884	304	105	162	79
Ivanteevka city	242	5	6	91	91	29	6	2	-	12
Klimovsk city	93	7	2	15	37	23	2	4	2	1
Kolomna city	115	10	6	35	32	12	8	6	-	6
Krasnoarmeisk city	19	-	1	8	6	4	-	-	-	-
Krasnoznamensk city	29	2	3	8	9	3	-	-	1	3
Lobnya city	531	5	35	175	175	105	17	5	3	11
Lytkarino city	62	-	3	21	12	20	6	-	-	-
Orekhovo-Zuevo city	16	1	-	6	4	4	1	-	-	-
Podolsk city	563	20	29	128	150	120	52	18	20	26
Protvino city	23	5	2	2	5	3	2	3	1	-
Puschino city	29	1	3	8	7	7	-	3	-	-
Reutov city	40	3	9	9	10	6	-	3	-	-
Roshal city	24	-	-	9	5	6	1	-	1	2
Serpekhov city	240	11	16	91	35	52	21	7	3	4
Troitsk city	26	2	1	4	7	8	-	-	-	4
Fryazino city	110	8	5	27	32	23	9	2	2	2
Electrostal city	31	1	5	7	9	4	3	-	-	2
Yubileinyi city	59	2	5	15	15	10	4	3	3	2

10. Summary

Ensuring data confidentiality becomes very difficult and expensive task. Taking into account that data collection during All-Russian Population Census 2020 will be held in the Internet, on paper and on tablet computers it will be a wider use of solutions based on new data protection algorithms and developed software solutions.