Ensuring balance between data access and data protection

Zoltán Vereczkei

Deputy head of the Methodology Department

Hungarian Central Statistical Office

E-mail: Zoltan.Vereczkei@ksh.hu

The study, on the one hand, presents the main elements and co-management of the requests for statistical data and the obligation to protect them and, on the other hand, introduces the data access channels of the Hungarian Central Statistical Office that provides balance between data access and data protection.

KEYWORDS:
Data access.
Data protection.
Data access channels.

DOI: 10.20311/stat2016.K20.en064

Dissemination and provision of access to official statistics have a key role in the statistical business processes of institutions responsible for the development, production and dissemination of official statistics. These institutions provide access for users to various types of statistical information, and thereby, data requests are considered as a driving force behind their statistical business processes. Among data requests, the ones that are made to gain access to data (especially microdata) for scientific purposes have a dedicated role.

The obligation to protect the data of data providers is also important in official statistics. It is fulfilled by legal, methodological, dissemination, and IT security safeguards.

As a result of the developments launched in the last decade, data access requests and the data protection obligation are handled in a balanced manner by the HCSO (Hungarian Central Statistical Office). User needs to access microdata for scientific purposes have now been part of the everyday statistical business process of the national statistical institutes, both at national and European level.

1. Data protection in the HCSO

The purpose of official statistics is to produce statistical information of good quality and to provide access to this information for as many users as possible. Dissemination of statistical information on the society, economy and environment is an integral part of the statistical business process, and data are made available to users within the European Statistical System by means of different data access channels. Consequently, the two key characteristics of statistical information are accessibility and good quality.

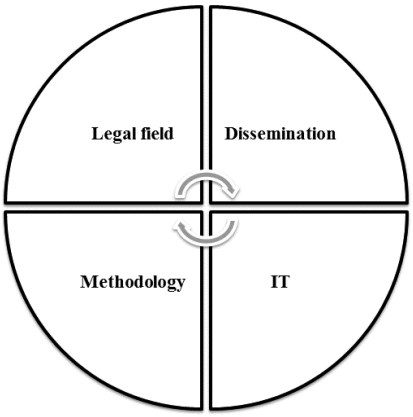
Within the scope of the statistical business process and data access services, data protection is a key element, which means the protection of the individual data of data providers (basically the management of identification and disclosure risk) in official statistics. A fundamental principle of data protection is that – apart from some specific legal provisions (*Nagy E.* [2015]) – the identification of statistical units (individuals) providing information for the purposes of official statistics is strictly forbidden. The event when a statistical unit is unambiguously identified is called *identification*. Another key concept of data protection is *disclosure* when previously unknown in-

formation on statistical units is disclosed. In line with this conceptual background, the aim of official statistics is to prevent disclosure of information on statistical units that would not be otherwise known. In order to fulfil this requirement, various methodological data protection solutions – summarized under the term of *statistical disclosure control* – are used, which require the cooperation of different areas in the statistical institutions. (See Subchapter 1.1.)

1.1. Areas of data protection in the HCSO

Data protection is ensured by the HCSO through the cooperation of four (legal, methodological, IT and dissemination) areas. (See Figure 1.)

Figure 1. Cooperation model of data protection in the HCSO



The legal aspects of data protection cover legal and operational provisions concerning the operation of the office, statistical business processes and statistical products. The obligation to protect statistical information is regulated by national and European laws. These legal safeguards guarantee the protection of information on data providers and on units described by the statistical information.

From this perspective, the most relevant pieces of legislation are the Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics, the European Statistics Code of Practice, and in national context the Act No. XLVI of 1993 on Statistics, the Government Regulation 170/1993 (XII. 3.) as well as the Act CXII of 2011 on the Right of Informational Self-Determination and on Freedom of Information. Since there are several studies on this subject (for example *Nagy E.* [2015], *Lakatos* [2015]), the present paper does not address the legal framework of data protection.

In addition to legislation, *methodological safeguards* also serve as a prominent area of data protection. They cover the development and use of such logical, mathematical, statistical solutions that guarantee that new information on statistical units is not disclosed by the disseminated statistical data.

The *IT area of data protection* embraces the development and use of technical and IT security-related techniques. Typical examples are the protection of communication channels, data access safeguards, management of data access rights and every other solution needed to ensure the expected IT security level of the HCSO.

Data protection in dissemination means that the conditions (such as table structures, the level of detail, etc.) imposed by the statistical outputs should be met in accordance with the data protection provisions, and the necessary data protection safeguards are implemented and applied properly within the operational environment of the IT systems used for dissemination (e.g. limiting the number of variables accessible in the dissemination database). Dissemination is an important area of data protection because the statistical product specifications actively influence the needs for and the usability of the methodological solutions serving the protection of outputs.

