WP. 47
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (vii): General statistical confidentiality issues

LEGAL, POLITICAL AND METHODOLOGICAL ISSUES IN CONFIDENTIALITY IN THE ESS

Invited Paper

Submitted by Eurostat[1]

[1] Prepared by Maria João Santos and Jean Marc Museux.

Legal, political and methodological issues in confidentiality in the ESS

Maria João Santos[1] and Jean Marc Museux[2]

[1] Methodology and Research Unit, Eurostat, L-2920 Luxembourg
[2] Living conditions and social protection statistics, Eurostat, L-2920 Luxembourg

Abstract. All member countries in Europe face similar problems with respect to Statistical disclosure control (SDC). They all need to find a balance between preservation of privacy for the respondents and the very legitimate requests of society, researchers and policy makers to provide more and more detailed information. This growing demand, due to developments of the information age and knowledge society is a common problem of the European Statistical System (ESS). SDC is also a critical issue for Eurostat because it is at the core of the delicate trust data providers have towards statistics compilers. It influences greatly the quality of EU statistics and consequently the relationship between Eurostat and ESS. In addition, the regulatory framework on statistics includes strict rules to ensure that the information provided by respondents is adequately protected from disclosure. In order to meet the European challenge the SDC problems connected to it have to be approached by all countries in the coming years. In the paper current Eurostat confidentiality issues and strategy are discussed.

Keywords. Keywords for index, separated by commas, without full-stop at end

1 Introduction

The objective of this paper is to provide an overview of the various issues related to confidentiality in a European wide perspective. It aims to give technical experts an idea of the difficulties raised by the multinational and administrative perspective which might not be perceived at first sight. The variety of perception, the lack of well defined standard is a source of diversity that renders standard confidentiality problem much more problematic at European level. This paper calls for a closer partnership between administrative and research community and for a strong scientific research input and responsibility in order to design best practices to feed legal reflection at European level.

2 Confidentiality legal framework

2.1 General framework

The right to privacy is a fundamental right. It includes the protection of the person in the context of personal data processing. That means for instance the right to receive certain information, the right to access the data, the right to have the data corrected, etc. Statistical confidentiality primarily aims at safeguarding privacy in the field of statistics and is a key to the necessary trust that has to be maintained between statistical bodies and respondents. Mutual confidence ensures accurate and reliable basic information and eventually high quality statistics.

At EU level, statistical confidentiality is addressed in the following legal acts:
- Council Regulation (EEC, Euratom) No 1588/90 of 11 June 1990 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities;
- Council Regulation (EC) No 322/97 of 17 February 1997 on Community statistics;
- Commission Decision 97/281/EC of 21 April 1997 on the role of Eurostat as regards the production of Community statistics;
- Commission Regulation (EC) No 831/2002 of 17 May 2002 implementing Council
- Regulation (EC) No 322/97 on Community Statistics, concerning access to confidential data for scientific purposes;
- Commission Decision 2004/452/EC of 29 April 2004 laying down a list of bodies whose researchers may access data for confidential purposes.

Statistical confidentiality is regulated at EU level only to the extent to which statistical activities carried out by Eurostat and the national statistical authorities for the production of Community statistics are concerned. Specific confidentiality regimes still coexist at national level and differences may appear with the EU statistical confidentiality regime. These differences are less on the substance (the general concepts are common to a very large extent) than on the perception of the issue (the national framework remains the frame of reference), which is equally important.

Thus, the existent statistical confidentiality regime is not unified in one regulation, which leads to difficulties of interpretation between MS and the Commission and renders difficult current work in different sectors. Improving the existing framework should contribute to avoiding repeated discussions and even in some cases obstacles when dealing with confidentiality issues in the context of the negotiation of sectoral regulations.

At the moment there is an ongoing reflection at Eurostat and at MS level on the need to revise the legal framework based in the principles of maximising the quality of European Statistics both produced by Member States and European Institutions and increase the possibility of secondary use of the data by the research community and the general public; while at the same time respecting the confidentiality mandate to preserve the direct or indirect disclosure of individual information.

