United Nations ECE/CES/2013/31

Economic and Social Council

Distr.: General
27 March 2013

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-first plenary session
Geneva, 10-12 June 2013
Item 4 (b) of the provisional agenda
How should national statistical offices respond - moving from risk avoidance to risk management

Micro-data: a crucial asset for statistical systems

Note by the National Institute of Statistics of Italy

Summary

This paper describes the strategic issues in improving access to micro-data, and how the National Institute of Statistics of Italy addresses the related challenges.

Official statistics aim to satisfy the information requirements of different categories of users, which are continuously increasing and diversifying. National Statistical Offices are expected to develop integrated systems of dissemination that permit to combine data from multiple sources and increase the completeness of statistical information. The development of communication technologies and, more importantly, the increased need for information, lead to more users requiring access to micro-data.

I. Introduction

1. The impact of technological progress on societies, economy and priorities of governments, and the increased volume and diversification of data sources freely available on the web have created new needs for statistical information. National Statistical Organisations (NSOs) and other statistical agencies offer the key source of official statistical information in every country. Originally created to serve mainly the interests of governments, NSOs are now expected to provide a wide range of products and services to a broad spectrum of users. These users differ greatly in their information needs and in their ability to manage statistical information. In this context, several challenges are faced by NSOs: they need to produce an increasing amount of consistent and relevant statistics; more timely data are needed on more topics than before; more detailed, spatial statistics are more often demanded; integration of data from multiple sources is a necessity for assessing cross-sectional issues; and finally new strategies and dissemination tools need to be developed to effectively supply information to different users.

2. Viewed from a customer management perspective, to achieve the maximum return on their (and public) investment, statistical agencies should plan, develop and align their products and services to the needs and expectations of key user groups. Further, they need to match the content and attributes of products to the idea of statistical literacy. In practical terms, this requires guaranteeing that the statistics incorporate different levels of detail, use graphic visualisation, are delivered through several channels, and are organized so as to respond to questions and concerns of different user groups. The same approach adopted for classical statistical products should be applied to micro-data access.

3. Students, public administration, private and public institutes need access to micro-data more often despite the complex procedures through which micro-data are provided by NSOs. If micro-data access has to become a key service for statistical agencies, then organisational competences, management processes, cultural norms, legislative frameworks and micro-data ownership need to be reconsidered (McMillan, 2010).

4. This paper reports on the experiences built in recent years at the National Institute of Statistics of Italy (Istat), presents the vision adopted by the Institute to respond to the challenges of micro-data access, and highlights general strategic issues to be addressed. In Section II, we report on the project that aims to create a network of data archives in Italy where Istat is a key player, and investigate the challenges facing NSOs as far as micro-data access is concerned. Section III shows how the modernization of statistical production has a strong impact on micro-data by aligning architectures with emerging standards and reference models for statistical information and processes. Section IV presents the current system of micro-data access services provided by Istat, its recent integration with the release of Public Use Files and corresponding emerging issues. Conclusions are presented in Section V.

II. Official statistics data archives

5. Defining a dissemination strategy has an influence on improving all stages of dissemination activities: characterising dissemination policies, designing products and services, preparing and presenting statistics, disseminating information on the website, and promoting and marketing products, services and “statistical releases” to users. However, if micro-data access has to become one of the key services for statistical agencies, then far more is needed than mere dissemination strategies.

6. Recently the input phase of statistical production, data capturing, has been transformed by new communication approaches (see, for example, Marske and Stempowski, 2008). The output phase should go through a similar reform where the user, for example the user of micro-data, is put into the centre of attention. This could be achieved by strategic changes: creating a single point for finding all relevant information, developing micro-data products and services tailored to user needs and providing services that facilitate connections with users.

7. To design a user-centric dissemination system for micro-data, the governance, the infrastructure and the legal framework need to be defined and put in place together with practical data access policies and flexible micro-data products and services. This also requires establishing different relations with micro-data users.

