UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

Statistiska centralbyrån
Statistics Sweden

STATISTICAL CONFIDENTIALITY AND ACCESS TO MICRODATA

Proceedings of the Seminar Session of the 2003 Conference of European Statisticians

UNITED NATIONS

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

Statistiska centralbyrån
Statistics Sweden

STATISTICAL CONFIDENTIALITY AND ACCESS TO MICRODATA

Proceedings of the Seminar Session of the 2003 Conference of European Statisticians

UNITED NATIONS

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

Statistiska centralbyrån
Statistics Sweden

STATISTICAL CONFIDENTIALITY AND ACCESS TO MICRODATA

Proceedings of the Seminar Session of the 2003 Conference of European Statisticians

UNITED NATIONS
NEW YORK AND GENEVA 2003

Note

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations.

FOREWORD

At the 51st plenary session of the Conference of European Statisticians (CES) held in Geneva, 10-12 June 2003, one of the two seminars was devoted to the issue of statistical confidentiality and access to microdata. This seminar was organized and chaired by Statistics Sweden in cooperation with the CES Bureau. Since the seminar was regarded as very fruitful, the United Nations Economic Commission for Europe (UNECE) and Statistics Sweden decided to make a joint publication of the proceedings of the seminar.

The Fundamental Principles of Official Statistics, adopted by the Economic Commission for Europe in 1992, include a principle, according to which "Official statistics provide an indispensable element in the information system of a democratic society…". Furthermore "…individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes". When discussing microdata, the main challenge for statistical offices is to ensure that the improved access to microdata will not undermine the confidentiality principle, both in reality and in the perception of the public and respondents to statistical surveys.

The introduction of this publication includes a brief summary of the discussions that took place at the seminar. Large differences between countries can be seen both regarding legal aspects and access to microdata, which become obvious when studying the publication. The need for further collaboration and sharing information was emphasized at the seminar. The Conference agreed to create a Task Force to develop a set of guidelines for dissemination of microdata and confidentiality protection.

Heinrich Brüngger
Chief, Statistical Division
UN Economic Commission for Europe

Svante Öberg
Director General
Statistics Sweden

Table of contents

INTRODUCTION

Confidentiality is one of the Fundamental Principles of Official Statistics. It is a top priority issue on the policy agenda of statistical offices and an indispensable element to maintaining the trust of respondents and thus ensuring the quality of data. The Bureau of the Conference of European Statisticians (CES) recognised the need to discuss confidentiality problems in statistical practice at the highest level and chose confidentiality and access to microdata as the topic of a special seminar of the 2003 plenary session of the CES.

The present publication provides all the papers, both invited and supporting, that were considered at the Seminar. It follows the programme of the Seminar, concentrating on the following four themes: (1) overview and use of microdata, (2) data confidentiality, (3) legal aspects of microdata, and (4) access to microdata. Each topic begins with the discussants' comments, which provide a good introduction to the issues considered. Current problems in confidentiality protection are analysed and some steps for future international cooperation in this area are identified. Special attention is paid to confidentiality problems in Central and Eastern European and the CIS countries.

Chapter I deals with confidentiality issues and access to microdata in general. The paper by Statistics Sweden (Matti Niva, Bo Sundgren and Ingrid Lyberg) provides a good overview of the main issues. Julia Lane, in her keynote speech, summarizes the benefits and risks associated with microdata access. Statistical offices experience increasing pressure from scientists and governments to provide access to detailed data. The wealth of available data would be invaluable for research, policy making and monitoring, local planning, etc. If statistical offices do not address these issues, researchers and governments may look for alternate data providers, which is a waste of public resources and will (probably) result in a lower quality of data. There are benefits also on the statistical offices' side: bringing in the needs of the research can help to improve the surveys, academic use of data can increase the prestige of statistical work and attract more highly qualified staff.

However, there are high costs and risks associated with microdata access. Setting up the necessary conditions and contracts, keeping the technical tools up-to-date, monitoring the conformance to confidentiality principles, etc. is a demanding exercise. These costs have to be fully born by the statistical offices while most of the benefits accrued arise from analysis undertaken outside the statistical agencies. Therefore, statistical offices often allow research access to microdata only on the condition that it provides a benefit to the statistical agency's programs.

Chapter II covers specific issues of confidentiality protection in Central and Eastern European and the CIS countries; these are considered in the papers by the UNECE secretariat, the Russian Federation (Vladimir L. Sokolin), Georgia (Teimuraz A. Beridze), Ukraine (Olexander

Osaulenko), Poland (Tadeusz Toczynski), Kyrgyzstan (Zarylbek Kudabaev and Nataliya Gudkova), and Lithuania (Sigitas Biciunas).

The UNECE paper presents the results of a survey carried out in the transition economies in January 2003. The survey shows that the main concerns of transition economies with regard to data confidentiality are: methods of access to microdata, legal implementation of confidentiality protection, methodological and technical standards, issues related to administrative registers, and improving respondents' perception of confidentiality protection.

The Russian paper elaborates on specific confidentiality problems in Russia which also reflect the situation in other transition economies. In many of these countries, there is still pressure from other government bodies to gain access to microdata. Emerging from a past where Official Statistics followed a completely different paradigm, the role of Official Statistics is often not fully understood by other government agencies. Raising awareness and training are needed to explain the confidentiality principles to respondents, users and the staff of statistical offices. Unresolved technical aspects, lack of special software and the low level of development of technological equipment sometimes make it difficult to ensure data protection at the required level. However, the statistical offices of these countries are committed to resolving the problems of confidentiality protection.

In Chapter III, several papers concentrate on the legal aspects of confidentiality. Eurostat (John King) presents the implementation of recent EU legislation for research access to confidential data (European Commission Regulation 831/2002). Access is controlled by strict rules and a contract should be made in each case, spelling out the necessary conditions and safeguards for confidentiality protection. The paper by Statistics Sweden (Birgitta Pettersson) gives an overview of the confidentiality legislation in the Nordic countries.

A basic principle in statistical legal frameworks is that data collected for statistical purposes may only be used for the production of statistics. A distinction is made between statistical and non-statistical use of data. Non-statistical use (scientific, historical, public planning) is regulated with specific rules and contracts, and must not be in contradiction with the purpose for which the data was collected. In the case of government ministries, the risk of non-statistical use of data is high. There can also be pressure to make specific exclusions from the confidentiality principle because of threats to national security.

Statistical laws change slowly but the national legal context is complex and can change quite frequently. Differences in various legislative acts might be a source for conflicting rules and obligations concerning the protection of confidentiality. This problem is more acute in the transition economies but it applies also to more developed economies. The legal framework provides the basis for confidentiality protection but is not sufficient in itself. An institutional body needs to be set up to take the decisions on how to implement the regulations, and to organise their implementation.

Chapter IV deals with various issues arising from access to microdata. The Australian Bureau of Statistics' paper (Dennis Trewin) provides an overview of the main methods of access to microdata: Public Use Microdata Files, on-site research centres and on-line access. It is becoming increasingly difficult to provide truly "safe data" using automatic procedures, so it is inevitable to rely more on "safe settings", including legal and administrative arrangements. A move from the paradigm of risk avoidance to risk management can be observed. The papers from Denmark (Lars Thygesen), Germany (Tom Wende and Markus Zwick) and the United States (Gerald Gates et al.) elaborate on issues, organization and approaches of access to microdata.

So far confidentiality protection has been mainly a national issue. However, in the context of EU and the increasing dissemination of data over Internet, it is also becoming an international one. There is a lot of international collaboration among the research community, and researchers can be very critical towards different access rules in different countries. Frequently, researchers are not allowed to access other countries' microdata because of the fear that confidentiality protection cannot be guaranteed. This raises the need to unify approaches internationally, and to agree on some core principles for the dissemination of microdata.

The Conference agreed that generic guidelines on confidentiality would be valuable for countries that are in the process of setting up the legal background and procedures for access to microdata, and would help in discussions with the user community. The principles should be general enough to be applicable in different countries and it is desirable to involve countries with different confidentiality practices and different levels of development in their preparation. In order to develop a set of guidelines for confidentiality protection, it was decided to create a Task Force chaired by Mr. Dennis Trewin (Australia). The Task Force will cooperate with the Working Group on Confidentiality (led by Eurostat and UNECE) that is dealing with the more technical issues of confidentiality.

Large differences between countries can be seen in the use of confidentiality measures, due to the different levels of development and due to varied national practices that have been formed over time. Regardless of the differences, the same problems are often encountered. Therefore, all countries stand to gain through collaboration and the sharing of information. For this purpose, some countries' guidelines and regulations on confidentiality can be found on the UNSD website on best practices in Official Statistics (http://unstats.un.org/unsd/goodprac/default.asp). Also available are the materials from several international conferences that have dealt with more methodological and technical issues of statistical confidentiality. Among these can be mentioned the joint UNECE/Eurostat Work Sessions, with emphasis on the application of disclosure control in statistical practice. For interested readers, documents from the UNECE/Eurostat joint meetings are available on the website: http://www.unece.org/stats/archive/docs.subject.e.htm (under programme element 2.1 Management of information technology infrastructure).

In the discussions at the CES plenary session, several important questions were raised in connection with increased access to microdata. For example, how to define scientific purposes? How to preserve the principle of equal access to data, one of the Fundamental Principles of Official

Statistics, while giving restricted access to microdata? How to inform the respondents about a potential new use of their data that was not foreseen when the data were collected? Should the data still be considered confidential when they are anonymised and modified in such a way that disclosure is virtually impossible? Are there differences in protecting the confidentiality of data on legal and physical persons? How to treat individual data that are publicly available from other sources? How do respondents perceive confidentiality? How much influence does the statistical offices' pledge of confidentiality protection have on response rates and data quality?