The proposed revision could pass by proposing amendments of the legal framework in several domains. In what concerns the transmission of confidential data, could be envisaged the modification of the provision given by art. 14 of Reg 322/97 on the transmission of confidential data without direct identifiers, towards a regime where the transmission and exchange should cover confidential data as defined objectively by Article 13 of Regulation 322/97, covering thus the full range of confidential data. This transmission and exchange should be allowed: between MS and between MS and Eurostat and whenever it concerns and to the extent it is necessary for the production and the quality of Community statistics.

The concept that publicly available information should not be considered confidential already covered by Art. 13 of Regulation 322/97 should be more systematically implemented, possibly through a specific legal act defining variables and fields that are publicly available according to accounting EU directives. In parallel Article 13 §, could be amended in order to ease its implementation by withdrawing the specification: “and remain available to the public at the national authorities”, which is seen as additional constraint for its implementation.

The wide acceptation of an objective basis for declaring data confidential and measuring disclosure risk would definitively ease legal progress in the field of statistical confidentiality. Scientific researcher’s authority in certainly required to put a cut off to the endless subjective discussion. Lawyers are waiting for a strong technical input in order to design harmonised legislation.

2.2 Access to researchers

There is a growing appreciation of the benefits of providing access to microdata for research and analysis. At the same time it is vital to protect data confidentiality. It is essential that new approaches are developed to meet these objectives which create conflicting pressures. The risks to confidentiality must be managed effectively. A key challenge is how to minimise the risks to confidentiality, including the perception of threats to confidentiality. Striking the right balance is vital.

Complex policy making requires multivariate causal thinking about policy alternatives, which in turn, require complex, multivariate, often longitudinal data. As the economy grows more complex and the population becomes more diverse, increasingly detailed data and data analysis are required for policies to match well with economic and demographic alternatives.

An effective public-private partnership between data collection institutes and the research community is a critical element in bringing analyses of complex data, particularly microdata, to bear on policy design and assessment. This partnership between NSI and research is of mutual benefit and is strengthened by continuous improvements in data access, both through public use data sets and through restricted data access modalities. The relationship between data use and data quality is the essential foundation for the common interest of the statistical system and the broader research community in broad and responsible access to data.

There is a need to explore new avenues of access of data to researchers and in parallel improving the current instruments.

Streamlining the implementation of Commission Regulation 831/2002

A detailed description and analysis of this legal act in the paper presented by John King and Jean Louis Mercy in the Work Session on Statistical Data Confidentiality held in Luxembourg on 7-9 April 2003. While this regulation sets important hopes for the availability of microdata to the research community, its implementation has faced several difficulties which have made its development progress at a slow pace.

The committee statistical confidentiality (CSC) of December 2004 has analysed the progress in the implementation of this Regulation and has agreed on the development of quick procedures to process the requests of researchers and to grant the eligibility of research institutions. These fast track procedures will the presented to the CSC on the next meeting in December 2005; their adoption will improve the timeliness and efficacy of the regulation.

There are two levels of access to microdata:

Level one: **Confidential data as obtained from the national authorities.** They allow only indirect identification of the statistical units concerned. This access is done through the use of a safe centre at Eurostat.

Level two: **Sets of anonymised microdata extracted from the above data**. They are individual statistical records which have been modified in order to minimise, in accordance with current best practice, the risk of identification of the statistical units to which they relate.

This access is done via distribution of encrypted CD-ROM according to contracts established between Eurostat and the corresponding institutions.

At present microdata for researchers for level two can only be provided for three statistical domains. These are the European Community Household Panel (ECHP, CVTS (Continuing Vocational Training Survey) and the Labour Force Survey (LFS). In addition, the Community Innovation Survey Working Group is now discussing criteria to distribute microdata files of this investigation. Furthermore, a task force has been set up to do the same exercise for the coming Survey on Income and Living Conditions (EU-SILC).

The necessary measures are going to be taken to propose adding other microdata sets to the ones mentioned in Commission Regulation 831/2002 such as SES (Structure of Earnings Survey).

The advantage of possibilities offered by this regulation is that researchers now have the possibility to have access to harmonized datasets spanning all Member States (MS), before gaining access to data for each of the MS has involved a lengthy process of making requests to each MS. This gives researchers opportunities for pan-European Union research and analyses. The table below presents a synthesis of the projects reported by those research institutions which, during 2004, submitted to Eurostat requests of micro-data of the European Household panel (ECHP).

