Working Paper No. 12
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 ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (iii): Emerging legal/regulatory issues

DEVELOPMENTS AT EUROSTAT FOR RESEARCH ACCESS TO CONFIDENTIAL DATA

Invited paper

Submitted by Eurostat[1]

[1] Prepared by Jean-Louis Mercy (jean-louis.mercy@cec.eu.int), European Commission, L-2920 Luxembourg and John King (email: johnrbking@waitrose.com).

Developments at Eurostat for research access to confidential data

Abstract

The background to a new Regulation (Regulation 831/2002 concerning access to confidential data for scientific purposes) is described and the main provisions of the Regulation are described. The implications of the regulation, for Eurostat, Member States, the research community and even data subjects, are considered. Eurostat's activities in implementing the Regulation are outlined together with an indication of some of the outstanding issues. Implementation of the Regulation has also raised some further questions on matters as diverse as the possibility of remote access to confidential data and the meaning of "scientific purposes".

About a year ago the European Commission adopted Regulation 831/2002 concerning access to confidential data for scientific purposes. This was a significant step in providing better access to confidential data for research. This paper describes some of the background to the regulation; outlines the provisions of the regulation and the steps Eurostat is taking to implement the regulation; discusses some of the implications of this work; and indicates some further questions arising from this work.

1. Background to Regulation 831/2002

Micro-datasets are becoming important because of increasing interest in accessing them by researchers. This interest has two related drivers. The first is an aspect of modern life—accountable government and transparency. This is reflected in an increasing interest in and demand for evidence-based policy, policy analysis, and monitoring policies and their impact. This kind of activity requires timely, detailed information and frequently requires more detailed analyses than are presently published by statistical organisations. Sometimes these analyses are seen as being outside the remit of national statistical organisations (NSIs) or even as activities that could compromise the perceived independence of NSIs. Indeed, these analyses are performed often by academic institutions or independent research institutions.

Detailed data are needed for these types of analyses. The obvious and most relevant source is often identified as the data collected and held by NSIs. Hence there is an increasing pressure on NSIs and other statistical organisations to provide detailed data on a wide range of topics. In particular, for the European Union (EU), pan-EU analyses and research are becoming more and more important. The same could also be said for the Euro-zone. So the need is for access to pan-EU datasets for this research. Eurostat holds many such datasets, and so it is seen, by analogy with the national situation, as the natural, simple and direct potential source for these datasets.

The second driver here is the changing nature of research itself. Much modern research cannot be satisfied with aggregate data—micro-data are needed for fine analysis and model building. Hand-in-hand with this there has been an evolution (perhaps revolution would be a more appropriate description) of research computing capacity—both hardware and software tools—and in the number of researchers and research institutions. These factors have considerably increased the demand for access to micro-data records for computing correlation matrices, estimating models and other analyses, depending on the context of the research topic.

Examples of the micro-data needs of researchers were given in papers by, for example, Westergaard-Nielsen and Blundell at the CEIES (European Advisory Committee on Statistical Information in the Economic and Social Spheres) seminar (19th seminar) on "Innovative solutions in providing access to micro-data" last October in Lisbon. Other examples were given by several of the speakers, including Dilnot, Vickers and Blundell, at the inaugural conference in December 2001 of the cemmap (Centre for microdata methods and practice) research centre in London.

At the same time, statistical organisations, both NSIs and supra-national and international institutions, are increasingly seeing making more use of the data held by them as an important contribution to society and as part of an obligation to make better use of their resources (data). But there are constraints on what

statistical organisations, particularly NSIs, can do and on how they can do it. The role of researchers and research organisations is thus an important one, and it is an increasing one too.

Because of its role of producing statistical information for the European Union, Eurostat collects data from the Member States (MSs)on many aspects of economic and social life. These data sets are, broadly, comparable across the MSs and use harmonised definitions. So the datasets held by Eurostat represent a rich and valuable resource for the Commission, the MSs, and potentially, researchers. The data collected and held by Eurostat are the subject of regulations. The regulations represent agreements between the Commission and the MSs on the purposes for which data are provided and conditions under which the data are provided—in essence, statements of what can and cannot be done with the data. The data are held subject to the requirements and conditions imposed by the MSs—this is stated explicitly in some of the regulations.

