United Nations

ECE/CES/2019/40

Economic and Social Council

Distr.: General
13 May 2019

English only

## Economic Commission for Europe

Conference of European Statisticians

**67th plenary session**
Paris, 26-28 June 2019
Item 2 (a) of the provisional agenda
**New data sources – accessibility and use**
**Session 1: Accessing new data sources**

# Application of alternative data sources in official statistics – the case of Hungary

## Note by the Hungarian Central Statistical Office

*Summary*

The challenges faced by official statistics in the twenty-first century are manifold. We are surrounded by systems that are becoming substantially and increasingly complex. Moreover, official statistics need to be able to capture new phenomena and complex realities (e.g. globalisation, global demographic trends or sustainable development) in a meaningful and timely fashion. In parallel with the emergence of new phenomena, new types of data also appear, offering an opportunity to further increase the relevance of statistics. To utilize these new datasets, we need to create data inventories and develop a data strategy that will facilitate finding new, potentially usable tools and processes to harvest the statistical potential, work out new methods, and improve statistical skills and capabilities.

The goal of this document is to highlight some of the major developments using non-conventional data sources for official statistics at the Hungarian Central Statistical Office, namely, an online cash register – for retail turnover, traffic monitor camera data – for tourism statistics, database of National Tax Authority – for labour marker statistics, web scraping – for consumer prices, etc. By efficiently using these alternative datasets, it is possible to reduce the administrative burden and increase the statistical potential, while supplying tailor-made information to decision-makers and improving the quality of official statistics.

This document is presented to the 2019 Conference of European Statisticians seminar on “New data sources – accessibility and use”, session 1 “Accessing new data sources” for discussion.

# I. Introduction

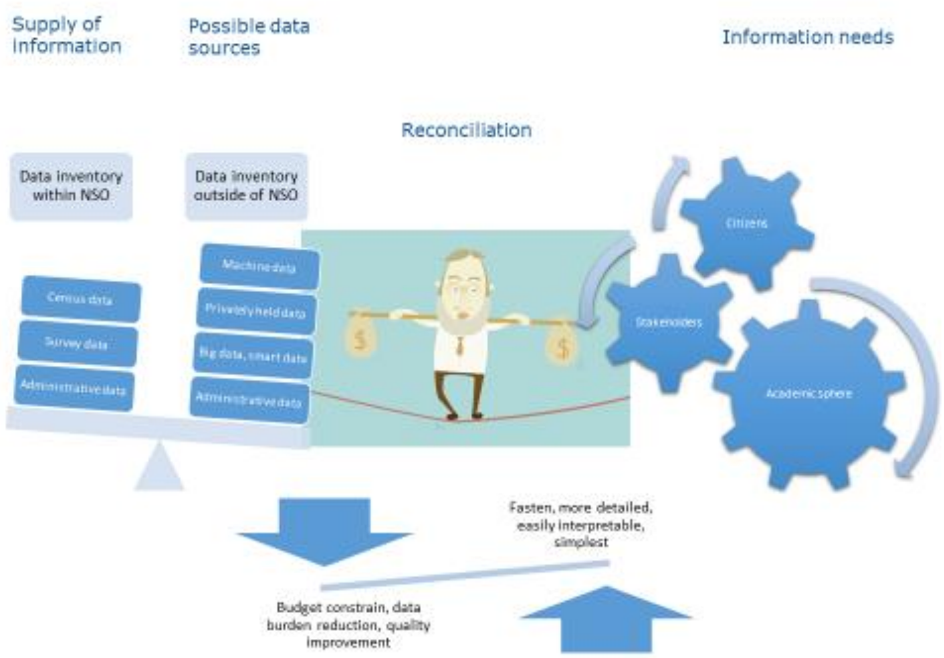
1. As a consequence of the data revolution, there is nowadays an abundance of data sources: questionnaire-based statistical data, census data, big data, smart data, machine data, administrative data, privately held data, and so on. On the one hand, there are many cases in which statistical domains are based on traditional data sources, which are already reaching their limits with respect to timeliness, relevance and compliance with the requirement to reduce the burden on respondents, which is causing decreasing response rates. On the other hand, non-conventional data are not generated for statistical purposes. The quality of new data sources is sometimes lower than of the traditional ones or the measurement of their quality also requires non-traditional solutions, and their usage causes more external vulnerability. Therefore, statistical authorities have greater opportunity than ever to utilize these non-conventional data sources, but national statistical institutes (NSIs) need to reconcile them with their high quality standards and to integrate them into their statistical production system.