Research contracts using ECHP data. Year 2004. Main Topics	
Studies of specific sub-populations	*Studies of specific phenomena*
Elderly	Mobility
Poor	Income inequality
Regions	Transition employment <-> unemployment
Long-term unemployed	Taxation, subsidies
Married women	Intra-family transfers
Female participation in labour	Inequality in income and education
Divorced	Wage changes
Temporary Workers	Education and Health
Persons at end of working life	Labour market participation and fertility
Youth	Childcare
	Discrimination

Regulation 831/2002 foresees (article 3) a fairly straightforward and simple request process for researchers from two categories of organisations:
1(a), i.e. universities and other higher education organisations established by Community law or by the law of a Member State; or
1(b), i.e. organisations or institutions for scientific research established under Community law or under the law of a Member State.

For “other bodies”, article 3 of regulation 831/2002 lays down the condition that they must first be approved by the CSC if they wish to make requests to access confidential data for scientific purposes. Commission Decision 2004/452/CE list other bodies that have been considered admissible. The prerequisite to achieve admissibility is that the institution has demonstrated that it fulfils a set of criteria. The CSC has approved these criteria at its meeting of 10 December 2004. Specific services of EU Institutions, which carry out statistical activities, may be considered eligible as researchers for access to specific confidential micro-files provided that the equivalent guarantees are provided. This follows the precedent established with the ECB and the Central Banks of Spain and Italy. Universities based outside Europe can also be considered as admissible; the University of Cornell (USA) was the first to be included in this list. The efforts will continue to extend the list of other bodies than can be regarded as admissible.

Establishment of bilateral agreements on licensing and delocalisation of safe centres
An important component of developing a new confidentiality protection system is the development of a safe centre network. At the moment the safe centre for the data sets mentioned under Commission Regulation 831/2002 is localized at Eurostat. Eurostat will discuss with the MS the possibilities to delocalise via the establishment of bilateral agreements the safe centres to MS or to create the conditions to establish licensing agreement with established institutions.

3 Methodological Issues

In general the legislation at national and European levels is fairly harmonised with respect to what is considered as confidential data. However, when implementing this legislation, the criteria used differ considerably from country to country. These criteria have sometimes an important historical weight; sometimes do not have a solid scientific basis; and in many cases lead to conservative solutions because real risks are not well mastered.

This diversity of interpretations is a consequence of the fact that there is no harmonised approach of disclosure risk. To agree on disclosure risk, one should agree first on the sensitivity of the data (how “private” are the variables in the file) and on the possibility to match these data with external sources, that is, to the presence of key variables or identifying variables. Second, there is a need to find a harmonised way to measure the risk. Methodological work is needed to reconcile the different approaches or to express preference for one of them.

It is obvious here the need to have common core criteria which, while providing a satisfactory harmonisation level, allow for a degree of flexibility to adapt to the

specific perception of the society in each country. This will also have the advantage of having a more solid internationally agreed basis that better justifies national choices made in the release of microdata.

Disclosure protection of EU aggregates

Most of the time, Eurostat compiles EU aggregates on the basis of national aggregates. These are accompanied with a confidentiality flag informing Eurostat that the information should be treated as confidential. In the best situation, Eurostat is also informed on the presence of dominance in these aggregates. However, meta information is not standardised and even sometimes there exists confusion between not publishable because of lack of reliability and confidential as meant in the legal framework.

To declare information as (primary) confidential, MS use measures of risk of disclosure of individual information (dominance rules, threshold rules) which are not harmonised. The level of protection can varies between MS depending on different perceptions of the level of disclosure risk and also simply of the perception of the damage of disclosure itself. Distinction is rarely made between variables themselves: some variables might be considered as non sensitive whereas other from the same record could be.

The lack of harmonisation of primary confidential rules causes major methodological problems at Eurostat level. Software packages for handling secondary confidentiality are not designed to deal with such a situation. For instance, the input required, mainly micro data, does not fit Eurostat situation which deals with aggregated data. Consequently, Eurostat, following the most stringent rules used by national authorities to protect EU data, is led to over protect data and not to release useful information for the user. The lack of harmonisation of disclosure protection measures between MS hampered thus the release of European data. This situation could be improved by rising the awareness of disclosure control issues and if statistical disclosure experts would issue a unified set of best practices accompanied with practical hints to implement them. This would be developed in a European perspective.