In the HCSO, data protection covers every legal, methodological, IT and dissemination solution, method and practice that is used to protect statistical information from unlawful access, that is, to prevent disclosure. Unlawful access is interpreted in its widest scope. Management of information in the HCSO serves only statistical purposes. Collecting individual information is a necessary "instrument" of official statistics but it is not its final purpose (*HCSO* [2014a]). Unlawful access to data or any divergence from their intended use regulated by legal safeguards (e.g. contracts) in data access services is not allowed.

Examining the data protection activity of the HCSO, one can conclude that it aims at enforcing every relevant legal provision by combining legal, methodological, IT and dissemination solutions. Data protection, however, results in the limitation of the use of statistical information (due to legal safeguards or by way of limited information content of disseminated statistical data), leading ultimately to loss in data quality.

To conclude, data protection is moving towards the regulation of the legal and operational circumstances of data access and, as far as the statistical disclosure control is concerned, towards the limitation of the content of information to be disseminated. This approach serves as one of the "pans of the data access-data protection scale".

1.2. Quality aspects in relations between data access and data protection

As previously mentioned, one of the key characteristics of data managed by official statistics is quality. In general, quality means fitness for use and is described by

quality components (*HCSO* [2014e]). These components (relevance, accuracy, timeliness, punctuality, accessibility, clarity, comparability, and coherence) are measured and considered as general criteria for all statistical business processes and products. They are not independent from each other: improvement of one of them may result in a decline of another.

Commitment to quality (measured by means of quality components) is a core value of national statistical services. They usually define these components in their quality policies and consider them as guiding requirements for their statistical business processes and statistical products; the quality policy of the HCSO – which declares the office's commitment to quality and defines the quality components – is definitely no exception to this. Being in accordance with the policy, the quality guidelines for the statistical processes of the HCSO define the quality criteria to be followed in every step of the statistical business processes (*HCSO* [2014c]).

There are different opinions on which quality components are the "most relevant" for users, and they are subject of debates. Some people believe that timeliness is the most important factor in our modern, technologically advanced world (the market values timely information and, in return, is "willing to tolerate" less accurate information) even though traditionally one of the most important values of official statistics is accuracy (that usually has a counter-effect on timeliness). The quarterly or monthly GDP estimates can be mentioned as an example of this dilemma, which are always changed when more detailed yearly information (national accounts) is available; nevertheless, both datasets fulfil important user needs.

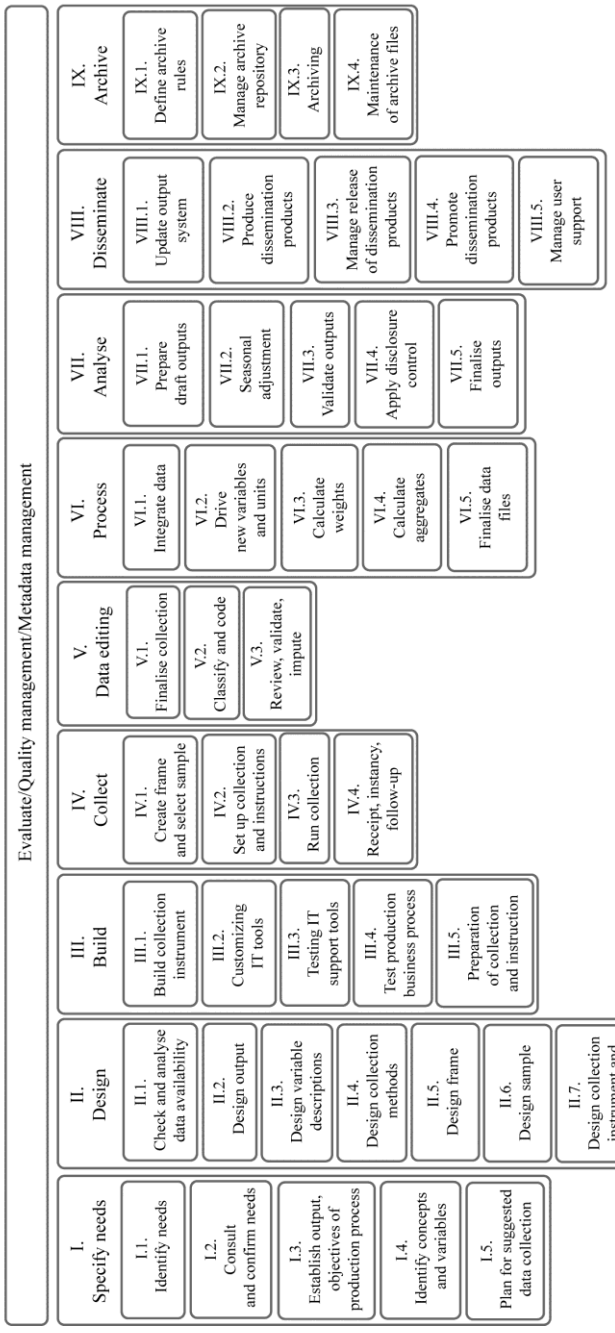
Good quality – or quality, in general – might be perceived differently by developers, producers and disseminators of official statistics. The stakeholders have separate "rankings" of quality components, and they require various "data access solutions" that provide sufficient flexibility corresponding to the different user needs.