2. Until now, statistical authorities have had a major competitive advantage in having almost exclusive access to input data from both statistical surveys and administrative records. This situation has drastically changed. In the age of globalization, data have growing market value. New data sources are accessible to almost everyone, offering new opportunities, and at the same time exposing statistical authorities to competition.

3. Furthermore, societies and economies are going through significant changes which require sophisticated statistical services. The traditional user needs have altered as well. New policy related demands, new scientific priorities emerge increasingly – fostering new needs for statistics (digitalization, globalization, Sustainable Development Goals, etc.).

4. The attitude of our current main user groups (decision makers in the public and private sector, academia, media and citizens) to data quality also seems to have changed. More and more users seem to value timeliness over accuracy. They expect statistics to be made available faster, as they have plenty of non-official alternative sources at their fingertips.

Figure 1
**Official statistics at the cross-roads**



5. All in all, we are at a crossroads. Hence, we need better data management, strategies, more process-based approaches, skills and tactics that will control costs, improve processes and timeliness in capturing the right, fit for purpose information.

6. As part of this strategy we need:

- A process-oriented approach, in which it is necessary to enhance the integration of more or less isolated existing data sources and statistical registers, the risk management of the statistical production process and further decrease the time needed to hit the markets, without making serious compromises on other quality dimensions, namely accuracy.

- Establish/renew a timely and historical data inventory – the largest wealth of NSIs – to efficiently handle growing data needs. The principles of the inventory, the range of data, the rules of updating and archiving need to be defined. This work also underlines the necessity of well-established methodological guidelines, metadata standards and detailed documentation. We need to recognize that the collected information is not only for the present, but for the future as well.

- More effective mapping and integration of administrative data sources into the statistical system. It is necessary to provide for the possibility that the use of administrative data may not only be carried out in an isolated statistical domain but should also be wider. The main goal is to make efficient use of potential data sources. The Fundamental Principles of Official Statistics should be taken into account when performing this task.

7. As part of its data strategy, the Hungarian Central Statistical Office (HCSO) is striving to sustain the quality of classical data collections, as well as to integrate administrative and alternative data sources into the official statistical processes, thus benefiting from the opportunities of the changing external environment. Where quality requirements enable to do so, and it is methodologically feasible, HCSO also aims to replace classic questionnaires with new data sources. The Hungarian Central Statistical Office is committed to facilitate a shift towards more diversified data portfolios.

8. We have given a brief outline of the strategic frame above, following which we deploy the alternative data sources into our statistical system. As part of our data strategy, we renew, among others, the retail and the border traffic data collection systems, using online cash register’s data and road traffic data.

## II. Specific developments to access a more diversified data portfolio

9. One of the main strategic aims of the Hungarian Central Statistical Office is to decrease the administrative burden on data providers by simplifying questionnaires and using already available, administrative data of other institutions. As a first step, we made an inventory on all currently potential available data sources.

10. By gathering the datasets that had already been taken over from administrative sources, we can investigate if there is any available data which could be used in lieu of a primary data collection. We analysed the intense of use of the administrative data, whether the data can:

(a) Be used as a replacement of an existing data collection;

(b) Replace some part of a statistical data collection;

(c) Be used for validation; or

(d) Be published without a building a heavy statistical process (e.g. publish information right after checking the data quality).