A. Governance

8. In Italy, the vision for a user-centric approach to micro-data access is promoted by a group of institutions. These institutions are forming connected data clearinghouses and related infrastructure that will provide the broadest possible access to publicly funded micro-data. This project is led by official statistics: Istat will be the hub of a joint venture that aims at developing a network of data archives that will deal with micro-data produced using public funding. In this network the Bank of Italy and Istat will manage micro-data for their respective topics, a third node will be in charge of research institutes micro-data and the Ministry of Education will encourage universities to contribute their micro-data. In the future, Istat should become the hub for micro-data access for data from government departments (like ministries) and the Italian National Statistical System.

B. Infrastructure

9. In Italy several public institutions provide micro-data access. Even though micro-data are provided, they are dispersed across many government institutions or research institutes. In terms of infrastructure, the Italian network of data archives will develop a coherent and single access to the micro-data products or services offered by public statistics. A web catalogue of all micro-data will be available and accessible to users both directly through a download or via requests for further specialised services (remote execution, remote access, etc.).

10. The infrastructure for remote micro-data access, developed by Istat, and the remote execution system in use at the Bank of Italy could jointly provide access to micro-data from the entire National Statistical System and other publicly funded data. This vision of a single access point via the web has been endorsed for example by the United Kingdom and is envisaged at European level by the Council of European Social Science Data Archives (CESSDA). Under the auspices of the 7th Framework Programme project Data without Boundaries (http://www.dwbproject.org/), 28 institutions are working on the coordination of existing infrastructures for access to official micro-data in Europe.

C. Legal framework

11. Istat, as a hub of the network of Italian Data Archives, should have the possibility of providing access to micro-data from other institutions of the national statistical system and managing remote access to micro-data from research institutes. A legal framework supporting such architecture is crucial. This approach shall become possible in Italy following the amendment of the Italian statistical law being prepared at present.

D. Data discovery, accessibility

12. A user-centric system allows for a quick and clear overview of the kind of micro-data that are available and under which conditions. User-friendly search protocols will be developed to help users to find the micro-data they need. Additionally, standardised metadata protocols, a milestone of any dissemination system, need to be carefully selected to guide data interpretation. Finally, the network of data archives will work towards harmonisation of micro-data access policies inside the Italian national statistical system to support the management of requests to different institutions.

E. Relationship with users

13. Besides the cooperation among the data archives, the focus of the network will be the collaboration with users and statistical literacy. The users of micro-data need knowledge in statistics, analytics, data analysis and computer programming. The process of data analysis should be understood in its simplest form by most citizens. It should, therefore, be taught to students and should be the core competency for the staff of public administration.

14. Having powerful tools for analysing statistical micro-data is not sufficient. Adequate competences should be developed for selecting suitable methods, applying them correctly and understanding and interpreting the results. Statistical literacy should be constantly promoted and supported by NSOs both inside their national statistical systems (government agencies, public administrations) and outside (undergraduates, master and PhD students, young researchers, etc.).

15. In response to the increasing need for quantitative literacy coming from the society and public administration, Istat established, in 2011, the Advanced School for Statistics and Socio-Economic Analyses (Scuola Superiore di Statistica e Analisi Economica, SAES). The school promotes statistical literacy and offers training programs on advanced survey techniques and statistical methodologies, tailored training, traineeships. The aim is twofold: to increase statistical knowledge and analytics in the public administration and to train future users of official statistics.

16. The network of data archives and SAES will collaborate to offer training programs for different needs. But most importantly the network will pave the way for changes in the relationship between producers and users of official statistics. Partnerships are currently being created to collaborate with Istat, to improve survey design, to propose changes to questionnaires used in data collection and to improve usefulness of data. Users will not just be data analysts but will be increasingly called upon for an active contribution toward the improvements of official statistics.

III. Micro-data: a product in its own right

A. Standardization and quality

17. In recent years, official statistics have been confronted with both accelerating changes in society and resource constraints that have led chief statisticians to look critically at the whole architecture of the statistical production process. In Europe this is stated in the *Communication from the Commission to the European Parliament and the Council on the production of EU statistics: a vision for the next decade (COM 404/2009)*. At the worldwide level, the initiative of the High-Level Group on modernization of statistical production and services (HLG) has strengthened the process of standardization and industrialization of official statistics.