Users can "get in touch" with the products of official statistics in different forms, through various channels of dissemination. Dissemination is, therefore, a key step in the GSBPM (Generic Statistical Business Process Model)[1], and this principle is followed by its Hungarian adaptation, the Hungarian Generic Statistical Business Process Modell (called ESTFM), too.

In Figure 2, one can observe that statistical disclosure control also has an important place in the ESTFM (presented as a separate activity under the process phase "Analyse"). However, it does not mean that data protection is present merely at this stage (before dissemination): data management for statistical purposes requires that data protection provisions be enforced throughout the statistical business process. (For the relevant regulation see *HCSO* [2013]).

[1] http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

Figure 2. Hungarian Generic Statistical Business Process Model



Source: HCSO [2014c] p. 11.

Statistical business processes must guarantee that quality criteria are respected and fully met. This purpose is also served by the quality guidelines of the HCSO, some of which require the enforcement of data protection provisions and the consistency between data access and data protection. For example:

> "Data requests must always be satisfied by using data access channels that are the most suitable in terms of data confidentiality and data access considerations. Users can receive information on data access channels and their operational characteristics from publicly available sources." (*HCSO* [2014c] p. 130.)
>
> "If tabular data are disseminated or anonymised microdata are released, efforts should be made to ensure data confidentiality in a fashion that entails the least possible loss of data, i.e. keeping disclosure risk to a bare minimum." (*HCSO* [2014c] p. 131.).

2. Requests for data access

The needs to access official statistical information have intensified (both in quantity and in content) in the last couple of years, and particularly, the number of requests to gain access to data (especially microdata) for scientific purposes has increased significantly. Traditionally, users need indicators, printed materials and tabular data, but the development of IT tools has led to an increased focus on electronic products and on access to databases. Owing to the development of IT and methodological tools, users' ability to process and analyse data has strengthened, and over the last ten years providing access to microdata files (datasets including detailed information of record level) for scientific purposes has become a priority both at national and European level.

The increased access to microdata has brought about a higher risk of confidentiality and the need for new statistical disclosure methods compared to those relevant to tabular data. Access to microdata files (except access to public use files) of the HCSO is limited to scientific purposes. Thus, a researcher accreditation procedure was introduced by the office in 2014 (in accordance with the practice of most European countries) to determine whether data requests serve scientific purposes. The following statements can be made about such requests:

> – Their number is increasing year by year.
>
> – Microdata files are requested for more and more statistical domains.

– The number of microdata sets compiled to answer specific user needs is also increasing. However, they are not included in the list of deidentified datasets available on the HCSO website. [2]

In order to achieve their goals, researchers want to gain access to detailed datasets of good quality so that their scientific results do not be distorted by statistical disclosure control methods. Thus, the "data access side" seeks to maximize the content of the datasets available. This approach serves as the "other pan of the data access-data protection scale".

3. Overview – needs for data access/data protection and changes in tendencies in the last decade of the HCSO

The changes in data requests and data protection needs show different tendencies compared to those of the former decade, while the elements and approaches of some areas have not changed and are still as relevant as they were ten or more years ago.

The principles of data protection were already known and respected in the beginning of the 2000s as well as at any time since the establishment of the HCSO (*Lakatos* [2015]). However, they mostly addressed the protection of tabular data while access to and protection of microdata sets were not regulated at that time.

The HCSO has been criticized by the scientific community several times for it's too strict data protection and data access practice; some people thought that it did not adequately support the needs of scientific research and researchers. One of the reasons behind this opinion was users' unawareness of the data protection considerations and practice of the office, while another reason was that the HCSO's practice did not conform to the increased needs for getting access to data (especially microdata) for scientific purposes.

Centralized methodological coordination of statistical disclosure control has been performed by the HCSO since 2003; however, the release of tabular data has already been regulated since prior to that. After the dissemination database was introduced and the needs to get access to microdata files increased, the requirements to harmonise and coordinate methodological solutions have become more evident. In 2003, the central methodological unit of the HCSO started to harmonise and modernize its statistical disclosure control methods and to standardize the relevant processes of the office. In parallel with these activities, it was also realised that new or changed user needs could be only met if the data access services were extended, especially, for the

[2] The list is available on the HCSO website at http://www.ksh.hu/safe_centre_access

scientific community (e.g. development of the dissemination database and the establishment of the HCSO Safe Centre) (*Szép* [2012]). To achieve this goal, a number of projects were started to standardize and further develop the statistical disclosure methods and practices whose results can be witnessed by users today.