11. The inventory enables us to analyse the ways of re-using the existent data to produce or supplement new statistics. From the 370 data collection exercises at this stage, we see changes in 21 cases for 2018. Regarding 2019, we noticed 20 amendments, which include 14 new administrative data takeovers. The majority of the new data takeovers is mainly related to the 2021 Population and Housing Census. However, some of them may replace primary data collections or are issued to meet the requirements of regulations in the European Union.

12. The introduction of new data sources into the statistical system can only be implemented gradually as a function of fulfilment of the quality requirements. In this case, experimental statistics play a crucial role in disseminating our primary results, and then, at a later date when the product is mature enough, it will become part of official statistics. We are committed to this step-by-step approach and to fostering the experimental statistics in HCSO.

## A. Monthly earnings statistics from new administrative data source

13. From 2019 onward, the monthly wage labour statistics data collection can be completely replaced by the available administrative data sources. The information on earnings (by Hungarian concept) are produced based on declarations on contributions received from the National Tax and Customs Administration. In regard to budgetary institutions, data received from the Hungarian State Treasury are produced and published, in parallel with cancelling the previous monthly data collection on labour. Cancelling the monthly data collection reduced the burden of data supply significantly, to less than half (regarding to the infra-annual data collection on labour data). The administrative data contains significantly more detailed data – on personal/job level – regarding every filled position and earnings. The calculation of such basic information, which was not previously accessible in the HCSO data collection, is made possible in this way. The value of median earnings, net earnings based on tax and contributions allowances belong here. The database increases with the availability of background information – unavailable in the previous data collections – such as employees' sex, age and occupational characteristics. The new information will be introduced gradually. Information not included in the declarations - such as wages and salaries by SNA concept, hours worked, detailed regular earnings data - is available quarterly since 2019 from the labour statistics data collections for enterprises employing at least 5 persons, budgetary institutions and non-profit organizations which are significant in respect of employment.

## B. Administrative data for labour cost index calculation

14. This project aims to enhance data quality according to the potential use of the administrative data. HCSO is planning to renew the estimation method of the quarterly labour cost index (within this the social cost and taxes) using administrative data. The tax declaration on contributions contains information also on taxes paid by the employer (e.g. social contribution tax, other taxes paid after benefits granted to employees).

## C. Use of administrative data via the ASP2 government project

15. The Application Service Provider (ASP) model has been largely used worldwide in the business sector and public administration as a cost-effective solution. The advantage of the model is that the user can access the software as a service provided by a remote service provider online via a simple web browser. This model can also be technologically and economically advantageous for municipalities to support the wide range of their functions. The scope of Hungary's Municipality ASP system is to simplify the administrative procedures of local governments, and to make the local government subsystems of the state budget transparent. The further goal of the Municipality ASP is to provide modern, integrated and cost-effective state of the art IT solutions for local governments, fostering standardised internal operation and a common platform-based provision of local e-Government services to citizens and businesses. In 2015, Hungary launched a pilot project entitled "Establishing a Municipality ASP center", with 55 local volunteering municipalities, and further 39 municipalities joining from January 2016. Following the successful pilot, the Government decided to launch "Municipality ASP 2.0" in 2016 to extend the system to national level on a mandatory basis (according to Government Decree No. 257/2016. (VIII. 31.) on the Municipality ASP). As of January 2019, over 90% of the Hungarian local governments are using the central ASP system. The leader of the project is the Governmental Information-Technology Development Agency in association with the Ministry of Interior and the Hungarian State Treasury. Many other governmental bodies are involved in different aspects

of the systems development, including the Hungarian Central Statistical Office. The objective of the continuous professional consultations between HCSO and the members of the ASP project is to replace certain statistical questionnaires submitted by local governments to HCSO. The automatic data transmission from ASP's integrated systems reduces their administrative burden by eliminating their obligations to fulfil specific statistical surveys. As of February 2019, HCSO plans automatic data transfer in case of two important statistical publications from ASP:

- A report on businesses with distributive trade activity
- An investment survey of municipals.