18. The HLG seeks to re-use and share methods, components, processes and data repositories and adopt a shared "plug-and-play" modular component architecture with the aim of increasing efficiency. The implementation of such an architecture can be made possible only by sharing common standards as represented by the Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM). The former is necessary for defining common components of the statistical process; the latter is a reference framework of information objects which enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process. Together they represent the starting point for defining a common language and for setting up an integrated statistical production system.

19. In Europe, a strategic task force, the Sponsorship on Standardisation, has been set up to advise the European Statistical System on how to pursue standardisation and integration (see Braaksma, et al. 2013). At Istat a project has been launched, Stat2015 (see Falorsi et al. 2013), whose aim is the standardisation and industrialisation of production processes based on re-use and on the adoption of a model founded on shared services in a service-oriented architecture (SOA) framework.

20. According to the vision launched by the HLG, validated micro-data stem from a standardized and harmonized process and are accompanied by appropriate metadata enabling their unambiguous use for different purposes. The validated micro-data are, in the GSBPM, singled out in a specific sub-process 5.8 (Finalize Data Files) to state the importance of the achievement: validated data are, themselves, the output of a standardized process. Moreover, adopting the perspective of a statistical process which is completely metadata-driven is useful for the quality of micro-data. With the use of standardised protocols, data and metadata can be linked as early as in the data capture phase -e.g. by means of the Data Documentation Initiative (DDI) and/or the Statistical Data and Metadata Exchange (SDMX)-. For micro-data to be considered a product in their own right their level of quality needs to be high.

B. Impartiality, transparency and public trust

21. Through its web-site, Istat, like other NSOs, disseminates many statistical products: indicators, tables, maps, graphs, and so on. Dissemination and communication of "value added" products requires making choices about what data or results to highlight and on which findings and implications to focus. Micro-data-based products may contribute to increasing the transparency of NSO's work and, more importantly, to increasing its credibility and public trust. Indeed, only access to micro-data allows users to replicate the statistics released by NSOs, perform analyses and comparisons, thus contributing also to the continuous adaptation of the statistical system to the society's information needs.

22. The high standards of quality, as well as strict ethical and professional principles at the base of the production of the statistical information should encourage NSOs to open their micro-data resources, in full compliance with confidentiality principles. Transparency, impartiality and neutrality can be increased by adopting micro-data as a product in its own right.

C. Micro-data integrated from multiple sources

23. Improving micro-data access requires more than quality and infrastructure considerations. It should be recognized that users need micro-data access to analyse complex phenomena not addressed by readily available statistics. This implies that data from a single survey may not be sufficient for capturing complex issues of the society. Here

are also the strengths of the system of official statistics: often NSOs are equipped with a legal framework allowing them to retrieve data/micro-data/registers from public administration for the purpose of statistical production. This makes the NSO a natural host for micro-data integration processes inside the national statistical systems.

24. At European level several projects are developing a common platform to share, process and redistribute information in the European Statistical System (Euro-Groups Register (EGR) project, and European Union Statistics on Income and Living Conditions (EU-SILC) data warehouses of micro data on income and living conditions). This means that technologies are already in place that can change the way micro-data are processed and accessed. This opens new ways to make use of micro-data, thus increasing their importance as a product in its own right. These technological and process infrastructures push the frontiers for micro-data (re)use while stressing once again the substantial requirement for micro-data to be considered as a product in its own right.

25. It is probable that production of statistical information based on integrated micro-data will become in short-term a best practice adopted by the NSOs. Simultaneously, integrated micro-data as the key source for examining complex phenomena will tempt more users to access micro-data in order to improve their statistical competence. It is obvious that such analyses are meaningful only when dealing with well-defined objects, i.e. only if primary micro-data are considered an official product.

IV. Different micro-data services for different needs

26. Together with tools and services, products represent the core of data archiving. In order to meet new and heterogeneous needs of data users, in recent years Istat has varied existing modes of access and has developed new products and services.

27. Currently the offer ranges from the freely downloadable aggregates queried through the corporate dissemination data warehouse (Istat: http://dati.istat.it/?lang=en) to customized data processing (https://contact.istat.it/Index.php?Lingua=Inglese).