Compared to the practice of the previous decade, one of the most noticeable development-induced changes is that managing microdata files and providing access to microdata-level information are now part of the HCSO's everyday operation. Although the office was certainly aware of the needs of both the Hungarian scientific community and the European Statistical System, the basic interest of researchers to get access to statistical microdata was not generally recognized.

Over the last few years, however, the representatives of both the scientific community and the national statistical institutions have expressed their expectations and defined their problems concerning access to data and confidentiality on different fora (*Harcsa* [2012]; *EC* [2013], [2014]). Their dialogue has contributed to the current practice of the HCSO, where the obligation to protect data and the needs for data access are handled together. Even though the main mission of the HCSO has been to disseminate official statistics on the state and changes of the society, economy and the environment since its establishment in 1867, the co-management of these two aspects became an integral part of the HCSO's practice only in the last ten years. Its most obvious signs are the increasing access to data and the development of the statistical disclosure control methods.

The new and renewed data access channels (especially the establishment and reorganization of the HCSO Safe Centre) as well as the dissemination of a great deal of information on data protection/data access options and practices to the general public are considered as important achievements of the office. All of these changes have greatly contributed to the dialogue between the HCSO and the scientific community (see the HCSO website,[3] *HCSO* [2014b], *Erdei–Horváth* [2004], *Kristóf* [2015], *Nagy B.* [2015], *Faragó* [2013]). Researchers have realized that the various data protection methods might have different effects on data quality and on the relevance of anonymised-dataset-based analyses (*Bartus* [2013]).

The co-management of data access needs and the data protection obligation is a common European practice and not a national "phenomenon". The large European developments of the past few years also show that their separated management cannot be an efficient solution for statistical organizations. The practices to be developed or redesigned should be based on the consensus of all stakeholders (especially that of developers, producers and disseminators of official statistics as well as data archives and researchers) and on international solutions and recommendations (*EC* [2013], [2014]).

[3] http://www.ksh.hu/data_requests_home

4. Balance between data access and data protection

In the light of the above, one might ask the question: "How can the balance between the needs for data access and the obligation to protect data that are somewhat contradictory be ensured?" One of the pillars of official statistics is gaining and keeping data providers' trust by making every effort to protect individual information. Consequently, compliance merely with data protection requirements make data access limited.

Considering the needs of users to get access to statistical data of good quality, the content and details of statistics to be disseminated should be maximized.

The goals are, therefore, to find balance between these two equally important objectives and to achieve a balanced state by official statistics. Thus, the main purpose of data access/data protection activities is to find an optimal solution. If both the needs for data access and the obligation to protect statistical data are recognized as equally important factors, one may not focus only on one of them. The official statistics, therefore, cannot be efficient if the needs for access to statistical data is not examined in detail and the practices used in statistical business processes are not changed accordingly. It is the "suicide of official statistics" if it only sets the objective of meeting data access needs and does not care about the justifiable data protection requirements.

In practice, the balance can be ensured by various data access channels, but the continuous redesign and refinement of the production methods of official statistics and the relevant quality guidelines are also crucial for success.

4.1. Guarantees of the data access-data protection balance in the strategic documents of the HCSO

HCSO Strategy 2020 is the best example for reconciling the requirements of data access and those of data protection. One of its seven strategic objectives ("Ensuring balance between data protection and data access") explicitly targets this issue, and sets out that "the basic purpose of official statistics is to provide the widest possible access to data. Besides, our statutory obligation (and our practical interest) is the protection of the individual data of data providers. We constantly strive for the development and application of new technologies and methods which facilitate the secure access to statistical data without damaging the level of data protection." (*HCSO* [2014d]). This objective addresses the use of modern statistical disclosure methods and practices, the development of integrated service-oriented data access solutions, the use of up-to-date IT security solutions, the provi-

sion of wide-scope access to datasets managed for the purposes of official statistics and the development of joint data access-data protection solutions.

In addition to the Strategy 2020, the quality guidelines and the Data Protection Policy of the HCSO also play an important role in striking the balance. The latter, for example, defines the nine data protection principles of the office, highlighting the importance of access to official statistical information for scientific purposes (*HCSO* [2014a]).