## D. Development of the Household Budget and Living Conditions Survey

16. HCSO has projects aimed at the reduction of response burden and improvement of income data quality. This includes the methodological development of the Household Budget and Living Conditions Survey (HBS+SILC) in terms of the use of innovative tools during the data collection recording of purchased items, use of scanners as online character recognition system for recording the purchase bills. Regarding the data collection of income data – in the framework of a grant project – the use of employee income for tax register data is under testing. Donor imputation from several external data sources was carried out due to high item non-response and generally underestimated employee income in the Labour Force Survey (LFS) in the framework of a grant project. According to the final report of the project, the most reliable data sources were proved to be the registers of return contributions by the Tax authority. After the Integrated European Social Statistics come in force, employee income of LFS will be constructed by the method developed in this project.

## E. Improvement in the use of administrative data sources in Census 2021

17. The 2021 census will represent a significant improvement in the use of administrative data. The Hungarian Census Act CI/2018 prescribes the name of the respondent as part of the obligatory data to be collected in the 2021 census, which is an important change from the anonymous censuses of 2011 and 2001. The name, together with data about address, sex and date of birth, makes it possible to connect census questionnaire data to administrative datasets directly. Based on the results of research supported by Eurostat grants, the 13 administrative data sources necessary for the census were defined and listed in the Census Act. The administrative data sets will serve as the basis of a statistical population register to be used in the census for various tasks. It will provide the census frame and will make it possible to check coverage during the data collection phase. It will also serve as the dataset to be used for validation, imputation and editing during data processing. After the census, the administrative population register supplemented with census questionnaire data will be the main data source to provide census type data on an annual basis, and it will take an important part in the development of population and other social statistics. It will also make possible to build person-based samples using an in-house register.

## F. The use of online Value Added Tax (e-invoice)

18. The reporting system for Value Added Tax (VAT) has been completely altered from 1 July 2018. The purpose of the introduction of the online reporting system and the development of the data handling system is to further the economy by confining tax fraud activity. The changes apply only to the invoices with the following attributes:

- The amount of output tax on the invoice is no less than 100,000 HUF
- The invoice has been issued towards an economic entity that has a domestic tax identification number and is not subject to EVA (Simplified Taxation of Entrepreneurs).

19. The summary was an important data source for HCSO, mainly for the foreign trade statistics, because this was the base to identify commercial partners. The Large Cases Unit

(LCU) section is also planning to use them whereas these data give insight into the commercial activities among the enterprises and these can be used for the analysis of acquired figures. Therefore, it is necessary for us to continue receiving information on the invoices transmitted to the online system. An experimental data reception is in progress aimed at specification of that variables and those thresholds obtaining manageable amount of reliable data by HCSO.

## G. Estimation of online cash register data in retail trade statistics

20. The estimation of retail and catering data in HCSO is based on a new methodology, in which data collection is supplemented with data from administrative sources. Based on the new methodology, we can generate these data based on more comprehensive population data compared to previous practices, resulting in reduced sampling error. On this basis, as a result of the new methodology using online cash register data, data provide a more reliable picture of retail and catering sales than the previous practice, which was rather based on sampling. Using these data, HCSO has increased the coverage of data providers while reducing respondent burden by approximately 80 per cent.

## H. Use of roadside camera data for the estimation of tourism basic population

21. The main goal is to improve the data quality of tourism statistics. Beyond the existing methods (manual border traffic counting), in addition to the National Toll Payment Services Plc. roadside traffic data, we also examine license plate data and the Hungarian National Police Headquarters vehicle traffic data. Using these new data sources, we produce estimations for certain typical subcategories of the population (e.g. transit, daily commuters), which can be used beyond the main objective (a more accurate estimation of the nationality distribution of the travellers) to validate the results of the data collection on “Tourism and other expenditures of foreigners in Hungary” and “Hungarian travellers abroad” recordings. The quality of tourism statistics is improved by the fact that the results can be used in statistical surveys of Hungarian tourism, which serves as a basis for the definition of tourism expenditures and is used for the estimation of the exports and imports of tourism for the calculation of the gross domestic product.