The principle of statistical confidentiality is effectively the contract connecting the statistician with all those providing their individual data, either voluntarily, as is frequently the case, or by legal obligation, with a view to producing the statistical data essential for the society as a whole. From the formal legal point of view most of the European countries have established legal provisions for statistical confidentiality a long time ago. At the European level, the principle has been enshrined in Article 285 of the Treaty establishing the European Community as a fundamental principle for Community statistics. Article 285 provides that the production of Community statistics shall conform to the principles of impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality. The confidentiality principle is therefore part of the European basic charter and has thus acquired the highest status in legal terms.

The principle has been further specified and data received, held, used and disseminated by Eurostat are controlled by a set of legislations that have developed since the Treaty founding the European Communities. In 1990, Council Regulation 1588/90 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Community set out basic rules and safeguards for the handling of confidential data. Subsequently, in 1997, the "Statistical Law"—EU regulation 322/1997 on Community Statistics—expanded on these basic rules. In particular, a legal definition of statistical disclosure was introduced. Article 13 states:

> "1. Data used by the national authorities and the Community authority for the production of Community statistics shall be considered confidential when they allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information.
>
> To determine whether a statistical unit is identifiable, account shall be taken of all the means that might reasonably be used by a third party to identify the said statistical unit."

This definition has replaced the former definition laid down in Regulation 1588/90 where confidential data were defined as "data declared confidential by the Member States in line with national legislation or practices governing statistical confidentiality." The notion of confidential data has consequently become an objective notion with a clear Community dimension.

Article 13 goes on to state:

> "2. By derogation from paragraph 1, data taken from sources which are available to the public and remain available to the public at the national authorities according to national legislation, shall not be considered confidential."

The Statistical Law also states that confidential data must be used exclusively for statistical purposes unless the respondents have unambiguously given their consent to the use for any other purposes (article 15). The law also makes provision for access to confidential data for scientific purposes (article 17).

With the agreement of all the MSs, the latter provision was used to provide simple access to data of the European Community Household Panel (ECHP). An anonymised micro-dataset was developed (by Eurostat in collaboration with the MSs) and made available under certain conditions to researchers.

The provision has also been used by several enterprising researchers who have wished to use pan-EU microdata for their research. The researchers have had to contact the national statistical authority in each

MS to request permission to access the data of that MS from a particular survey. Eurostat is then authorised to provide access to data of the MSs so agreeing. There has been mixed success with this approach, depending on the type of survey or data requested—sometimes MSs deny access to their data.

2. What Regulation 831/2002 sets out to do

Regulation 831/2002 implements certain provisions of the Statistical Law (regulation 322/97), particularly articles 17(2) and 20 (1). Essentially, Regulation 831/2002 sets out simplified procedures under which access to confidential data for scientific purposes may be granted. For many researchers it attempts to remove some of the access burden implicit in the Statistical Law, although access is still subject to comment by the national statistical authority of each MS and to various conditions. The regulation refers to four important sources:

- European Community Household Panel (ECHP);
- Labour Force Survey (LFS);
- Community Innovation Survey (CIS);
- Continuing Vocational Training Survey (CVTS).

In summary, researchers must belong to research institutions and organisations within the MSs (other researchers or organisations have to go through a more lengthy approval process). A detailed proposal must be prepared stating the purpose of the research and details of the data to be used. Safeguards for the secure holding of the datasets will be necessary and controls on access by individuals will be required. Agreement to conditions and safeguards will be through a contract with the researchers' institution. There is no right of access to confidential data under the Regulation. In addition, MSs can withhold the data of their country from any particular research request. Access to confidential datasets can be on the premises of Eurostat with checks on the output and results to maintain confidentiality; or access can be through distributions of anonymised micro-datasets. Agreement by the researchers to conditions and safeguards will be through a contract with their organisation.

Incidentally, the new Regulation 831/2002 now provides a legal definition of anonymised micro-datasets. " "anonymised microdata" shall mean 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."

3. Implementing Regulation 831/2002 at Eurostat

For Eurostat, the implications of the Regulation and putting it into practice are considerable. But there are precedents and experiences to build on. For example, the European Community Household Panel survey (ECHP) has already paved the way—initially by providing some controlled access to confidential microdata and, more recently, by creating and making available anonymised micro-datasets. Similar approaches are being developed and extended—to the other surveys mentioned in the regulation and to a wider range of researchers.

New procedures are being developed for receiving research requests, evaluating the researchers and their requests, and for setting up contracts. Procedures for consulting the national statistical authorities of the MSs, as required by the regulation, are being developed. New contracts have been developed and "confidentiality undertakings" have been drafted. The contracts will be between Eurostat and the researcher's institution or organisation. This means that there must also be a contractual relationship between the researcher and his or her organisation. The regulation does not permit access to confidential data by individuals as individuals.