4.2. Data access channels of the HCSO

The IT developments of the last decade as well as the new IT tools and extended knowledge of users have contributed to the increase in data requests (both in quantity and complexity). Due to their various needs, the data access system of the HCSO has to be quite complex, and several data access channels are needed. These channels created with this purpose in mind are one of the most important guarantees of statistical data access.[4]

Researchers – and the scientific community, in general – are a special group of users requesting access to official statistics. As other members of the European Statistical System, the HCSO also focuses on the use of official statistical data for scientific purposes and on maintaining balance between data access and data protection (*HCSO* [2014d]). However, it does not mean that users outside the scientific community cannot access statistical information or their needs are less important to official statistics. Nevertheless, the available alternatives to get access to official statistics are differing depending on whether the user does or does not belong to the scientific community. Figure 3 describes the six data access channels of the HCSO, summarizing their differences.

The data access channels are classified into two groups: access *1*. in the so-called Safe Environment and *2*. in any other environment.

4.2.1. Access to official statistics – in all environments

The HCSO fulfils most data requests by providing tabular data. This data access channel does not require the use of a safe environment. The common feature of the release of tabular data and public use files is that users indeed have the datasets (the conditions of data transmission to a third party can be limited) and can work with them in their own regular working environments (the datasets can be downloaded). Consequently, the release of these data is associated with a higher disclosure risk

[4] http://www.ksh.hu/data_requests_home

compared to that of data access in a safe environment. Since users have more tools and means to use data in this case, ultimately, a stricter data protection procedure is needed.

There are three data access channels that can be operated in all environment: release of tabular data, access to public use files and access to anonymized microdata sets. While the last one is limited to scientific purposes, the other two are open to all users.

The release of tabular data means access to aggregated statistical information by users (typically) on the HCSO website (in form of database files, reports, etc. or in other forms [e.g. paper documents]). It also includes a service when non-regularly-produced tabular data are provided upon user request. These requests can be submitted by anyone, and their scientific purpose is not examined.

The HCSO is required to examine every case of tabular data release, and, if needed, to apply statistical disclosure control methods to protect data. The goal is twofold: to protect the information of tables by preventing data disclosure and to maximize the information content to the greatest extent possible.

In addition to finding a balance between data access and data protection, the harmonization of statistical disclosure control methods is also an important factor. The HCSO has a standard practice for using statistical disclosure control methods, based on methodological recommendations.

The office applies cell suppression for the protection of tabular data. In this technique, information to be protected is removed from the tables and is replaced by a standard sign. With due regard to the data protection provisions and methodological considerations, examination of disclosure issues is carried out for all datasets, and, if needed, primary cell suppression should be applied. In the latter case, the need for secondary cell suppression[5] should be also considered. The HCSO applies the following common rules when using the cell suppression technique:

– All sensitive table cells have to be protected.

– When applying cell suppression, the information loss must be reduced to a minimum, which can be implemented by the following:

– suppression of the fewest number of cells possible;

– if there are several alternatives to suppress the minimum number of cells in a table, that alternative should be selected where data of the fewest data providers are suppressed;

– avoidance of the suppression of sums since they usually present key information for users.

[5] "Statistical disclosure control method applied to tabular data when additional cells apart from the ones treated by primary cell suppression are suppressed in order to ensure the protection of the concerned tabular data." (*HCSO* [2013])

For more information on the protection of tabular data see "Guideline for researchers" (*HCSO* [2014b]) on the HCSO website.

Access to anonymised microdata sets. Anonymised microdata sets are made accessible to researchers for scientific purposes on DVDs and on other media, or through electronic channels. For the protection of anonymised microdata sets, there are legal provisions in place, some of which regulate the management of disclosure risk (e.g. prohibition of integrating datasets with those that were not included in the data request forms of users, prohibition of transmitting datasets to a third party, destruction of datasets after the scientific purpose is fulfilled, etc.). These datasets also have to be fully protected for disclosure avoidance purposes.

Applying statistical disclosure control methods on datasets always results in quality loss (changes in the detail of datasets) and consequently, in bias of conclusions (e.g. estimates for scientific purposes) drawn from these datasets. Therefore, the selection of the appropriate disclosure control methods is of key importance. There are two main goals of protecting microdata sets:

– To minimize the disclosure risk;
– To achieve the "best possible quality" of the research objective (when using anonymized datasets).

Consequently, this data access channel requires coordination between the demands of researchers and data protection requirements.

Access to public use files. This is a relatively new data access channel in the practice of the HCSO that is available for users to access microdata-level information. There is an extensive literature on public use files, and there are also several approaches to manage such files in the data access systems. For the theoretical background that is not addressed by the present paper, see *Kristóf* [2015].

It is important to highlight, however, that one cannot draw conclusions of "good quality" on a statistical population based on public use files. They are mainly used to prepare research in a safe environment and to fulfil education-related needs.

4.2.2. Access to official statistics – safe environments

In the other group of data access channels, users may access detailed microdata files in a controlled environment. Safe environments provide more guarantees to control access to data, thus, it is a preferred solution in official statistics for accessing datasets of higher disclosure risk.