The discussion provided some answers, but a debate on these and many other confidentiality related questions will certainly continue. Hopefully, this publication will give statistical offices ideas for solving the confidentiality problems allowing increased access to microdata and will provide a basis for future discussion by building on the experience acquired.

CHAPTER I: OVERVIEW AND USE OF MICRODATA

I.1 SUMMARY OF DISCUSSANTS' MAIN POINTS

by Len Cook, Office for National Statistics, United Kingdom

Introduction

We cannot now doubt the weight of the argument that inadequate use of microdata has high costs. This may be because we now see as everyday the expectation that microdata analyses will usually inform policy analysis, and policy evaluation. Perhaps official statisticians have always been too cautious. Whatever the reason, the analysis of microdata is now an expected capability for statistical researchers in many fields, particularly economics, geography and social research.

The cost in failing to make full use of microdata is most significantly seen in inadequate identification and analysis of policy options. Benefits that may be perceived more readily by official statisticians from effective microdata access to official statistical survey records include a reduction in what, in many countries, is a clearly inferior survey operation when conducted outside of the official statistical environment.

Enabling research on microdata is a strong part of any official statistics system. In the USA and the UK, this is a long-standing distinguishing characteristic of the statistical system. In both countries, a high level of microdata access has enabled a well-informed user community to extend the analytical strength of available data and has often led advances in the organization, management and design of statistical sources. The ESRC data archive in the UK, established some 30 years ago, has, through custom and practice, created a powerful experience in trust and value in analysis. There is also EDINA, established in 1996, with 6,000 registered users in 200 institutions. It hosts the digitised boundary data for the 1991 census, and is an excellent resource for GIS work. Further work is done by the Cathie Marsh Centre for Census and Social Research in Manchester.

Benefits from microdata access

In her address to the 2003 plenary session of the Conference of European Statisticians, Julia Lane argues that furthering of microdata access would build trust in Official Statistics, through the following means.

- By allowing better research into more complex questions, including marginal effects: the academic community has a range and depth of expertise that would be very difficult, perhaps impossible, to replicate within the official statistics community. Without access to microdata that allows complex questions to be answered, the UK would be in a situation where subject experts in the wider research community cannot obtain the relevant data. Research expertise builds richness on top of the statistical statements of official statisticians. If the public see the outputs of academic research as an enhancement of official statistics, and not as something different or separate, then trust in official statistics is likely to be enhanced. Much microdata-based research draws on common research models and are not 'hit and miss' studies. These common models need to be effectively specified, so that they become reflected in the survey design for statistical surveys.

- Through scientific safeguard: both official statisticians and the academic researchers are part of the same scientific community, and therefore both parties should welcome replication of outputs. Often, meaningful replication requires microdata access.

- Through feedback about quality: academic research can stretch microdata to its limits of fitness for purpose. Near these limits, a lot can be learned about quality. If a mechanism was in place to report quality issues, then the processes of instrument design, collection, capture, editing, imputation, etc. could be improved - thereby improving quality of official statistics, and trust in them.

- Through the development of a core constituency: in the UK, we may need to better communicate exactly what microdata studies achieve. A good example is Angela Dale's explanation of what the SARs from the 1991 population census in the UK have achieved.

Risks in microdata access

Where there are risks in microdata access, the extensive experience in the UK and the USA would suggest that the damage to public confidence has been less than that caused by several other issues.

- Health Ethics Committees, whose cautionary excesses now are at risk of truncating sound and proper statistical surveys, through seeking to constrain statistical practices as though they had the same effect as medical experiments (Alder Hey, Hepatitis, have increased the pressure to protect confidentiality through not obtaining information in the first instance).

- Problems in maintaining access to an increasing number of "Hard to reach" households.

- Concentrations of 'alienated' communities, who simply avoid wanting to be included in the statistical measures on which we base decisions that affect them.

- An increasing general survey burden on households.

Broadening the statistical system through microdata access

- A great variety of partly connected institutions engage in publicly credible research.
- Strong community research institutions continue to be founded on analysis of data.
- 'Evidence-based policy' initiative of government has foundations in and beyond the government research community.
- Sound ethical base for protecting confidentiality exists, maintaining well-founded practices that protect obligations to users.
- Influences the public climate and debate about information access, by valuing and advancing access to information for research and public service, compared to finding fraud and illegal activity, from specific identification. We do need a strengthened legal base for this, an issue which ONS and the Statistics Commission are looking into further.
- Data Protection legislation is strongly valued and plays a huge part in cross agency function.

Increasing trust through microdata access, by managing disclosure risks

Any form of disclosure, particularly through a matched dataset, would disproportionately damage all this good work. In the end, it is the producers of official statistics who carry this risk, and it is not really possible to transfer it with the data. Therefore, the question remains as to how producers of Official Statistics manage this risk - by process ('contracts' of trust with users of microdata) or by design (continuing use of statistical disclosure control techniques).

A future microdata access strategy might advance in several directions:

- seek to expand options for access, recognising that even more datasets are disproportionately more value as unit record data sets (time use, longitudinal studies), in-house custody or externally run analysis;
- National Statistical Offices should create a place where record matching can take place that would be only accepted for statistical purposes, so we can create further information/research resources;
- in principle, we may need to distinguish between political acceptability of any sort of matching, and the technical and statistical dimensions, and manage these differently;
- find ways of facilitating greater access to identifiable unit records, particularly longitudinal business studies;
- recognise degrees of trust needed for different forms of access.

Legal / international framework

Some issues that arise from the papers we have for discussion include the following.

- A condition for the use of statistical microdata records might be that there is no incomparability between the purpose of the research, and the purpose to which the data was collected. (In UK data protection legislation, this is expressed as statistical purposes being 'not incompatible' with the original purpose.)
- The balance between Freedom of Information (Press), Legislation and Data Protection legislation poses real challenges (Swedish experience).
- Need to also consider the local impacts of overseas legislation and EU legislation (US Safe Harbour Principle).
- We are increasingly likely to need to focus on trust in researchers, and strength of obligation and capacity to enforce. This will not be consistent with any obligation for equal treatment of users.
- National Statistical Offices need a continued evaluation of practices, given new domestic and international experience.
- The US and the UK have had very long experience with releasing microdata on a large scale. In both these countries, as with those that have less active traditions of microdata release, user expectations are also for more access.
- There is an international community of microdata users, who share knowledge of country practices and developments perhaps even more effectively than do statistical offices.

- introducing change to microdata access practices (in any nation) involves assessment of cost, scale, traditions, laws and cultures.

Questions for the future

(i) Equity of access is critical to official statistics. How easily can this principle of equity of access be applied to the results of research?
(ii) The phrase 'for statistical purposes' is well defined. How can we give clarity of a similar strength to the phrase 'for research purposes'?
(iii) Access to unit records involves a matter of trust. Trustworthiness is not equally distributed across the population! When trustworthiness is made a condition of access, it is no longer possible to have equity of access for research. How do we resolve this?
(iv) Some projects, such as the matching of administrative data sets, are only done in National Statistical Institutes. How can the National Statistical Institutes better capture the reverse benefit of this? How can the credit for this work be recognised in funding and feedback for suppliers of the data?
(v) It is important to remember that National Statistical Institutes do not just compile datasets, but report and analyse the progress and trends of society. Given this reporting role, how far can synergies be expected between Official Statistics and research projects?
(vi) The timing and form of statistical releases are managed carefully in order to ensure the impartiality of the release process. Do researchers who reinforce the credibility of their work by making use of official statistics require similar obligations when releasing their results?
(vii) National Statistical Institutes have considerable power to protect the confidentiality of statistical records. The National Statistician has responsibility for exercising that power, and maintaining the level of protection when it is challenged (through legal and other means). When researchers have custody of statistical unit records, how far do we need to give them the capacity to meet the same challenges that National Statistical Institutes can face, if harassed by bodies with some legal authority?

I.2 USES OF MICRODATA: KEYNOTE SPEECH

by Julia Lane, The Urban Institute, United States

Introduction

The mission of national statistical institutes is to collect and disseminate data. Decades ago, this meant producing books and reports primarily consisting of tabular data – designed to answer pre-defined questions. The increasing complexity of 21st century society, however, has put increasing pressure on such institutes to produce microdata – designed to allow policy analysts and researchers to pose and answer questions of their own choosing. This pressure creates both opportunity and challenge. On the one hand, the relevance and stature of statistical agencies can be enhanced by their dissemination of data that policy makers can use to answer complex questions quickly. On the other hand, the well-known confidentiality challenges to the creation of public use files and other access modalities have been exacerbated by the development of new types of microdata, as well as substantial computing and technological advances.