At the end of the day, the facilities to be provided under Regulation 831/2002 have to be user-friendly and have to provide a service to the research community. Eurostat sees consultation with the research community on their requirements, in terms of both data and facilities, as very important. Equally, Eurostat must explain the constraints to the research community and attempt to develop both appreciation

and acceptance of them. Close interaction with the research community, to understand its needs and interests and to explain the constraints, is a relatively new activity for Eurostat. However, recently contacts with CEIES, ESF (European Science Foundation), and other international research bodies took place; the dialogue started.

But this is not entirely new territory. Researchers' expectations and needs have been referred to above. There are examples in several MSs and elsewhere of facilities available to researchers. The Luxembourg Income Study provides an example, close to home, of internationally comparable datasets with remote access by recognised researchers. Some MSs, for example the United Kingdom, have lengthy experience of developing anonymised micro-datasets for research use by academics and research institutions. In the United States access to confidential data is provided through Research Data Centres of the Census Bureau. But this kind of access is not common to all countries—there are differences in practice, expectations, culture and legal frameworks.

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 Committee on Statistical Confidentiality if they wish to make requests to access confidential data for scientific purposes. "Other bodies" are those specified in article 3. 1(c) of the regulation. Essentially, these bodies are organisations that do not fall under the categories of 1(a) and 1(b) above and which have not been commissioned by departments of the Commission or of the administrations of the Member States to undertake specific research.

Regulation 831/2002 does not itself state criteria that should be taken into account by the Committee in forming its opinion. But there are some requirements in the Regulation and in Regulation 322/97 which indicate factors for consideration. Specifically, these are:

- prevention of non-statistical use (Regulation 322/97 arts.10 and 18 and Regulation 831/2002 art. 8 (1));
- access for scientific purposes (Regulation 322/97 art.17 and Regulation 831/2002 art.1); and
- protection of the data (Regulation 831/2002 art. 8 (1)).

In addition, the principles of transparency and fairness mean that criteria should be clear and known.

The Committee on Statistical Confidentiality decided that the following factors should be taken into account when forming its opinion:

- the primary purpose of the organisation;
- the organisational arrangements for research in the organisation;
- the safeguards in place in the organisation;
- the arrangements for dissemination of results of research.

Eurostat is now translating these conditions and factors into operational procedures. For example, the prior question of "admissibility" of an organisation to have the standing to make a request (regardless of the merits of the research request itself) has been specified in a series of questions (a questionnaire) covering:

- Identification and primary purpose of the organisation
- Brief description of the research project(s)
- Organisational and financial arrangements for research within the organisation

- Security in place in the organisation
- Arrangements for dissemination of results of research

This information will be passed to the national statistical authority of each NSI for it to express an opinion. This will probably be done usually through a written procedure in order to make the process reasonable fast.

The regulation provides for access by researchers to confidential data on the premises of Eurostat. There is also provision for similar access on the premises of national statistical authorities of the MSs if the level of the security and checking facilities are the same as those at Eurostat. Access of this type is often referred to as controlled access or access through a "Safe Centre".

4. Implications for Member States and national statistical institutions

The Regulation encourages the NSIs of MSs and Eurostat to work closely together in developing a system for providing access to confidential data for scientific purposes. This is a very wide-ranging set of activities—from agreeing ways of checking and protecting the outputs of research; agreeing on safeguards and controls for the data and ways of creating anonymised micro-datasets, to procedures for handling research requests and consulting each other. These processes are currently being designed and will be discussed with the MSs. The safeguards, controls and methods will build on existing approaches and methods. These will reflect existing national practices, but may require some adaptation. For example, one MS has an established procedure for considering research requests a few times a year. Yet the regulation requires that each MS must respond to a notification of a research access request within six weeks. Again, one MS has an established process for approving access requests by researchers and institutions of that country. But the regulation allows access by researchers—not only of other MSs, but also by researchers and organisations outside the EU.

Although there is a requirement that each MS be informed of each research request, there is a presumption in the Regulation that MSs will agree to give access to their data, provided that all the conditions and requirements specified have been met by the researchers.

There may be implications for NSIs in the way data are collected. In particular, if the uses to which the data may be put have to be specified to the respondent, then the research usage envisaged under the Regulation may have to be included. This is discussed further below.