Safe Centre access, remote access and remote execution belong to this group of data access channels. The first two present actually the same conditions for users, while a different kind of controlled environment is faced with during remote execution.

Safe Centre and remote access. The Safe Centre is such an IT-controlled, safe environment that guarantees that no disclosive information can get out of it. Under the current rules, microdata files themselves cannot be transferred out from the Safe Centre of the HCSO but research outputs can be transmitted after an output checking procedure (*HCSO* [2014b]).

This type of access enables researchers to access microdata files having such a high level of detail that is not guaranteed otherwise. The risk of having them accessing detailed microdata is controlled by an obligatory researcher accreditation procedure and legal disclosure control tools. When using Safe Centre or remote access, researchers can conduct their research on the most detailed microdata sets. As a result, their research outputs may be of the best quality. Due to the nature of safe environments, no statistical disclosure control methods (no anonymization) are needed; although the direct identifiers of statistical units are removed, usually no other methods are used (unless it is justified e.g. when some statistical units [for example companies in a monopolistic position] can be easily recognized even if their direct identifiers are removed).

Safe Centre access and remote access are considered as two separate data access channels, even though they operate basically the same way. The difference is only of organizational nature. Based on their preferences, researchers can also get access to microdata sets for scientific purposes at an access point that is geographically closer to them, thereby eliminating the need to travel. The remote access service is, therefore, an "extension" of the Safe Centre, and its main purpose is to ease the burden on researchers by saving time and travel costs. The Safe Centre of the HCSO is in Budapest, however, the office does not intend to limit its service to those researchers who work/live close to the capital: there is another remote access point in Szeged, too, that has been operating since 2014.

Remote execution. Even though remote execution is considered as access to microdata in a safe environment, it is completely different than the former ones. The main difference is that researchers cannot actually see the datasets when using remote execution service and do not have to visit an access point either. Instead, they have to provide the specifications and/or syntax files of their research they are willing to conduct to the HCSO, whose staff prepares the outputs for them.[6]

Remote execution guarantees that researchers can conduct their research on datasets of the same quality as those of the Safe Centre or remote access. In their case, the research outputs have to go through a mandatory output checking procedure before making them available for further use.

[6] http://www.ksh.hu/remote_execution

Figure 3. Data access channels of the HCSO

Release of tabular data – Data are available to the general public on the HCSO website and in yearbooks and reports – Data are also compiled upon user requests – Acceptance of the terms of use is needed	**Access to anonymised microdata sets** – "Standard" datasets are available – Data are also compiled upon user requests – Signature of a contract and a confidentiality commitment is needed	**Access to public use files** – It is available since March 2014 (first files are available on the Population and Housing Census 2011) – Acceptance of the terms of use is needed
Safe Centre access – "Standard" datasets are available (free of charge) – Data are also compiled upon user requests (payment is needed) – Signature of a contract and a confidentiality commitment is needed	**Remote access** – It is provided in a safe environment – available since March 2014 (from the access point in Szeged) – Signature of a contract and a confidentiality commitment is needed	**Remote execution** – Researchers cannot „see" the datasets – The HCSO produces the requested research outputs based on the specification provided by researchers – Signature of a contract and a confidentiality commitment is needed

4.2.3. Comparison of the HCSO data access channels in terms of data access and data protection

The table compares the six data access channels of the HCSO regarding data access and data protection.

In the table, the *strength of primary statistical disclosure control* refers to the impact of the methods used on the dataset. "Strong" means that the datasets are thoroughly checked from data protection point of view, and statistical disclosure control methods are applied when it is necessary. Since the data access channels concerned are available to all users, various scenarios for statistical disclosure control cannot be considered; the quality of the output dataset (usually the level of detail, and thus the "depth" of research) is affected by strict statistical disclosure control methods. When the strength of the primary statistical disclosure control is medium, full-scale statistical disclosure control is applied but, depending on the intended use of data agreed to by the HCSO, various disclosure control scenarios can be drawn, so the possible bias effects of the statistical disclosure control methods on the dataset can be managed (key variables are protected only when it is absolutely necessary [e.g. when the protection of the dataset otherwise cannot be ensured]). When the strength of the primary statistical disclosure control is "weak", there is no primary statistical disclosure control (anonymization), but the research outputs are thoroughly checked in an output checking procedure prior to their release to researchers.