Finding creative ways to address the fundamental tension between data dissemination and the protection of respondent confidentiality goes to the core of each statistical institute's mission. Failure to do so has tremendous costs to society. An example might serve to illustrate the point. I have worked with the World Bank on and off for over a decade, in a number of less developed countries. One common characteristic of the statistical institutes of the countries in which I worked was a reluctance to provide access to microdata – and in every case, this led to incomplete analysis and wasted resources in countries that could afford them least. In one case, the country in question was concerned about the low labour force participation rate of women – which had hampered development for over a decade. Several policy options were on the table – including providing free child care, flexible work-weeks, and subsidized education. However, no microdata analysis had been undertaken. Although labour force surveys were regularly fielded, they were not even released to the Ministry of Human Resources or the Ministry of Education. We analysed the microdata and found that, even after controlling for education, industry and occupation, women were paid 60% less than men – and had been for the ten years in question. Our conclusion, which would have been apparent to any analyst working with these data, was that the country in question would have been better served by investigating the sources of these earnings differentials, rather than investing in the expensive set of options initially identified. Had the country in question permitted broader access to the microdata a decade earlier, the appropriate policies could have been in place much earlier.

That microdata are important is not news to any of you. Indeed, Eurostat has recently issued a new regulation (831/2002) to codify access to confidential data[1]. What I would like to discuss is how can statistical agencies determine the "optimal" amount of microdata to release – and find creative ways to increase this optimum? As an economist, my answer is that an accurate assessment depends on the benefits derived from the use of such data, the costs, and the trade-off

[1] See Jean-Louis Mercy and John King's paper "Developments at Eurostat for Research Access to Confidential Data", Joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg (Luxembourg, 7-9 April 2003) Working Paper 12.

between the two. My goal in this paper is to attempt to explicitly delineate these benefits and costs, identify changes and summarize the consequences and opportunities for statistical agencies.

The benefits of microdata use – and why they are increasing

The benefits associated with microdata access are myriad. The most obvious is that microdata permit policy-makers to pose and answer complex questions, but others are also apparent. Access to microdata permits analysts to calculate marginal, rather than average effects; it acts as an important scientific safeguard, because it permits others to replicate important findings; it creates a virtuous cycle of knowledge for the statistical institute because data use inevitably reveals data quality and processing anomalies as well as new data needs; and finally, it creates a core constituency for the statistical agency itself. I will illustrate each of these points with an example.

a) Microdata permit analysis of complex questions

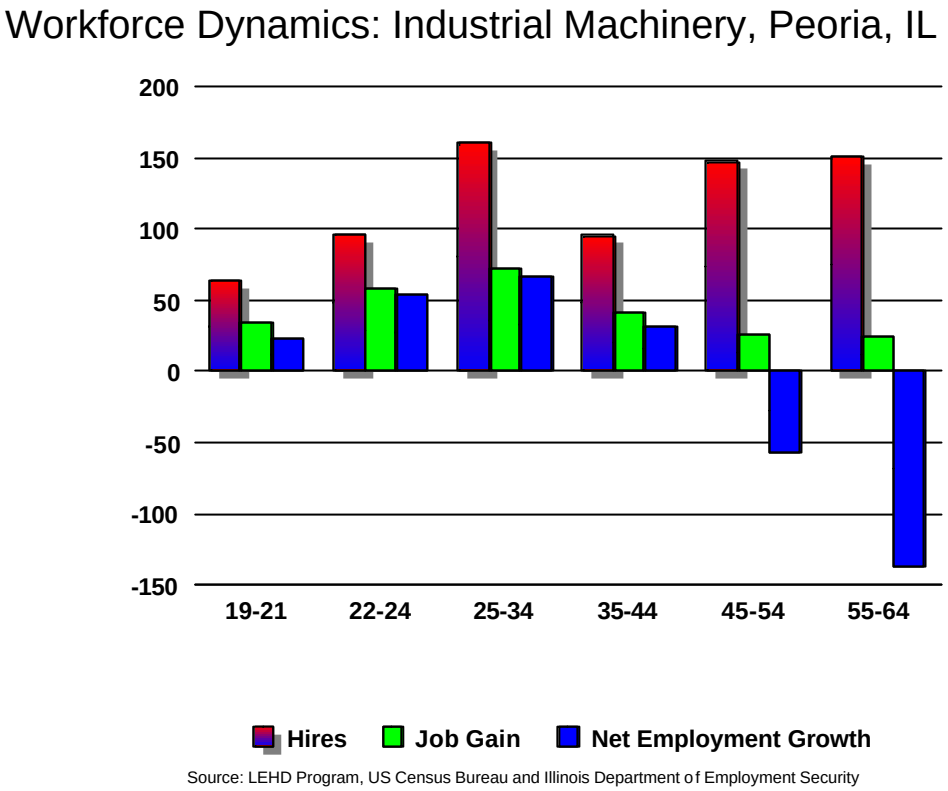
One of the most important findings in economics over the past decade has been that the analysis of aggregate statistics does not give policy makers an accurate view of the functioning of the economy. Indeed, the creative turbulence which is the hallmark of the United States economy, and a major contributor to its success, is not apparent from macro level indicators. Analysis of microdata suggests that the widespread reallocation of factors of production from one firm to another firm even within narrowly defined industries is a major contributor to U.S. productivity growth – more important than investment in equipment and structures[2].

As an example of the importance of this phenomenon, policy makers in Illinois asked me to examine employment changes in a detailed industry – industrial machinery – in a detailed geographic area - Peoria, IL. Aggregate statistics indicated that this industry had lost a total of 20 jobs in the previous year. An analysis of the microdata, summarized in Figure 1, revealed a very different picture. The net employment loss of 20 jobs was the sum of positive employment gains for workers 44 and under, and employment losses for workers 45 and older. In fact, in net, about 160 jobs were reallocated from older to younger workers. The microdata revealed even more reallocation than this. If we simply tabulate up the job gains from expanding and new firms, there were over 250 jobs gained for workers of all ages (including older workers). The gross job reallocation, achieved by summing up 250 jobs gained and 270 jobs lost, exceeds 520 jobs. The worker flows are greater yet. Over the same period, over 710 workers were hired and 730 separated – for a total of 1400 workers reallocated.

The importance of knowing that even quite small net job changes can represent enormous job and worker reallocation is non-trivial information for policy-makers so that the productive potential of this reallocation process can be realized to its fullest. In this case, for example, the analysis showed Illinois policy makers that the aging of the industrial machinery workforce would lead to a demand for trained workers to replace oncoming retirements.

[2] Foster, Lucia, John Haltiwanger and C.J. Krizan (2001). "Aggregate Productivity Growth: Lessons from Microeconomic Evidence." *New Directions in Productivity Analysis,* (eds. Edward Dean, Michael Harper and Charles Hulten), University of Chicago Press, (forthcoming).

Figure 1



Source: LEHD Program, US Census Bureau and Illinois Department of Employment Security

The new challenge that this increasing value of microdata poses to statistical agencies is that the microdata sets that permit such in-depth understandings of the economy – which involve the longitudinal linkage of firm and worker data over time – are also very large and complex, and often involve the integration of administrative and survey records. External researcher access is often the only way to create such data – because many of the decisions require subject matter knowledge as well as statistical expertise.

b) Calculating marginal rather than average effects

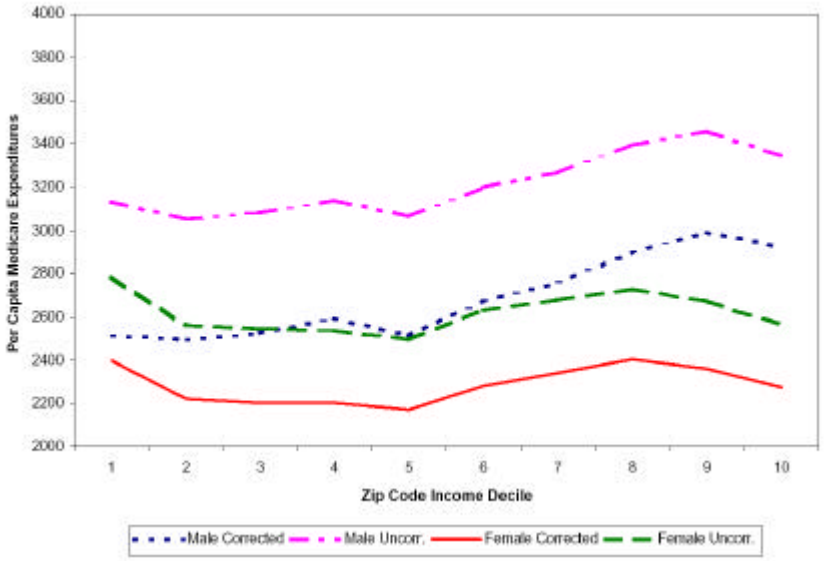
The ability to estimate marginal effects goes to the heart of the use of microdata. Microdata enable analysts to do multivariate regressions, whereby the marginal impact of key variables, controlling for other factors, can be isolated.

An excellent example is provided by a recent study[3] which investigates the distributional impact of Medicare. The importance of this healthcare program for the elderly population is difficult to overstate – it cost $220 billion (in 1998) and its costs are growing faster than Social Security. Understanding program use, and the correlation of this with income and health, is critical to understanding the effects of the program.

[3] Lee, McClellan and Skinner "The Distributional Effects of Medicare", MBER working paper 6910, January 1999.

The microdata reveal important facts about program use that, again, would not be available from an analysis of aggregate data. Program use is heavily skewed – a very small proportion of the elderly population account for a very large proportion of expenditures. Program use is very persistent: those who account for a high proportion of expenditures in one year are highly likely to be heavy users in subsequent and preceding years. Even more interesting, however, is the effect of examining the relationship between income and expenditures, which is described in Figure 2.