28. As for micro-data, besides providing anonymised unit level data from social surveys available at request to everyone, named *file standard*, Micro-data Files for Research (MFR) have been developed for scholars containing much more detailed information (http://en.istat.it/dati/microdati/file_microdati.html#file_ricerca).

29. Finally, the network of points of access to the Research Data Centre (RDC) (Laboratorio per l'Analisi di Dati Elementari), allows researchers to analyse any micro-data produced by Istat in any of the 18 Istat regional offices across Italy (http://www.istat.it/en/information/researchers/analysis-of-individual-data). This network based on a thin-client system (network of servers) allows researchers to access the original confidential data through a secure channel.

A. Public Use File and the philosophy of re-use

30. The Istat system of micro-data access will be developed further during 2013 with the adoption of a public use file (PUF), named mIcro.STAT. These files are samples of individual data with well-defined characteristics:

(a) They are freely available on the web (no need to sign any access agreement);

(b) They allow users to make inferences on the phenomena related to the data;

(c) They are usually released under no restrictions or conditions on their use.

31. Sub-sampling techniques are used, together with other protection methods, for the production of PUFs (see Hundepool et al. 2012). Sub-sampling techniques reduce the risk of disclosure by increasing the uncertainty on the number of population units sharing the same score on identifying variables. In order to increase efficiency in the production of diversified products, Istat has followed a different approach.

32. mIcro.STAT is a perfect example of the "re-use" of micro-data. Indeed, for the PUF production two elements are re-used: an already existing product, the micro-data file for research (MFR) and the sampling competences of Istat. In order to gain efficiency in the production of PUF, a new methodological solution has been developed based on sub-sampling from the corresponding MFR (see Foschi et al., 2012). The PUF and the MFR therefore share the same structure. Such a hierarchical structure of the two data files greatly simplifies assessment of the disclosure risk and reduces information loss associated with the anonymisation procedure. Another advantage of this approach is that it preserves the hierarchical detail of data and the internal consistency of the two sets of records. This allows for a reduction in the cost of preparation of the files, therefore increasing efficiency.

33. The combined use of different sampling techniques e.g. multivariate multi-domain allocation (see Bethel 1989) and balanced sampling (see Deville J., Tillé 2004), allows for a simultaneous controlled reduction of the disclosure risk and the preservation of pre-defined data utility indicators. The goal is to provide a PUF satisfying as many risk and utility requirements as possible.

B. Reasons for public use files: democracy and statistical literacy

34. The development of a public use file that shares the same details, quality and complexity of the corresponding file for research purposes is based on the principles of democracy of access and the right to research. Furthermore, only by allowing students to be trained on complex official data will statistical literacy increase in the country.

35. Teaching files usually contain very few variables, basic information, few observations and show a simplified structure of data. To overcome these limitations, students need to learn how to practice using real survey micro-data, how to produce statistical analysis out of raw data and, more importantly, how to build knowledge out of data. These features are essential if we want to increase the statistical literacy of the population in general and of students in particular.

36. When the PUF satisfies predefined quality standards, it could positively contribute to statistical literacy. The value added of a PUF is a direct result of its quality standards defined as the ability to simulate real applications. The file dimension, expressed as a number of records and a number of variables, provides a first quality indicator. Moreover, since the data production process and data quality are not extensively discussed in lectures on statistics, any PUF could contribute to the reduction of this gap. At the same time, a large number of variables would favour the development of critical reasoning on the variables' meaning, their operational definition, the surveyed phenomenon, etc. The precision and accuracy of the estimates that could be derived using the PUF would significantly improve learning of statistical methodology. Table 1 shows which mIcro.STAT characteristics suit the process of knowledge extraction.

37. Finally, the activities related to the dissemination of PUFs include the definition (possibly internationally shared) of a license for use as well as tools and services to be developed for analysing micro-data products. On the one hand, micro-data offer the greatest possible flexibility when analysing a phenomenon. On the other hand, by definition, micro-data are not user-friendly and do not offer the possibility of immediate results. Only the design and development of adequate tools and services for micro-data analysis would

improve the usability of data. A clear licensing process and the development of new services are strategic issues when dealing with micro-data access facilities.