Comparison of the data access channels of the HCSO

Aspect	Release of tabular data	Access to anonymized microdata sets	Access to public use files	Safe Centre access	Remote access	Remote execution
Strength of primary statistical disclosure control	Strong	Medium	Strong	Weak	Weak	Weak
Data access conditions	Open to everyone	For scientific purposes only	Open to everyone	For scientific purposes only	For scientific purposes only	For scientific purposes only
Pros	Flexible	Researchers can "take away" the dataset; they can work with it in their usual working environment	Freely downloadable with the acceptance of terms of use	Possibility to work with the most detailed microdata sets	Possibility to work with the most detailed microdata sets	Possibility to work with the most detailed microdata sets
Cons	Possible bias due to anonymization	Possible bias due to anonymization; legal obligations (use for the agreed purposes only; need to destroy the dataset after scientific purpose is fulfilled)	Due to the strong statistical disclosure control, the datasets are useful only for their dedicated purposes	Need to visit the HCSO Safe Centre in Budapest	Need to visit a remote access point (currently: Szeged)	Researchers do not see the dataset; it is usually a long procedure to discuss and update the specifications provided by researchers
Quality of research conducted using the dataset	Medium	Medium	Low	High	High	High

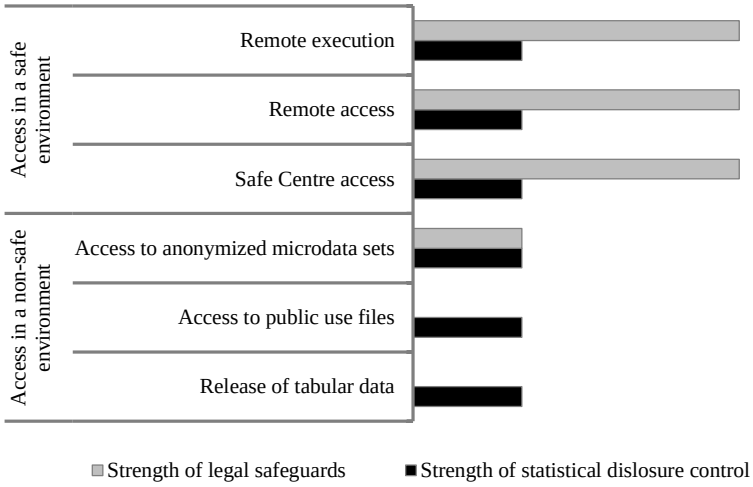
In the table, the "*Quality of the research conducted using the dataset*" aspect refers to the quality (usually the level of detail) of the accessible datasets. The most detailed data are available in safe environments for scientific purposes. The use of such datasets, however, requires extensive preparatory work from researchers (explanation of their research, researcher accreditation, detailed documentation, etc.).

Overall, there are several data access channels available to users (especially researchers) who intend to use microdata sets for scientific purposes. If they want to use these data in their regular working environment and are willing to accept the fact that the datasets are anonymized (that results in loss of quality), they can get access to anonymized microdata sets. But if researchers want to conduct their research on the most detailed microdata sets possible, Safe Centre access, remote access or remote execution services are available to them.

4.2.4. The principle of balanced risks

Over the last ten years, the HCSO has had an explicit focus on the so-called "balanced risks" principle to keep the risk of disclosure constantly to a minimum in all data access activities. This practice has also been instrumental in the development of the six data access channels of the HCSO. Figure 4 shows the implementation of the principle in the channels.

Figure 4. The principle of balanced risks and the data access channels of the HCSO



The balanced risks principle requires that the access to the HCSO datasets through the six channels should be harmonized in relation to data protection and users' requirements. In practice, the balance is ensured by statistical disclosure control methods and legal tools. As a general rule, the stronger the legal safeguards are, the less strict statistical disclosure control is needed, and vice versa.

Full advantage of the options provided by legal safeguards should be always taken to minimize the need for statistical disclosure control as it usually results in the reduction of the scope of disseminated data and, ultimately, in quality loss.

As it is shown in Figure 4, statistical disclosure control is minimum when users access data in a secure environment. In this case, data protection is guaranteed by strong legal (and technical) safeguards, thus, statistical disclosure control is limited only to the removal of direct identifiers and to the mandatory output checking procedure.

In data access channels (release of tabular data, access to public use files) where access is not limited to scientific purposes, the power of legal tools is typically weak (the acceptance of the terms of use is the only instrument which can be enforced).

Therefore, the statistical disclosure control is stronger for these channels in order to guarantee the protection of statistical information.

4.2.5. Key elements of balance: cooperation and consensus

In addition to the strategic goals, methodological guidelines and data access channels already mentioned, the balance between data access and data protection can be further strengthened by the active discussion and consensus of stakeholders. There are several examples of such cooperation in the European Statistical System where the three key stakeholders (national statistical institutes, data archives and the researcher community) are involved in (*EC* [2013], [2014]).

Although the methodological experts of the HCSO always focused on the cooperation with the scientific community in the last ten years, the balance between data access and data protection requires more active and efficient "teamwork".