Figure 2 Income and Medicare Expenditure



Source: Lee, McClellan and Skinner, 1999

Briefly, it is clear from an examination of Figure 2 that the marginal effect of income on expenditures is broadly positive for men, but that the relationship is not only much flatter for women but women spend less. The marginal effect of correcting for health status (whether or not the individual died during the analysis year) at all income levels is also evident. Thus this analysis of the microdata provides a quantification of the marginal effects of the key contributing factors to health expenditures: sex, income and health.

This example controls only for demographic effects – yet the increasing complexity of economic activity requires the production of data that can be used to separate out not just complex demographic interactions, but also economic and spatial effects. The expansion of research on the human dimensions of environmental change has increasingly meant that researchers want to include the contextual variables surrounding an individual - the schools they go to, the neighbourhoods they live in, the firms they work for, and the people with whom they interact. As Rindfuss points out, "Linking data on people and their environments is at the very

core of IHDP"[4]. The imperative to identify marginal effects in such an environment will put tremendous pressure on statistical agencies.

c) Scientific safeguard

Access to microdata is critical to ensure that other scientists can replicate important research. This acts as an important discipline device for both government statisticians and academic researchers. That there is overwhelming temptation for scientists to misrepresent results is, sadly, evident from the all-too-frequent news stories of data fabrication. That there is similar pressure on statistical institutes should be taken as self-evident.

I will give one example. I used to teach a PhD level Applied Econometrics class when I was on the faculty at American University in Washington DC. Because it was offered at night, and the university was close to downtown, I had many students from international organizations who were brushing up their quantitative skills. I structured the class so that the first few weeks were spent on discussing techniques for dealing with "dirty" data – a problem of which they were only too well aware, and students would take turns regaling the class with first-hand anecdotes. One particularly popular example was from a gentleman who had worked in his country's population division, charged with providing annual population estimates. Apparently, another division first estimated GDP, and since per capita GDP was an important criterion for determining international financial aid, his main task was to make sure that the denominator was high enough to keep per capita GDP appropriately low. While external access to the underlying microdata might not be a panacea to cure cases like this, it certainly might increase the level of accountability – and reduce the amount of "dirty" data.

Constant vigilance in this area is important. When the gains to monopoly power over information are great, in terms of either political or professional prestige, it would be naïve to think that there were no malfeasance in even the most pristine of agencies. The consequences to the statistical system of such malfeasance can be devastating if unchecked.

d) Data quality

Although statistical institutes expend enormous resources in quality assurance to ensure that they produce the best quality product, there is no substitute for actual research use of microdata to identify data anomalies. Indeed, there is general recognition of the direct correlation between the quality of a national statistical institute and that institute's openness to external research in international agencies, such as the World Bank. The United States Internal Revenue Service (IRS) and the United States Census Bureau have actually formalized the role of researcher use of selected tax microdata to improve national statistics. Because this inter-agency agreement only permits the IRS to release selected microdata[5] to the Census Bureau in order to improve the economic and demographic censuses, surveys and inter-censal population estimates

[4] Ronald Rindfuss "Confidentiality Promises and Data Availability" in IHDP Update, 02/2002, Newsletter of the International Human Dimensions Programme on Global Environmental Change.

[5] As in many countries, selected tax data form the heart of the Census Business Register – the business sample frame – and play a critical role in developing inter-censal population estimates.

researchers who use Census Bureau's tax-derived microdata must document the benefits according to the following criteria:

- Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census or estimate;
- Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census or estimate;
- Enhancing the data collected in a Title 13, Chapter 5 survey or census. For example:
- Improving imputations for non-response;
- Developing links across time or entities for data gathered in censuses and surveys authorized by Title 13, Chapter 5.
- Identifying the limitations of, or improving, the underlying business register, household Master Address File, and industrial and geographical classification schemes used to collect the data;
- Identifying shortcomings of current data collection programs and/or documenting new data collection needs;
- Constructing, verifying, or improving the sampling frame for a census or survey authorized under Title 13, Chapter 5;
- Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;
- Developing a methodology for estimating non-response to a census or survey authorized under Title 13, Chapter 5;
- Developing statistical weights for a survey authorized under Title 13, Chapter 5.

A sterling example of how this can work is a new project between the Census Bureau and researchers at the Sloan Industry Centers. The Sloan Foundation has invested heavily in case study research of a number of industries, five of which (semi-conductors, software, retail trade, finance and trucking) are involved in this project. The Sloan researchers work directly with Census staff – and their rich industry specific knowledge should lead to contributions ranging from help with industry classification to identifying new survey questions that could hone in on the driving forces of change in their industry.

Statistical institutes operating in an environment where the blurring of firm and industry boundaries is accelerating, where the differentiation between place of work and place of residence is increasingly unclear, and where the engine of economic growth has changed from measurable machines and equipment to the much less measurable workforce quality will increasingly need to turn to external researchers for guidance.

e) Development of core constituency

The funding of a statistical agency depends on the development of a constituency and greater use of data – which includes the creation of new products from existing data - creates a constituency beyond that of those who access the data. More analysis, more publicity and more insights lead to a greater understanding of the value associated with products produced by the statistical institute – with associated funding benefits. I have been associated with at least one statistical agency that resolutely opposed any access to its microdata by anyone other than its

own staff. This resulted in extremely strained relations with other ministries and the development of pseudo-statistical agencies within those other ministries that developed and fielded their own surveys without appropriate sampling frames or survey development or statistical method expertise. Not only did this generate (in my opinion) bad data for decision making, but it also seriously threatened the long-term financial viability of the statistical agency. Specifically, the ministries directly competed with the national statistical institute for funding, and co-ordinated strong resistance to any funding increases for that institute.

The value of a core constituency goes beyond the (admittedly crass) funding aspect. The quality of staff that can be hired is directly correlated with the prestige and visibility of the institute, and the perceived quality of work that can be done within its walls. External researchers, who are often academics, also advise and counsel students about career opportunities. Cultivating this network is an important first step to developing a high quality staff – maintaining the dynamic interaction between staff and their mentors can create an ongoing virtuous cycle of information exchange and education.

The costs of microdata use – and how they are changing

One of the most boring things about economists is that they will tell you that nothing in life is free. I am no exception. The most obvious costs of microdata use include the cost of providing access, potential reputational costs and the costs associated with re-identification of the sampled entities and the concomitant potential disclosure of confidential attributes. These are the costs that must be weighed against the benefits of providing access.

a) The cost of providing access

Clearly the cost of providing access depends on the modality, and several have been developed by statistical institutes across the world – public use microdata, remote access sites, research data centres and licensing. The agencies' explicit costs for each of these methods are substantial in terms of staffing, support and documentation. The costs to users vary dramatically – public use data are clearly the lowest cost option, while the explicit and opportunity costs of accessing microdata research data centers are substantial.

The most important of these modalities – and the one subject to most change - is public use microdata. Statistical institutes have worked very hard to make these available, with dramatic success. It is not an overstatement to say that since such data were first created over 30 years ago, they have had a major impact on decision making. Indeed, decisions are often made in developing countries based on results from European and North American public use data sets. Funding decisions for some entire data collection activities are predicated on the existence of public use microdata. However, the cost and feasibility of producing high quality public use datasets is unfortunately increasing. A combination of technological advances in computing capacity, computer linking software and increased online availability of administrative data threaten their very existence[6].

[6] See, for example, Chapter 1 in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, co edited with Pat Doyle, Laura Zayatz and Jules Theeuwes, North Holland, 2001.

Dealing with the threats to public use files is an area in which much needs to be done – and one in which statistical agencies can join forces. One under-investigated area is the effect of the choice of different disclosure protection techniques on data quality. The lack of agency focus on this is evident: agencies that pour resources into producing top quality data - for example, survey design to improve response quality, and response follow-up to reduce attrition bias – will spend much less on the decision to top-code, data-swap or suppress information. While this lack of focus was rational in a less technologically savvy era, it is unlikely that statistical institutes will continue to be able to be so sanguine. I hope that agencies will increasingly rely on technical statistical analysis to make decisions about the appropriate level of data quality/data protection.

One of the most attractive technical developments, in my opinion, is that devoted to creating inference-valid synthetic datasets[7]. These datasets often use multiple imputation and other Bayesian techniques to create datasets with the same analytical structure as the underlying protected data. They can be used by researchers at a remote site to develop an understanding of the structure of the datasets, use simulated data to develop code and even estimate basic relationships before sending the code to the secure site to estimate the true underlying relationships. The quality of this approach is evident in Figure 3 – using French data, Abowd and Woodcock show that there is almost no difference between results estimated using some forms of synthetic data and real data. Other forms of synthetic data suffer some analytic difficulties but they appear to be manageable.