## I. Use of new data sources for commercial and other business accommodations

22. The establishment of the National Tourism Data Provision Centre changes the statistical data provision of commercial and other business accommodation establishments in a fundamental way, since their data are transferred electronically to the Centre on a daily basis instead of HCSO. Currently, the electronic data receiving application (KARÁT) of HCSO is being planned and tested; the first data for commercial accommodations are expected in August 2019. Data for other business accommodation establishments will be more detailed with the integration of this type of accommodations into the Centre compared to present data collection of HCSO, which is in line with the growing need for this kind of information.

## J. The National eHealth Infrastructure

23. The National eHealth Infrastructure (EESZT) has been developed in order to utilize the opportunities of eHealth by linking health service providers and providing a single communicational space for them. EESZT makes health data and documents more easily available, facilitates a more efficient data management and analysis by members of the healthcare sector. In collaboration with the experts of National Healthcare Services Center, HCSO is looking for future opportunities for statistical use of the data stored in the system.

## K. Development of cooperation agreement with business partners to take over and use real estate advertising information

24. HCSO plans to elaborate the model of cooperation with business partners to use their databases. We aim to achieve mutually beneficial cooperation where HCSO contributes with its professional knowledge and provides some kind of validation of business databases. A possible partner is a real estate website which has developed a series of data processing techniques to facilitate the statistical utilisation of data. To have a better knowledge of how real estate websites cover the whole rental housing market, HCSO completed a representative survey on rented homes in the end of 2018. In this way, matching information of the survey with business data may have multiple uses for both partners. Although the Law on Statistics in Hungary allows HCSO to use business databases, HCSO wishes to arrive at an agreement that is advantageous for both partners. For a real estate business partner, the greatest advantage might be the validation of their database and a precise estimation of its coverage, as in housing market the information has an extremely important role.

## L. Cooperation with the governmental building administration in order to harmonize the advantages of electronic systems developed on both sides

25. In the last few years, all national organisations and public administration units have implemented electronic systems and shifted towards electronic administration solutions. In parallel with HCSO's electronic data collection system (called ELEKTRA), the governmental building administration launched several electronic projects to modernise and facilitate the work of building authorities and also to make builders official duties cheaper, faster and more effective. The housing construction statistics are collected from local building administration units via ELEKTRA. Connecting ELEKTRA with electronic building administration systems would relieve the workload of administrative staff in local building authorities.

## M. Participation in the work of the project led by the National Healthcare Services Center for the digitalisation of the death certification

26. One of the most important goals of this project is the digitalisation of death certification (DC), which is currently working based on paper. The special feature of death certifications is that several institutions (health care providers, state and municipal bodies) participate in the completion of the certificates. Moreover, the filling out of the questionnaire is for different purposes and is carried out in several paper copies. The users of each of the copies are the following: the Hungarian Central Statistical Office, the registrar, the notary (in addressing inheritance matters), the relative of the deceased (or the organizer of the funeral), the doctor who issued the certificate, and the Public Health Department of Government Office. In the project, it is possible to establish connection between the interfaces and integrate the data of the health care systems and other systems supporting the administrative processes. The introduction of digitalisation will further improve the timeliness and the accuracy of death statistics, reduce the administrative burden of the data providers and diminish the time spent on data entry significantly. The new e-DC system is expected to be introduced in 2020.