5. Conclusions

Fulfilling data protection requirements and the data access needs of users is a key objective of all national statistical services and that of the HCSO, too. Striking and stabilizing the balance between data access and data protection is one of the goals of the HCSO Strategy 2020 and also an important demand of official statistical experts and the scientific community, declared during their discussions.

HCSO ensures the balance by its quality guidelines, Data Protection Policy, new regulations on the modernization of its data access-data protection system and the six data access channels.

The data access channels are one of the most important results of the developments taken place at the HCSO over the past few years. Based on the balanced risks principle, these channels require statistical disclosure control and legal tools of different levels. There are several alternatives for users to get access to various kinds of datasets of different quality (level of detail). In line with the strategic aim of the HCSO, researchers may also choose from several data access channels to get access to statistical information, based on their own data access needs.

In order to ensure balance between data access and data protection in the long run, the realization of the strategic goals of the HCSO, set until 2020, are crucial. Active and effective discussion between official statistics and researchers is also of key importance in the further improvement of data access services and data protection tools and also in setting future research goals.

References

BARTUS, T. [2014]: The effect of data swapping procedures on regression estimates – Evidence from a simulation study. *Hungarian Statistical Review*. Special No. 18. pp. 3–20.

EC (EUROPEAN COMMUNITY) [2013]: *Data without boundaries*. Work Package 6: Enlarging Cooperation: Conferences & Training Sessions. Deliverable D6.3. First Regional Workshop. 30 April. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d6-3_regional-workshop-report.pdf

EC [2014]: *Data without boundaries*. Work Package 6: Enlarging Cooperation: Conferences & Training Sessions. Deliverable D6.6. Second Regional Workshop. 31 October. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d6-6_regional-workshop2_report_final.pdf

ERDEI, V. – HORVÁTH, R. [2004]: Az adatfelfedés elleni védelem statisztikai eszközei. *Statisztikai Szemle*. Vol. 82. No. 8. pp. 705–727.

FARAGÓ, M. [2013]: Védett táblázatok morfológiája – Optimális másodlagos cellaelnyomás számítógép nélkül. *Statisztikai Szemle*. Vol. 91. No. 10. pp. 947–970.

HARCSA, I. [2012]: „A mikroadatok hozzáférésével és az adatok felfedés elleni védelmével kapcsolatos kérdésekről" címmel szervezett műhelykonferenciáról. *Statisztikai Szemle*. Vol. 90. No. 11–12. pp. 1162–1164.

HCSO (HUNGARIAN CENTRAL STATISTICAL OFFICE) [2013]: *27/2013. HCSO Regulation on Data Protection*. Budapest. http://www.ksh.hu/docs/szolgaltatasok/adatigenyles/hcso_regulation_on_data_protection.pdf

HCSO [2014a]: *Confidentiality Policy of the Hungarian Central Statistical Office*. Budapest. http://www.ksh.hu/docs/bemutatkozas/eng/avpol_web_eng.pdf

HCSO [2014b]: *Guideline for Researchers – Instructions about conducting research using the Safe Centre of the Hungarian Central Statistical Office.* Budapest. http://www.ksh.hu/docs/szolgaltatasok/adatigenyles/guideline_for_researchers.pdf

HCSO [2014c]: *Quality Guidelines for the Statistical Processes of the Hungarian Central Statistical Office*. Version 3.1. Budapest. http://www.ksh.hu/docs/bemutatkozas/eng/minosegi_iranyelvek_eng.pdf

HCSO [2014d]: *Hungarian Central Statistical Office Strategy 2020*. Budapest. http://www.ksh.hu/docs/bemutatkozas/eng/strategy2020.pdf

HCSO [2014e]: *Quality Policy of the HCSO, 2014*. Budapest. http://www.ksh.hu/docs/bemutatkozas/eng/minpol_web_eng.pdf

KRISTÓF, P. [2015]: Nyilvános mikroadatfájlok összeállításának főbb jellemzői, különös tekintettel az adatvédelmi szempontokra. *Statisztikai Szemle*. Vol. 93. Nos. 11–12. pp. 1112–1139.

LAKATOS, M. [2015]: Az adatvédelem és a statisztika kapcsolatának jogi szabályozása. *Statisztikai Szemle*. Vol. 93. Nos. 11–12. pp. 1017–1050.

NAGY, B. [2015]: A célzott adatcsere módszere a térstatisztikában. *Statisztikai Szemle*. Vol. 93. Nos. 11–12. pp. 1152–1169.

NAGY, E. [2015]: A statisztikai adatvédelem és -hozzáférés szabályai az Európai Unióban. *Statisztikai Szemle.* Vol. 93. Nos. 11–12. pp. 1051–1069.

SZÉP, K. [2012]: 2011 októberében volt 10 éve, hogy a KSH-ban létrejött egy központi módszertani egység. *Statisztikai Szemle*. Vol. 90. No. 4. pp. 319–333.