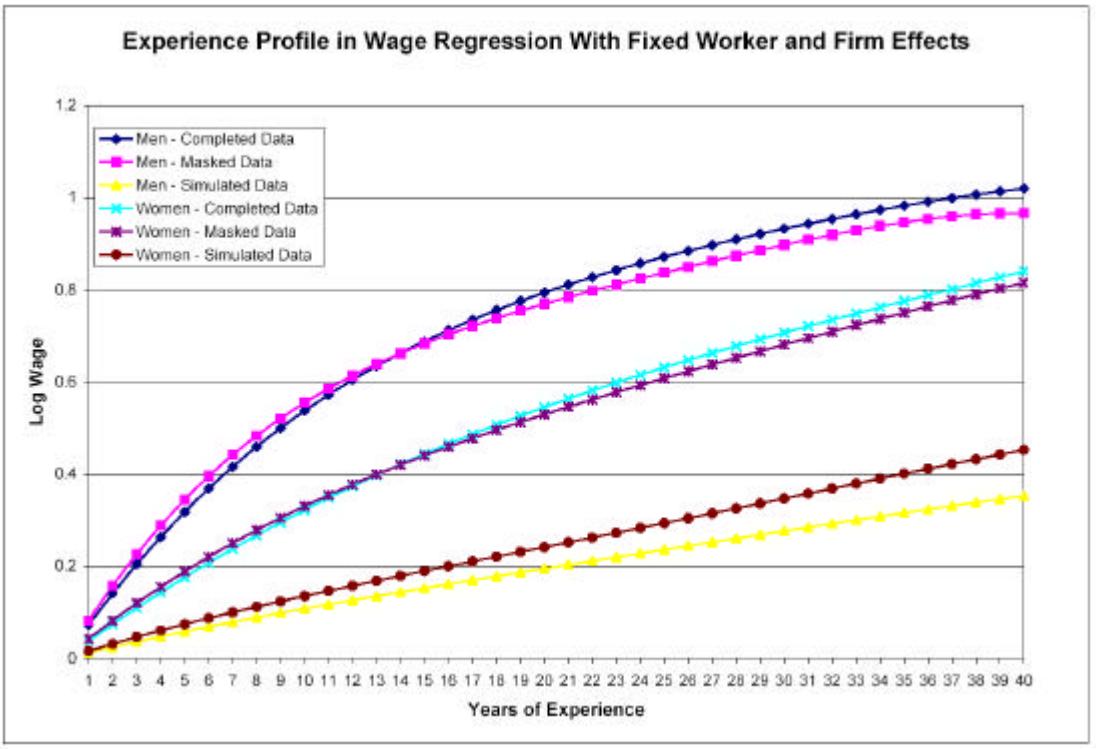
b) Reputational costs

Another very real cost associated with outside researcher access to national statistical institutes is that of reputation. The production of official statistics is the mandated reason for their existence – and the typical agency expends enormous effort making sure that published statistics with their imprimatur are the national gold standard. As a result, each agency is understandably concerned that research results using data with their imprimatur, and without their expertise, could be misconstrued as "official" – and be misused. Anecdotal evidence from developing countries reinforces this – there is a substantial fear that international officials can misuse the data, misinterpret it, and produce incorrect – possibly irreproducible - results that take inordinate amounts of time to clarify.

It is possible to manage this type of damage. The World Bank's Living Standards Measurement Survey (http://www.worldbank.org/lsms/) has extensive tutorials, software packages and "how-to" manuals to make sure that researchers working with similar datasets know what they're doing. An alternative approach was taken in the U.S. in the form of the recent "Information Quality Act" which requires the U.S. Office of Management and Budget to develop government-wide standards for data quality. Interestingly, that act distinguishes between "ordinary" and "influential" information – the latter including "influential scientific, financial or

[7] "Disclosure limitation in longitudinal linked data", Abowd and Woodcock (2001) in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, North Holland, Amsterdam, 2001.

Figure 3 Comparison of results based on synthetic and real data



Source: Abowd and Woodcock (2001)

statistical information" that will "have a clear and substantial impact on important public policies or important private sector decisions" (67 FR 8452). Even more tellingly, influential information should be reproducible by qualified third parties (though exceptions apply).

c) Disclosure of respondent identities

The ultimate cost to an agency is for an external researcher to disclose the identity of a business or individual respondent. While the penalties for this are typically substantial – ranging up to 10 years in jail and a $250,000 fine in the U.S. – the consequences of such a breach could be devastating to respondent trust and response rates. As trust in the government appears to be declining, statistical agencies might well also be concerned that respondent trust in their ability to protect respondent confidentiality is declining – and that this might only be exacerbated by permitting widespread researcher access

I need hardly tell a group of statistical agency heads that the only way to find out whether such perceptions are likely an important issue is to collect data and analyse it! There has been some research attempting to quantify the order of magnitude of the relationship between trust and response rates, and the trends over time in the U.S. (by Eleanor Singer for respondents to demographic surveys, and Nick Greenia for respondents to economic surveys). Indeed, a resolution was adopted at a UNECE confidentiality workshop in Skopje, Macedonia in 2001 to

move forward with a joint European endeavour to quantify the effect of researcher access on perceptions, but I am not clear on how much progress has been made in actual implementation.

Summing up

It is clear that statistical agencies will increasingly be challenged to provide more access to microdata. I would argue that this should not be seen as a necessary evil, but rather a chance to fulfil a critical societal mission. However, since increased access does not come without increased costs, it would seem reasonable for a conference such as this to see whether the costs might be reduced by combining efforts. Some areas in which joint research and development might provide substantial dividends, for example, would be:

i) the creation of inference-valid synthetic datasets;
ii) the protection of microdata that are integrated across several dimensions (such as workers/firms/geography);
iii) the quantification of the risk/quality trade-off in confidentiality protection approaches;
iv) the effect on response rates of increased microdata access.

I will close with two quotes. The first is from Chap T. Le and James R. Boen in *Health and Numbers: Basic Statistical Methods*. " There are aspects of statistics other than it being intellectually difficult that are barriers to learning. For one thing, statistics does not benefit from a glamorous image that motivates students to persist through tedious and frustrating lessons...there are no TV dramas with a good-looking statistician playing the lead, and few mothers' chests swell with pride as they introduce their son or daughter as "the statistician."" The reason I give you this quote is so that your feelings will not be hurt when I tell you that my children's reaction when I told them I was going to a Conference of European Statisticians meeting was, to say the least, underwhelming! However, I was delighted to be invited – because as Sir Francis Bacon noted "Knowledge is Power" – and your mission is to disseminate the data that underlies that knowledge. I firmly believe that the work you do is fundamental to the functioning of society – and will become increasingly important in an information driven society. I very much hope that your focus on microdata today will bear fruit in the form of providing the optimal amount of researcher access to microdata in each of your respective statistical agencies.

I.3 STATISTICAL DATA CONFIDENTIALITY AND MICRODATA

Invited paper by Matti Niva with Bo Sundgren and Ingrid Lyberg, Statistics Sweden

Introduction

The main challenge to a National Statistical Institute (NSI) regarding statistical confidentiality and microdata is to strike a balance between the confidentiality protection and increased use of microdata. As increased use of microdata implies improved possibility of providing better data to meet the needs of users, this balance lies at the heart of official statistics which should "*…provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data…*"[1]. Simultaneously, *"Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes*"[2]. In seeking this balance it is inevitable to combine different measures and actions; both legal, technical, administrative and methodological dimensions should be covered.

This paper aims to present some of the main issues faced when these different dimensions are addressed. Firstly, some ideas regarding the prospects of using microdata are presented. The next section deals with confidentiality and microdata. The legal issues are addressed thereafter and finally, different organizational approaches regarding access to microdata are briefly discussed. This also roughly corresponds to the outline of the seminar.

Use of microdata

Many simultaneous developments have increased the possibility to use microdata for research purposes. These include technological advances in hardware, software, data documentation and the Web. Modern PCs now have the processing capacity for advanced and large microdata sets. This implies that the NSIs can quite easily make their large data sets available to the researchers. This should be seen as an important part of the mission of an NSI: to assure that the wealth of microdata stored can be fully utilized by researchers and other legitimate users.

Traditionally, aggregate statistics were published according to what the NSI deemed important, although the users of course had an influence upon such decisions. The provision of tabular aggregate statistics also meant a clear limitation regarding how official statistics could be used in social and economic research.

The next step in the development of providing value added of the data stored at NSIs to users was to introduce statistical databases consisting of aggregated data matrixes and allowing the user to compile his own statistics to a large extent.

[1] UN Fundamental Principles of Official Statistics, Article 1, UN Statistical Commission 1994.
[2] Ibid, Article 6.

The access to microdata implies a major step further as researchers and other users themselves can choose the data suitable for their research. This has also had implications to theory developments in social and economic research. Many researchers can witness the importance of the use of microdata in analysing what the consequences of policy measures may be (e.g. Erikson, p.2). Theoretical explanations of aggregate conditions can thus be supplemented with analyses of mechanisms at the individual level with the help of statistical data of the NSIs.

The availability of large amounts of longitudinal microdata implies new analytical possibilities. For example a matching of different microdata files for several years opens new possibilities for dynamic analysis. This type of research based on microdata has been increasingly common during the past decades. This development is also obvious in economic research. A typical example from labour market economics is to link employee to employer data for analysis of both the supply and the demand side of the labour market (Westergaard-Nielsen, p.2).

NSIs can also integrate several microdata registers and create new databases. Normally, however, a lot of statistical work must be carried out to make the quality of data acceptable. Statistics Sweden has compiled some databases of this kind. The longitudinal database "Louise" with anonymised microdata on individuals and families regarding their education, income and employment might serve as an example. It should be added that this database includes annual data on all adults in Sweden from 1990 and is updated each year. Such an integrated database offers rich possibilities to carry out different analyses. An alternative to an integrated database is to link several microdata registers to each other on an ad hoc basis for specific purposes.

The increased availability of microdata combined with IT developments has also led to a new type of approach: data mining or, more broadly speaking, knowledge discovery in databases. This possibility is especially interesting given the possibility of multi-database data mining (Torra et al).

For the NSIs, the increased use of microdata implies value added in the form of better use of the data stored at the NSIs, and should also improve their legitimacy vis-à-vis respondents and the larger public. It also implies that the investments made in official statistics give a higher return.