Table1
Characteristics of the public use file (mIcro.STAT) and micro-data files for research and their relationship with the corresponding process of knowledge extraction

mIcro.STAT	Knowledge extraction	MFR
	Problem definition – hypothesis formulation	
✓	Data reading (format, documentation, classifications)	✓
✓	Initial data analysis	✓
✓	Use of statistical methods (modelling, cluster analysis, etc.)	✓
✓	Interpretation of the results	✓
✓	Audit of the results	✓
✓	Comparisons	✓
✓	Consequences	✓
✓	Hypothesis re-definition	✓
✓	Report e presentation of the results	✓
✓	Development of statistical application	✓
	Development of socio-economic theories	✓
	Policy decision making	✓

V. Conclusions

38. Analyses of micro-data are invaluable resources for government planners, market analysts, academia researchers and citizens to take adequate evidence-based decisions. NSOs need to respond to the challenges of improving access to their high quality micro-data for scientific research purposes as well as for society as a whole, while meeting the requirements of the Fundamental Principles of Official Statistics.

39. The great diversity of user groups requires a differentiated approach in terms of statistical data dissemination, i.e. to identify the requirements on one side and the statistical supply on the other. In a dynamic approach, it requires: knowing users' information needs, adjusting the supply of information to such needs and adapting to the information society's priority changes.

40. Modernisation of statistical production makes new ambitious objectives feasible for NSOs. For instance, it facilitates moving from sector innovation to continuous systemic innovation, and thus supports a strategic shift for NSOs to develop from data producers to providers of statistical products and services. NSOs can be the leaders in designing data archives for national micro-data.

41. The creation of data archives cannot be reduced to mere infrastructure or a catalogue of files available. A culture and knowledge around micro-data has to be developed. Statistical literacy need to be increased and NSOs should be proactive in providing knowledge based services for micro-data access that answer users' demands on the basis of principles of democracy (PUF) and meritocracy (MFR).

42. The crucial issue that remains to be solved is the governance of all these processes. Products and tools can be developed, systems can be designed and built, architectures can be adapted but a strong leadership and governance of these developments is essential for successfully managing micro-data access.

VI. References

Braaksma, B., Colasanti, C., Falorsi, P.D., Kloek, W., Martinez Vidal, M. and Museux, J.M. (2013). *Standardisation in the European Statistical System*, NTTS Conferences on New Techniques and Technologies in Statistics, Brussels, 5-7 March, 2013, available at: www.NTTS2013.eu.

Bethel, J.W. (1989). *Sample Allocation in Multivariate Surveys*. Survey Methodology, Vol. 15, pp. 47-57.

Deville J., Tillé Y. (2004). *Efficient Balanced Sampling: The Cube Method.* Biometrika. 91(4). 893-912.

Falorsi, P.D., Barcaroli, G., Fasano, A. and Mignolli, N. (2013). *A Business Architecture framework for industrialisation and standardisation in a modern National Statistical Institute*. NTTS Conferences on New Techniques and Technologies in Statistics, Brussels, 5-7 March, 2013, available at: www.NTTS2013.eu.

Foschi, F., Casciano, C., Ichim, D. and Franconi, L. (2012). *Designing Multiple Releases from the Small and Medium Enterprises Survey*, In J. Domingo Ferrer and I. Tinnirello (eds) Proceeding of PSD2012, Vol. 7556, Lecture Notes in Computer Science, pp 200-215. Springer, Berlin/Heidelberg.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and De Wolf, P.P. (2012). *Statistical Disclosure Control*. Wiley.

MacMillan, P. (2010). Unlocking Government: *How Data Will Transform Democracy*, Commonwealth Innovation, Vol. 16, 2, pp. 13-17.

Marske R. and Stempowski D.M. (2008). *Company-centric Communication approaches for Business Survey Response Management.* Proceedings of Statistics Canada Symposium 2008. Data collection Challenges, Achievements and New Directions.