## N. Further development of data transmission from the electronic registration system and the introduction of electronic data collection of vital events involving the institutions concerned

27. HCSO is entitled to collect population statistics on the basis of Section 30 of the Act CLV of 2016 on Official Statistics. According to the law, the registrars are the one of the most important data providers of vital events. The electronic registration system was introduced in Hungary on 1 July 2014; afterwards there were significant changes in the data collection of the statistics of vital events as well (improvement of the data content of the data

transmission from the electronic registration system and the development of the IT system required for regular data transmission). Currently, during the registration process, HCSO takes over all of the vital events data available from the electronic registration system. However, further improvements are needed to the transmission of the complete electronic data collection of vital events statistics, as some of the demographic information required for vital events statistics is not included in the registration system, e.g. education, economic activity, occupation, health data of a new-born child and a mother. For the examination of opportunities for further development besides the registrars, the involvement of other institutions – primarily health care institutions – will be necessary in the data service. The introduction of full digitalisation will further improve the quality, the timeliness and the accuracy of vital events statistics, the administrative burden of the data providers and the time spent on data entry will significantly diminish.

## O. Development of data transmission of births abroad

28. HCSO is constantly looking for the ways of using more administrative data sources for official statistics. For this, we have to keep track of the various national legislative changes and find new data sources as well. According to the modified legislation as of 1 January 2018 regarding the family benefits, two of the Hungarian family benefit forms (maternity allowance, support of young people’s life start), also became available to Hungarians living outside of Hungary. Applications for maternity allowance after Hungarian citizen children are born abroad are assessed at the Hungarian State Treasury. The primary purpose of our development is to take over the available data about children born abroad with Hungarian citizenship from the maternity allowance system of the Hungarian State Treasury. This would give us access to data that was so far not available. Data will be received on an annual basis at the end of May of the following year of the submission of applications. The most important data to be received are the following: the child’s place and time of birth, sex, citizenship, the mother’s residence, place and time of birth, citizenship. After the processing of the data, a report will be published as well, in order to satisfy the growing demand for data on live births occurred abroad in the recent period.

## P. ESS.VIP.BUS.ADMIN project to improve the measurement of international migration using administrative data sources

29. As seen above, HCSO utilizes many administrative data sources. The general aim of this action is to make further steps in the use of administrative data sources for the production of official migration statistics. One of the main focus is to expand the use of National Health Insurance Fund (NHIF) data in measuring international migration. NHIF is already a main data supplier in the production of migration statistics, currently for measuring emigration flows of Hungarian citizens. The NHIF database is not used to its full potential yet. We plan to use the database of non-Hungarian citizens and make use of the longitudinal measurements in order to fulfil the increasing need for a more dynamic measuring of migratory processes. The following actions are to be conducted in the project:

- Building a database suitable for longitudinal analysis based on data from NHIF
- Identifying a cohort of people involved in international migration using NHIF data
- Examining the extent of circular migration in the cohort using the definitions and recent measurement recommendations of UNECE and Eurostat.

## Q. Microdata exchange in the field of migration statistics between Statistik Austria and Hungarian Central Statistical Office

30. According to the available statistics, Austria is the second largest country of destination of Hungarian citizens emigrating from Hungary. The measurement of emigration and especially the measurement of migration of persons with the right of free movement means one of the biggest challenges for the statistical offices of the ESS. It is in line with the

interest of persons who emigrate to register in the administrative registers of the destination countries because this registration is connected with some rights or benefits, but these migrants rarely deregister when leaving their country of origin. To get a more detailed and clearer picture of the migration between Austria and Hungary, HCSO is planning to conduct a migration-related data exchange in two phases. In the first phase, HCSO is changing aggregated statistical data with Statistik Austria on NUTS 2 level and in the second step we are planning to carry out a microdata exchange. The exchange of the aggregated data is going to contain data of all type of citizens who takes part in the migration process, but the microdata exchange is going to contain only data of Hungarian and Austrian citizens. The main objective of this data exchange is the improvement of the migration statistics based on administrative data sources. The microdata exchange is strongly supported by the ESS for the improvement of official statistics. According to the experiences of this project HCSO is going to be able to better validate its international migration statistics.