Confidentiality

One major issue entwined in all use of all microdata is confidentiality. To put it bluntly, all use of microdata, even when anonymised, might imply a threat to confidentiality. Although a violation of confidentiality regarding microdata use has in fact hardly ever occurred in the NSIs' data based research projects, the confidentiality protection is still and should be a major issue and concern. The very positive track record so far is partly due to the efforts of the NSIs. Another probable reason could be that researchers dealing with microdata have their own human capital at stake. It is also customary for the microdata issue to facilitate contacts between the NSIs and the research community.

The need for privacy and integrity regarding statistical data is an old issue. One of the early perceptions of the need to strike a balance between the right to privacy and the increased need for information was put forward by Vincent P. Barabba in 1974 when. From the point of view of the US Census Bureau, he stated that "*…there is an inherent conflict in gathering data from individuals. The conflict is between the individual's right of privacy on the one hand, and on the other, government's use of mandatory processes to obtain the information it needs for valid purposes*" (Barabba, p.34).

The issue of confidentiality has been developed by and reflected in the documents of the international statistical community. In the ISI declaration on professional ethics from 1985 it is underlined that "Statistical data are unconcerned with individual identities" which implies that "*…identities and records of cooperating (or non-cooperating) subjects should …be kept confidential, whether or not confidentiality has been explicitly pledged*" (ISI, 4.5 Maintaining confidentiality of records). Further, statisticians should prevent their data from being published "*… in a form that would allow any subject's identity to be disclosed or inferred*" (ISI, 4.6 Inhibiting disclosure of identities).

The UN Fundamental Principles of Official Statistics are also very clear on this point: "*Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes*" (Article 6). In many countries a national code of practice has been developed for the NSIs. Normally, confidentiality protection is one of the corner stones of such declarations. This is also the case in the recent UK National Statistics Code of Practice, where the principle of protecting confidentiality is one of the main commitments of the National Statistics (National Statistics Code of Practice).

Confidentiality protection is clearly a crucial commitment of the NSIs. For ethical reasons, the NSIs are, of course, concerned about the integrity of the citizens as well as business establishments, but the commitment to confidentiality protection also has a specific explanation. The NSIs must be fully trustworthy in this respect to be able to gather data from respondents.

However, bearing in mind that use of microdata always might imply that the confidentiality is at risk (there is no such thing as completely safe microdata), the real issue is to strike a balance between increased information and confidentiality. NSIs normally estimate the risk of disclosure in different areas and phases of the statistical process regarding different types of uses and try to minimise such risks. An array of different methodological solutions has been put in place by the NSIs.

Some of the other issues regarding confidentiality protection are the specific features regarding the provision of microdata on businesses as opposed to individuals. Another is the timeframe, i.e. should confidentiality protection apply regardless of time (that is, forever), or should it last for a lifetime or even less?

In Central and Eastern European countries, the institutional and legal situation regarding statistics has changed significantly in the transition process. Data confidentiality in these countries may have some special features (e.g. legislative situation, implementation of one-way

flow principle[3], implementation of confidentiality protection throughout the statistical system). In a study carried out by the ECE secretariat, it was concluded that generic guidelines for statistical confidentiality would be valuable (see Chapter 2).

Recent developments regarding the need for improved national security may pose the problem of undermining statistical confidentiality by security concerns. If this would imply an unlimited access to microdata that has been reserved for statistical purposes, the credibility of the whole statistical community could be at stake.

So far, confidentiality protection has to a large extent been a national issue, but in the EU-context it becomes an issue for the EU institutions, such as Eurostat as well. New possibilities along with new tensions appear.

Legal issues

The arrangements for confidentiality protection can be based on legal acts and/or rules and regulations applied by the NSI. The legislative situation varies across countries and regions. If there is a Statistics Act of relatively recent date in a country, it normally contains regulations on statistical information. One central principle is usually that data collected for statistical purposes, regardless of whether it has been collected in accordance with prescribed obligation or is given voluntarily, may in principle only be used for the production of statistics. However, the data may also be used for research purposes under certain preconditions.

Normally, other legislation, such as The Personal Data Act, applies to the production of statistics and the release of microdata. In the EU-context, the so-called Data Protection Directive of the Council and the European Parliament (No 95/46 of 24 October 1995) is important as it strengthens legal protection of individuals with regard to automatic procession of personal information (applied to computerised personal data and data held in structured manual files) related to the individuals in question. All the Member States should have a corresponding national Personal Data Act based on the EU Directive. This might also imply that other authorities such as the Data Inspection Agency have a final say regarding the use of microdata on individuals.

When it comes to purely statistical confidentiality legislation in the EU, there is the EU Council regulation on Community statistics (No 322/97) according to which both national authorities and Eurostat shall protect confidential data. The main principle is that confidential data obtained exclusively for the production of Community statistics shall be used for statistical production only, unless the respondents have given their consent to its use for other purposes in an unambiguous way. The recent implementation regulation under the regulation on Community statistics concerns access to confidential data for scientific purposes (Commission Regulation 83/2002). This regulation is on rules concerning conditions permitting access to confidential data from four different EU-surveys. According to this regulation, scientific researchers might have such access if a contract regulating the terms of access is signed with Eurostat. However, it

[3] The one-way flow principle implies that the NSI has access to administrative records kept by ministries or other government agencies at the microdata level, whereas the possibility for a reverse flow of microdata subject to statistical confidentiality, whether from statistical surveys or from any other source, is strictly excluded.

remains to be seen how Member States will proceed with the implementation of this legal arrangement in practice.

Regarding data possessed by public authorities, one point of departure could be quite the opposite of confidentiality protection. It can be claimed that, for the sake of democracy, all information created within the public sector should be public. Accordingly, all decisions taken by a public authority and background documents facilitating these decisions, including correspondence of the public authority, should be made public. This principle of transparent public administration applies in Sweden. However, all exceptions to this principle of publicity due to motivated secrecy etc. must be stated in specific laws such as the Statistics Act.

Access to microdata – different approaches

One of the main challenges facing NSIs regarding microdata is to provide access to users in different ways. This access can in principle be organized in a number of ways and normally the NSI itself should find a suitable and feasible solution considering the prevailing institutional and organizational circumstances. At the same time, a lot of benchmarking has taken place during recent years.

One method of providing controlled access to microdata has been to compile anonymised Public Use Microdata Files (PUMFs). This solution was introduced by Statistics Canada in the early 1970's (Boyko & Watkins, p3). A rigorous processing is carried out before release of PUMF to reduce the probability of disclosure. Since the outset of this program, more than 350 PUMFs have been created and several other countries have chosen similar solutions. The PUMFs have been valuable for researchers in universities and government departments. Some of the problems related to their use have been relatively high costs, especially in the 1980's and early 1990's (use of mainframe, pricing policy) and the fact that the anonymisation process decreases the value of the data (ibid, p.5).

In many countries the delivery of de-identified microdata to researchers and other legitimate users outside the NSI is still the main method of release of microdata. There were around 200 such cases in 2002 at Statistics Sweden and the number of releases is rapidly increasing. If the released microdata is detailed register-based data, it is normally in fact not anonymous. It is obvious that the NSI in such cases must base the approval of the release of microdata on prevailing legislation and other regulations stipulating the confidentiality rules. In Sweden attempts to re-identify data are criminalized.

Because of the sensibility of microdata and the possibility of re-identification, the NSIs in many countries do not allow off-site use of microdata. Instead the NSI creates an on-site Research laboratory for the researchers. This is the case, for example, in Denmark (Access to microdata in the Nordic countries, p.14). This option also includes a solution where the NSI sets up and runs a Research Data Centre, e.g. at a University. In both cases, it is easier for the NSI to check that the confidentiality is not violated.

A still more cautious solution is to allow the use of microdata only by the NSI staff. In some cases the researcher becomes a staff member for a period of time in order to carry out the

microdata based research. Also, the NSI might have a policy of inclusion of research staff to be able to exploit the wealth of microdata for external clients.

However, it is becoming more common to authorise given research institutes to have on-line access to anonymised microdata of the NSI. This solution has been chosen in Australia, in Denmark and in Portugal. The data sets available on-line might be limited due to their sensitivity and also modified according the specific needs and orientations of the research institutes concerned. It is obvious that the remote access systems such as Internet-based on-line access are highly appreciated by researchers. It might also be attractive to a growing number of NSIs as this solution allows a certain control of the use of microdata.

The question of pricing is also relevant when discussing access to microdata. It is quite common for the NSI to have already been funded with appropriations for the major part of the work to compile and maintain microdata registers. If so, it would seem reasonable that the price charged corresponds to the extra costs incurred in the release of microdata. Such costs can, of course, be defined in a number of ways, depending on the calculation principles applied. However, the pricing of the release of microdata might also be based on other principles such as market pricing or even a free-of-charge basis under certain conditions.

Concluding remarks

This paper has underlined the fact that a balance must be struck between the protection of statistical confidentiality and improved access to microdata. It has also shown that a number of legal, administrative and technical measures must be combined to reach such a goal. This also implies that there are many different ways of reaching this balance. The statistical community could also work for a common policy and agree upon core principles regarding access to microdata. The CES seminar will hopefully contribute to a richer understanding of the options available.

References

Access to Microdata in the Nordic countries (2003). Statistics Sweden.

Barabba, Vincent P. (1974). The Right of Privacy and the Need to Know. Proceedings of the Social Statistics Section, American Statistical Association 33.

Boyko, Ernie & Watkins, Wendy (2002). Safe Data, Safe Places: No Either/Or Solutions. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'.