Procedures for anonymising data and for protecting outputs from direct access to confidential data must also be developed by each NSI and agreed with Eurostat. In practice, a common approach by all NSIs will provide better protection and more useful datasets. There are some areas that will require further research and consideration as they are little developed or understood at present. These include problems of the disclosure potential of results from modelling. We understand well the problems of disclosiveness in tabular data—and methodology for this exists and is also still developing—but we have a less clear idea of the problems, let alone the solutions, arising from modelling. An intuitive restriction is to suppress information about residuals—even though they are of great statistical interest to researchers—because they give information about outliers which are often the rare data subjects. But we need also to know about the disclosive potential of parameter estimates—particularly when a series of similar models are run and compared.

5. Implications of Reg. 831/2002 for the Research Community

The implications of the new Regulation for the research community illustrate the nature of the partnership between statistical organisations and the research community. On the one hand the Regulation opens up new opportunities and on the other hand it imposes tight disciplines and limitations as the price of the opportunities.

First the research community must accept that they have no right of access. Then, researchers will have to accept that they will have a responsibility to maintain and uphold the confidentiality of data they

access. The limitations and safeguards may be more restrictive than those prevailing in the researchers' universities and those they have come across with other datasets, but they must be adhered to. The documentation of the Research Data Centres (RDC) of the US Census Bureau is voluminous but thorough. In particular, the sections on the different cultures of the RDC and universities make interesting reading. They also provide a warning that there should be no presumption of a common culture or purpose. Researchers will also have to accept that yet another body will have the right to ask detailed questions—not only about the research and its purposes, but also, in the case of anonymised micro-datasets, about how the data will be held and access controlled. And that the researcher's responses will be passed to the NSI of each MS for consideration. In addition, following access to confidential datasets, prospective results must be provided for checking before publication or other release.

In return, most researchers will have simpler access to datasets spanning the MSs. Hitherto, under the provisions of the Statistical Law, gaining access to data for each of the MS has involved a lengthy process of making requests to each MS. This will give researchers opportunities for pan-European Union research and analyses. The Regulation covers four important datasets: it is expected that, in time, access to other datasets will also be provided.

6. Implications of Reg. 831/2002 for data subjects

Although the purpose of the Regulation 831/2002 is to improve access to data for researchers, there are implications for the data subjects who provided the original information. This information was given to statistical organisations in their own countries, as part of a voluntary or compulsory statistical enquiry. Or the information may have been taken from existing administrative registers as part of a statistical enquiry. In turn, the statistical organisation passed the data to Eurostat after removing information allowing direct identification of the data subjects. In this connection there are also additional implications for those statistical organisations. The principle underlying statistical data collection is that of *informed consent.* The principle is that the data subject has a right to know what the information will be used for and who will see their information. The argument here is that if there is a new dimension—new users, new uses—to the use of the information, then the data subject should be made aware of it. In some MSs it may be necessary to change the laws under which data are collected in order to specify the uses to which the data can now be put.

As part of the statistical enquiry the data subject should be informed that the information provided will be used for statistical purposes, and that this may include research undertaken by external researchers in addition to the routine direct purpose of the statistical organisation. Under the Regulation, researchers may be from institutions within the Member States, or indeed from institutions outside the EU, not just from institutions within their own country. At present practices vary in the Member States in this regard, so it is not easy to indicate what will have to change. And in practice little may need to change—the existing forms of consent may well cover, implicitly, access by researchers from another country for statistical research.

It is a question of degree, balancing along the implicit–explicit axis with the "informed" aspect of the consent. This may require some field research, including qualitative research, among data subjects. It is an important part of the contract between the data subject and the statistical organisation and will be seen by the latter as a factor affecting response rates to voluntary enquiries.

7. Some questions arising

What do we mean by "scientific purposes"? This is a question that has already arisen. For some this is synonymous with "academic". But even so, what is to be included? Recognised, post-doctoral researchers are presumably undertaking scientific research (even if some of it may also be "commercial"). But below this apparently clear-cut category distinctions are more difficult to make. And the focus of the debate tends to centre on the qualifications or status of the researcher rather than on the actual "scientific" nature of the research proposed. What, then, of Ph.D students? Much of the work

undertaken by doctoral students is at the forefront of scientific knowledge, so is presumably scientific. And much research undertaken for a Masters degree will be supervised by a recognised researcher/scientist and may form part of a larger project with a clear scientific purpose. Should undergraduates have access "for scientific purposes" for projects and for familiarisation with large complex datasets? After all, training may be regarded as a scientific purpose. The same training argument could be made at higher levels. It may be difficult to draw this line. Pragmatically, the line may well be drawn on legal rather than scientific grounds—does the person desiring access have a contractual relationship with the institution, so that penalties for non-compliance with conditions for access can be invoked.