### R. Web-scraping in consumer price statistics

31. The business process of calculating the consumer price index for Hungary is currently not based on automatically collected data, although some IT tools have been already developed for web-scraping. Initial research on how to use web-scraped price data for the development and production of official statistics has started recently in the Hungarian Central Statistical Office; the question of calculating weights and some other problems with sampling outlets and products are still open issues. Since web-scraped price data has more additional information about the quality of the products than original survey data, it makes new opportunities for quality adjustment and focusing on this aspect of the index calculation. With these data we would test hedonic regressions as a method of quality adjustment. Based on the limited experiences of HCSO with web-scraping solutions, finding/developing effective IT tools to support web-scraping methods is not self-evident and easy. It poses a challenge for the whole project.

### S. Scanner data in consumer price statistics

32. The business process of calculating the consumer price index for Hungary is currently not based on scanner data. From 2019 we introduced a new data transmission regarding scanner data into our National Statistical Survey Programme; therefore, reporting is mandatory for the designated retailers. We succeeded in concluding a cooperation agreement with a retail store chain about their data transmission on sales and volume data. In this current development phase, the greatest challenge is the interconnectivity of collected data from the traditional and new data sources and the implementation of appropriate methodological and IT solutions.

## III. Conclusions

33. This document intends to highlight some of the major developments in HCSO to share the experiences in using non-conventional data sources to reduce the burden of data providers and to improve data quality. For the effective use of new data sources, we need data strategy not only at Member States level, but at international level as well.

34. In the current situation of statistics characterized by a strong dynamism, the need for varied and timely information at multidimensional level increases and official statistics are called to respond effectively by ensuring and transforming its data management systems. For this reason, NSIs need to renew their business model to remain relevant in a fast-changing world where advanced technologies and communication impose challenging structural reforms in the statistical production chain. Experimental statistics are an important stepping stone in introducing new data sources. The topic of prioritization is also crucial in the reform process, because there is an increasing gap between external requirements and resources of statistical institutions. We do not have resources to waste, we have to optimize them. We must ensure that our strategic directions reflect the most important needs. On the other hand, achieving modernization is a task that combines different skills. Key challenges within the

implementation of data strategy include changes related to legislation in line with the data access and (re)use, as well as to continue ensuring high data quality together with integrity and suitability of statistical methods within the innovative methodological framework. Furthermore, investing in communication and dissemination is vital both in increasing statistical culture among users and in fostering international collaboration while engaging stakeholders in decisions and creating partnerships with the scientific community, universities and the private sector in order to enrich global synergies.

35. The driving force of data strategy is the adoption of the rapidly changing external context and the ability of continuing the production of top-quality statistical information. Various challenges emerge from the change in the demand for statistical data, from the increasing wealth of information available and from the availability of new methodological and technological tools. Official statistics need to continue its shift from mere "provision of numbers" towards "creation of new knowledge and more added value". Using the current opportunities of the twenty-first century and the data strategy, we could expand the statistical horizon of our future.

## IV. References

Baldacci, E. (2013). Innovate or perish, Italy's Stat2015 modernisation programme. 59th World Statistics Congress, Hong Kong, 25-30 August.

ESSC (2018): ESS priorities beyond 2020, https://ec.europa.eu/eurostat/documents/7330775/8463599/ESS+priorities+beyond+2020+final.pdf/42ea415b-0b40-418b-8c76-25a8e219365c

Giorgio Alleva (2015): Adding Value to Statistics in the Data Revolution Age Italian National Institute of Statistics – Istat, Rome, Italy, ISI Congress 2015, Rio de Janeiro

HCSO (2017): Másodlagos adatforrások használata a statisztikában, http://www.ksh.hu/docs/hun/xftp/idoszaki/pdf/muhelytanulmanyok11.pdf

McCandless, D. (2010). Data is the new soil. TED talk on visualising data, Columbia Journalism Review.

Søren Schiønning Andersen, Áron Kincses, Cristina Pereira de Sá (2018): What user needs do we want to satisfy? – VIG backround paper