CES Seminar on Data Confidentiality (2003). Note by the ECE secretariat. CES/BUR.2003/27/Add.1

Declaration on professional ethics (1985). International Statistical Institute.

Erikson, Robert (2002). The right to privacy and the right to information. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'.

National Statistics Code of Practice. Statement of Principles (2002). National Statistics, UK.

Perpétuo, Fernanda (2002). Statistical Information System for Researchers. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'.

Sundgren, Bo (2001). Statistical Microdata – Confidentiality Protection vs. Freedom of Information. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, FYROM.

Torra, Vincenç & Domingo-Ferrer, Josep & Torres, Àngel (2003). Data Mining Methods for Linking Data from Several Sources. Paper to the joint ECE/Eurostat work session on statistical data confidentiality. Luxembourg 7-9 April 2003.

UN Fundamental Principles of Official Statistics (1994). Adopted by the UN Statistical Commission.

Westergaard-Nielsen, Niels (2002). Linking employer-employee data. Paper to the 19th CEIES seminar 'Innovative Solutions in Providing Access to Microdata'.

CHAPTER II: DATA CONFIDENTIALITY

II.1 SUMMARY OF DISCUSSANTS' MAIN POINTS

by Tadeusz Toczynski, Central Statistical Office, Poland

Comments on the report of Mr. Sokolin from the State Committee of the Russian Federation on Statistics (Goskomstat)

Confidentiality of statistical data is one of the Fundamental Principles of Official Statistics. Why is it so important?

- Statistical offices which implement the principle of protecting individual statistical data are building trust between statisticians and respondents;
- Reciprocal confidence is one of the most important conditions of high quality statistics;
- Confidence should be clearly strengthened by legal acts regulating mutual relations and responsibilities.

The confidentiality of data on physical persons differs from the confidentiality of data on legal persons. In practice and by law, statistical information collected on physical persons is much more strongly protected. It will be interesting to see in what direction the Law on Commercial Secret (recently submitted to the Parliament) will develop. Respondents themselves can define which information should be considered as confidential and statistical organs will have to ensure the confidentiality of that information. The question that arises is: does this mean that other information can be disseminated to the public?

Regarding access to microdata by government bodies in the Russian Federation, a broad group of government institutions may have access to individual data under special circumstances and following special procedures, which means that they can probably access individual data. Can they also request data on physical persons?

Access to statistical data for researchers is now considerably limited. Some ways of providing microdata to users on a larger scale are under consideration.

The problems which Russian statisticians are facing are the following:

- the absence of a special law on statistics;
- the existence of clauses in various legislative acts which authorise certain government bodies to request individual microdata;
- statistical data are regarded as an information source for administering the territory, and even for controlling activities of particular business entities;
- technical difficulties in data transmission by means of a telecommunication system.

These problems do not concern only Russia, but also countries in transition which are trying to improve their statistical laws.

Comments on ECE Statistical Division's ad-hoc survey on data confidentiality problems

A short questionnaire comprising 11 questions was sent out by the ECE Secretariat, and 24 countries in transition economies replied. However, I think that we would have a wider picture if the questionnaire had been sent to all of the ECE region countries.

What are the results of the survey?

In 18 out of 24 countries, the principle of statistical confidentiality is defined by law (they have a special law on statistics) which provides safe protection for the statistical office from any requests to release data permitting disclosure of information on individual units.

In 21 countries there is sufficient legal protection of microdata of business entities, irrespective of their legal form, but in only 12 countries is the one-way flow principle for microdata between other parts of government and the statistical office provided by law. In some countries the legal basis is sufficient but there are many problems in implementing the law – there is pressure from officials to provide access to microdata.

Only 7 countries are obliged by law to disseminate disaggregated results that allow to infer information on single economic units to which the principle of statistical confidentiality should apply.

There are problems in small countries. However, even in Poland we continuously monitor aggregated data at the level of NUTS 3 and 4 (the administrative division of the country starts at the lowest level of more than 2500 communes - gminas). We are obliged, even if not really by law, to supply basic statistical data at territorial level.

Legally possible access to microdata for researchers is reported by 15 countries. However, most countries reported receiving few requests by researchers. This situation is not totally clear to me. In Poland we receive many requests but access to microdata, even non-identified, is strongly limited. In practice it embraces only the use of common studies conducted together with the Central Statistical Office.

The questions formulated by the 19 ECE member countries gave a positive picture of the respondents' perception of the statistical office's guarantee to keep their information confidential and to use it only for statistical purposes. In such circumstances, an important and difficult question is raised: why are we so strongly involved in discussion on special treatment of such a group of users like researchers, and what would be a balance of our profits and losses when we expand access to microdata for research purposes?

Is there is a difference between researchers and government research institutions? There is no difference. Neither group have access to data although, in Poland, such government institutions as the Government Centre of Strategic Studies, the National Bank or the Ministry of Economy also prepare strategic papers and conduct statistical studies using the same concepts, methods and tools as researchers. We should trust researchers, they do have a special right. If they conduct statistical studies properly, they should produce the same data as do we with our

statistical inquiry. What about giving access to data for commercial purposes? What is the difference between research and commercial purposes, and who will be able to judge? Will our respondents accept a law stating that we can pass microdata to researchers? (no such law exists to date). I do not think so. We have not asked them yet.

The increasing demand from researchers for microdata makes the challenge of maintaining the confidentiality of microdata more difficult. The main challenge to a NSO is to strike a balance between the confidentiality protection and the increased use of microdata for statistical and research purposes. In this context a significant amount of balancing is necessary when formulating legislation to protect personal integrity as regards personal data.

According to recent EU regulations, scientific researchers might obtain access to data if a contract regulating the terms of access is signed with Eurostat. The main principle is that confidential data obtained exclusively for the production of Community statistics shall be used for statistical production only, unless the respondents have given their consent to their use for other purposes in an unambiguous way. But other approaches can be found, e.g. in Poland where, in order to gain the full confidence of respondents and to encourage them to transmit complete and reliable information, it is absolutely forbidden to use data for other than statistical analysis. Additionally, the law on official statistics does not stipulate that respondents could be asked to consent to the use of the data for other purposes. Obviously, this varies from country to country. In the Nordic countries, legislation gives more possibilities to the statistical service as regards data processing.

Problems for further discussion

- The problem of "informed" consent – The principle underlying statistical data collection is that of informed consent. The principle is that the business entity which provided the original information has a right to know what the information will be used for and who will see it. The argument here is that if there are new users and new uses of the information, the entity should be made aware of it. In some countries it may be necessary to change the law under which data are collected in order to specify the uses of the data. It should be an important part of the contract between the entity and the statistical organization that can be seen by the organization as a factor affecting response rates.

- The problem of definition of "scientific purposes" – It is difficult to draw the line between what is destined for scientific purposes and what is destined for commercial ones. Doctoral students or undergraduates also use scientific knowledge. Training purposes can also be regarded as scientific. Can we be sure that researchers asking for data will not use them for pure business activities?

- The problem of anonymised confidential data – If direct identification is not possible and the risk of indirect identification is minimised following current best practice, the data are not disclosive and are no longer confidential. On the other hand, such data are always confidential no matter how much they have been modified?

- The communication strategy with respondents – Clear rules are needed to create confidence, especially with the respondents. There is not only the need to inform the researchers about the new legal provisions but also about why the constraints or inconveniences they may provide are necessary. We should remember that the value system of researchers is different from that of official statisticians.

- As we can read in the Swedish paper, the balance must be struck between the protection of statistical confidentiality and access to microdata. A number of legal, administrative and technical measures must be combined to reach such a goal. As there are many different ways of reaching this balance, the statistical community could work for a common policy and agree upon core principles regarding access to microdata.

Closing remarks

Problems of data confidentiality will never disappear from our everyday practice. Therefore, it is necessary to create common procedures which would be used by all NSOs. Lack of severity in the protection of values formulated in Fundamental Principles could destroy the confidence of our respondents. Staying in the middle of the "crossroads" and waiting for somebody to show us the correct signpost could result in the loss of our independence and credibility. Presidents of NSOs are obliged to protect statistical data. In Poland, we began closing access to individual data from government bodies and individuals even before the law was passed in 1996.

II.2 DATA CONFIDENTIALITY - A SURVEY OF TRANSITION ECONOMIES

Supporting paper by the ECE secretariat

Introduction

In order to obtain an overview of the specific concerns that exist in transition economies concerning data confidentiality issues, an ad-hoc survey was carried out in January 2003. The survey was conducted with the help of a short questionnaire comprising eleven questions.

The statistical offices of the following twenty-four countries submitted answers: Albania, Armenia, Azerbaijan, Bulgaria, Croatia, Czech Republic, Estonia, Georgia, Hungary, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Poland, Republic of Moldova, Romania, Russian Federation, Serbia and Montenegro, Slovakia, Slovenia, The former Yugoslav Republic of Macedonia, Turkmenistan, Ukraine and Uzbekistan.

The text below provides a summary of the results of the survey. Detailed information by country is provided in the attached tables.

Implementation of the principle of statistical confidentiality (Table 1)

The survey showed that the definition of the principle of statistical confidentiality in the legislation provides safe protection for the statistical office from any requests to release data that permit either direct or indirect disclosure of information about individual units in eighteen out of twenty-four countries (Question 1). Only Azerbaijan and the Russian Federation reported that such protection is not provided through the definition of statistical confidentiality in the legislation, while Kazakhstan, the Republic of Moldova and Ukraine reported that such protection only exists partially. In Croatia, a new EU conform legislation is about to be implemented.