Disclosure protection of micro data

To some extend the same holds when Eurostat has to design, in collaboration with MS, anonymisation of micro data to be released to researchers. Despite they share common objectives:

- the need to follow Regulation principles on the right for privacy,
- the need to maintain the trust the respondent have in the statistical system,
- the need to monitor the release so to avoid confidentiality breach),

the differences in the perception of the risk and the lack of a universal measure of risk render the possibility of a consensus very thin. Part of the problems lies in the absence of knowledge of real risk.

This situation would be improved if once again, European experts would agree on a set of measure and threshold to be used by practitioners. This probably needs more comparative research to be developed on the existing measures and on the tuning of methods. In parallel, more research could be carried out on the measure of the actual measure of risk. Computer scientist could design protocol to crack released European database, which in turn could be used to develop appropriate protection measures.

4 Conclusions and future perspectives

With respect to medium term perspectives, three main components are already identified:

SDC in the ESS

The Setting up of a Centres and Networks of Excellence (CENEX) in Statistical Disclosure Control. CENEX originates from the idea of sharing the work between different institutions within the ESS (European Statistical System) more efficiently, by providing adequate organisational solutions and institutional framework for modern types of cooperation and specialization of work.

In the latter case, the sharing of work between different Member States in a CENEX will creates synergies since each participating NSI will concentrate on specific areas and the product of this work will be beneficial to all NSI ultimately leading to the increase of the quality of ESS statistics. It is moreover essential for the generation of comparable statistical information across countries, and on the European level that similar methods and tools are used to protect confidentiality in the published information. As long as member states compile their statistics using different statistical disclosure control (SDC) methods, the compilation of European statistics is very much hampered.

The pilot CENEX on SDC was defined to address in a first phase the following objectives:

- Set standards for the protection of micro-data sets, based on disclosure risk assessment methods and criteria.
- Improve tabular data protection techniques and develop harmonized criteria
- Extend and develop SDC software tools, both for micro and tabular data, so as to fit the specific production and dissemination environments of ESS.

Eurostat plans to evaluate and further develop the CENEX approach to harmonise SDC practices in the ESS, promote the definition and use of best practices, bring up to level SDC software tools in the ESS and the remote access to microdata.

Public use files

Public use files (PUF) are the most accessible; widely and freely used microdata products made available by statistical institutes, but their value for policy for much policy relevant research is limited. Nevertheless these files are useful for some research purposes, as teaching aids and are a good advertisement of a statistical institute. Continued distribution of public use files is threatened by the increased re-identification risk associated with both technological advances in linking software and widespread availability of administrative records. During the last decade researchers have developed increasingly sophisticated methodologies for restricted data products. The development of a methodology for generating synthetic or virtual data is a relatively recent activity. A key objective of the method is to preserve faithful representations of the original data so that inferences from the synthetic data are as consistent as possible with the inferences that would be drawn from the original data. One attractive feature of the synthetic data approach is that it can be used to create multiple public use files from the same underlying data – targeted at different audiences. The methodology of synthetic files as a measure to replace public use files need to be further researched.

The work at the sectoral level to establish the criteria for establishing public use files (such as the on going work of the EU-SILC TF on anonymisation and establishment of public use files) will continue to be promoted in the future via the establishment of sectoral TFs that will define PUF for each survey.

Monitored remote access to microdata

A sensible approach for facilitating high quality research is to maintain the data in a secure, restricted remote access environment.

Monitored remote access has the advantage that a researcher does not have to go to a safe centre to make use of confidential data and output is returned relatively quickly. This approach to develop remote access procedures, which has the advantage of reducing researcher burden, involves substantial investment in hardware and software. This approach has been gathering momentum and is now operational in Europe in the Netherlands, Denmark and Sweden. It will be studied with MS the possibility offered by the 7th Research Framework Program in the field of research infrastructures to further develop such an approach at European level.

Some of the requirements and targets specified in laws are not fixed but are moving over time. There is thus a requirement on NSIs and on Eurostat to review practices and methods from time to time. It has been presented some of the more long term

threads to be followed in the future regarding the modification of the current legal framework. In parallel were described concrete axes of implementation reflecting the orientations of Eurostat in short to medium term with respect to confidentiality. Eurostat hopes to develop fruitful synergies with experts and NSI along these axes.

References

Jean-Louis Mercy and John King "Developments at Eurostat for research access to confidential data" Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003) Working Paper 12.