Remote access. For some, the concept of a physical "safe centre" has already been overtaken by events—new technology. Attending a physical "safe centre" for access has several drawbacks: cost, ease of access (Luxembourg is neither the easiest or cheapest place to visit), probable time lapse between running programs and receiving results, lack of spontaneity in performing analyses, convenience of access. For this reason many researchers seem willing to trade access to "real" confidential data through "safe centres" for anonymised micro-datasets—for the convenience of access on their desks.
But there have been several successful "remote access" facilities to confidential data. The leading example of this was the Luxembourg Income Study (LIS). Using the approach and software designed for that, another research consortium—the pay inequalities and productivity (PiEP) project—has developed procedures for remote access to the Structure of Earnings Survey data at Eurostat. In both cases there are trade-offs in order to obtain access. The LIS uses a highly reduced dataset of relatively few key variables, some of which are reduced to categorical variables. The PiEP accepts some reductions in the dataset and some restrictions on outputs—no tabulations, no information on residuals, and so on. These restrictions are designed, with the agreement of the NSIs of the MS, to reduce to an acceptable minimum the risk of identification or of disclosure.

The challenge is to provide remote access to researchers. A non-technological issue is whether the regulation can be interpreted as permitting this type of access. Under this type of arrangement the processing and analysis of the data would be performed on the premises of Eurostat. The controls—on individuals, authorised access and on outputs—that would be used in the case of a traditional "safe centre" could be the same. But access would be much easier. However, this issue has to be further investigated before putting it on the agenda of the Committee on Statistical Confidentiality.

Value of the research to the data provider. The US Census Bureau has an explicit pre-requisite that the research proposed should be of value to the Bureau—indeed, the research *must* (a legal requirement) provide a benefit to Census Bureau programs. The draft Protocol being discussed in the United Kingdom includes similar wording, but the legal basis for this is not clear. Regulation 831/2002 has no such requirement.

Are anonymised data confidential data? If the anonymisation process reduces to a minimum the risk of identification or of disclosure, is the anonymised dataset "confidential"? This question seems to bring together, for a healthy debate, the classificatory legal approach and the pragmatic statistical approach. If direct identification is not possible and the risk of indirect identification is negligible (minimised in accordance with current best practice)—anonymised data—then the data are not disclosive (or potentially disclosive) and so not confidential. Or are such data always confidential no matter how much they have been modified?

8. The future

One aspect of the legal requirement on NSIs and Eurostat that needs further consideration is indicated in article 13 (2) of the Statistical Law. This states that, "By derogation from paragraph 1, data taken from sources which are available to the public and remain available to the public at the national authorities according to national legislation, shall not be considered confidential." This indicates a conundrum that pervades statistical confidentiality: information obtained by an NSI through a statistical enquiry is treated as confidential even if the information is publicly available and even if the data subject itself proclaims the information. In some countries there is a let-out for the NSI—if the data subject releases the NSI

from the confidentiality requirement. But without this, much effort goes into protecting (mainly economic) statistical data that is publicly available. Perhaps the answer is for statistical enquiries to be in two parts—the publicly, and often statutorily, available information about a company; and the information to be protected as confidential.

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. For example, "anonymised microdata" are defined in regulation 831/2002 in terms of "…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". Clearly, current best practice changes over time, so must also the procedures used. Again, the Statistical Law requires that "account shall be taken of all the *means that might reasonably be used* by a third party to identify the said statistical unit". The means available to third parties will also change over time—greater access to other databases and more powerful computers and software. A further example is provided by Regulation 1588/90. Eurostat must offer the same confidentiality guarantees as the NSIs for the transfer of data to Eurostat from MSs. Developments in the MSs in this regard will need to be reflected in Eurostat's procedures.

9. Conclusions

Developing the draft regulation and getting it approved by MSs through the Committee on Statistical Confidentiality and by the Commission entailed considerable effort by many people. But that approval is only part of a larger process of implementation and creating the processes and facilities, to say nothing of the datasets themselves, in order to provide the research access to confidential data. The process of implementation has raised further questions, both statistical and legal, that need consideration.