A second question addressed the protection through statistical confidentiality of units other than natural persons and private households, notably private and recently privatised companies, irrespective of their legal form (Question 2). This was reported to be the case in twenty-one countries. Only the Russian Federation reported that this kind of legislation is not available, while the legal situation was reported to be unclear in Moldova and also in Croatia. However, Croatia indicated that they were in the process of implementing EU conform legislation.

The one-way flow principle for microdata between other parts of the government and the statistical office is established by law and implemented in twelve transition economies (Question 3). A “legally sufficient situation, but many practical problems” was reported by five countries (Albania, Estonia, Kazakhstan, The FYR of Macedonia, Uzbekistan), while seven countries reported that the “legal situation is insufficient” or that the “principle is not established in the law” (Azerbaijan, Georgia, Kyrgyzstan, Republic of Moldova, Russian Federation, Serbia and Montenegro, and Ukraine).

Countries were also asked to report on how the principle of statistical confidentiality is applied by other producers of official statistics, notably by regional statistical offices that are not or are not fully part of the central statistical office (Question 9). Independent of the legal situation, virtually all countries reported that the regional offices follow the principle of statistical confidentiality. In some cases, the regional offices only collect data and do not disseminate information, or they only disseminate summary information. Many countries mentioned that the principle of statistical confidentiality applies to all producers of official statistics. Only Georgia reported some violations of the principle at the regional level, while Kazakhstan reported some violations by other government bodies producing statistics. Moldova reported on some problems related to the requests of local government authorities for individual data on economic units. However, Moldova also indicated that the confidentiality principle for legal persons is strictly followed.

Data processing tasks for statistical and administrative purposes (Table 2)

The survey was also concerned with the issue of whether the statistical office is in charge of data processing tasks for administrative purposes or is responsible for managing administrative registers, and how strictly such activities are separated from the statistical activities in terms of organisational structure and IT.

Half of the countries reported that they maintain both administrative and statistical registers, and only three of them explicitly reported that activities on administrative and statistical registers are strictly separated in terms of organizational structure and IT.

Confidentiality issues related to dissemination (Table 3)

As regards the obligation of the statistical office to disseminate disaggregated results that allow for inference about single economic units to which statistical confidentiality would be applicable, only seven out of the reporting twenty-four countries indicated that such obligations exist (Estonia, Kazakhstan, Kyrgyzstan, Russian Federation, Serbia and Montenegro, Slovenia and The FYR of Macedonia).

Other confidentiality issues related to the potential disclosure of individual units in disseminating results can be summarised in two groups: (i) small area statistics: regional data and/or small countries; (ii) sectoral business data where one entity is the sole producer.

Access to microdata and how to deal with these requests (Table 4)

In most transition economies (15), it is legally possible for researchers to access microdata for their own statistical purposes. Only six countries (Georgia, Hungary, the Republic of Moldova, Turkmenistan, Ukraine and Uzbekistan) reported the exclusion of this option legally and thus do not provide access to microdata. Two countries (Kazakhstan, and Serbia and Montenegro) reported an unclear legal situation and, as a consequence, do not provide access to microdata. In the Russian Federation, there is no legislation concerning access to microdata. However, microdata are provided under procedures established by the statistical agency.

In most countries, few requests by researchers are received. Only four countries reported receiving many requests for access to microdata.

For those countries where access to microdata is legally possible, the procedures appear to be quite similar, especially since most of these countries are pre-accession countries and are thus targeting the implementation of Commission Regulation No. 831/2002 (EC): Access is given only to specific institutions and for specific purposes (scientific research). Often, an agreement is signed containing the exact conditions for using the data.

Respondents' perception of confidentiality protection (Table 5)

Almost all countries (19) gave a positive picture about the respondents' perception of the statistical office's guarantee to keep their information confidential and to use it only for statistical purposes. Two problems related to the perception of respondents could be identified: non-awareness of respondents of the confidentiality protection through the statistical agency, and the non-trust of respondents regarding confidentiality protection through the statistical agency.

Conclusions from the ad-hoc survey (Table 6)

Transition economies were also asked to indicate which of the confidentiality issues mentioned in the questionnaire they consider to be most important in order to improve the present situation. An overview of the results is shown in Table 6 where countries are grouped into pre-accession countries and CIS and other countries. The priority issues can be summarised as follows:

- issues related to access to microdata;
- the legal implementation of the principle of confidentiality: of utmost importance in CIS and other countries;
- the need for methodological and technical standards in the pre-accession countries;
- issues related to administrative registers: of priority to CIS and other countries;
- respondents' perception of confidentiality protection: although almost all countries reported a positive attitude of respondents' towards confidentiality issues, there seems to be room for improvement.

Table 1: Implementation of the principle of statistical confidentiality

	Question 1	Question 2	Question 3	Question 9
	Is the principle of statistical confidentiality defined in the legislation in such a way that it provides safe protection for the statistical office from any requests to release data that permit either direct or indirect disclosure of information about individual units?	Are units other than natural persons and private households, notably private and recently privatised companies irrespective of their legal form, protected by statistical confidentiality?	Is the one-way flow principle for microdata between other parts of the government and the statistical office firmly established and implemented?	How is the principle of statistical confidentiality applied by other producers of official statistics, notably by regional statistical offices that are not or not fully part of the central statistical office?
Albania	yes	yes	legally sufficient situation, but many practical problems	The regional offices are part of the central statistical office; regional offices only collect data and send it to the national office for further processing
Armenia	yes	yes	principle established in law and implemented	The regional offices strictly follow the principle of statistical confidentiality
Azerbaijan	not at all	yes	principle not established in law	Regional offices only disseminate summary information
Bulgaria	yes	yes (natural and legal persons)	principle established in law and implemented	The principle of statistical confidentiality is applied by all official producers of statistical data: Central Office and regional offices, and other bodies
Croatia	only partially (a new law conform with EU legislation is under way)	unclear legislation (legally no, but de facto yes)	principle established in law and implemented	Other producers of official statistics follow the legal provisions on statistical data confidentiality
Czech Republic	yes	yes	principle established in law and implemented	The regional SOs are part of the CSO; their publications are supervised by the CSO

Table 1: Implementation of the principle of statistical confidentiality (cont.)

	Question 1	Question 2	Question 3	Question 9
Estonia	yes	yes	Legally sufficient situation, but many practical problems (non-compliance of different legal acts; different interpretation of legal acts)	There are no regional statistical offices; the other main producer of official statistics (the Central Bank) applies the principle of statistical confidentiality
Georgia	yes	yes	Legal situation insufficient	In general, the legislation is to be followed by all producers but there are some violations at regional / local level
Hungary	yes	yes	Principle established in law and implemented	Act on statistics applies to all persons and offices dealing with statistics
Kazakhstan	only partially	yes	Legally sufficient situation, but many practical problems	Regional and national statistical offices respect the principle of statistical confidentiality but there occur certain violations from other government bodies that also produce statistics
Kyrgyzstan	yes	yes	If necessary, statistical agencies have access to microdata in ministries and government agencies; if necessary, microdata from statistical agencies are given for analyses to ministries and government agencies	The confidentiality policy applies in the same way to all producers
Latvia	yes	yes	Principle established in law and implemented	Local offices do not disseminate data, this is the responsibility of the Central Office; other producers follow the principles of the Data Protection Law and Statistics Law
Lithuania	yes	yes	Principle established in law and implemented	Law on Statistics applies to all producers of official statistics

Table 1: Implementation of the principle of statistical confidentiality (cont.)

	Question 1	Question 2	Question 3	Question 9
Poland	yes	yes	Principle established in law and implemented	Regional statistical offices are an integral part of the centralised system and follow the statistical act
Republic of Moldova	only partially	unclear legislation	Legal situation insufficient	The regional statistical offices are governed by the Law on Statistics of the Republic of Moldova. There are problems related to the requests of local govemment authorities for individual data of economic units. The confidentiality principle of legal persons is strictly followed.
Romania	yes	yes	Principle established in law and implemented	Regional offices follow the rules established at national level
Russian Federation	not at all	no	Principle not established in law	Despite the absence of a law on statistical reporting, the statistical agency undertakes steps to ensure confidentiality of statistical data; it guarantees confidentiality to reporting units. A procedure is set up on the provision of data to third parties - this is only possible if the reporting unit agrees (except for the cases foreseen by law)
Serbia and Montenegro	yes	yes	Legal situation insufficient	The principle of statistical confidentiality is applied in the same way as for the statistical agency
Slovakia	yes	yes	Principle established in law and implemented	There are no independent regional statistical offices
Slovenia	yes	yes	Principle established in law and implemented	There are no regional offices; all producers of official statistics must follow the National Statistics Act

Table 1: Implementation of the principle of statistical confidentiality (cont.)

	Question 1	Question 2	Question 3	Question 9
The former Yugoslav Republic of Macedonia	yes	yes	Legally sufficient situation, but many practical problems	Regional offices are part of the State Statistical Office; microdata are exchanged between producers of statistical surveys with everyone liable to statistical confidentiality; for providing or exchanging data, co-operation contracts are also signed
Turkmenistan	yes	yes	Principle established in law and implemented	Regional offices are part of the national one
Ukraine	only partially	yes	Legal situation insufficient	Regional offices follow the same rules as the national one
Uzbekistan	yes	yes	Legally sufficient situation, but many practical problems	Regional and national office are regulated by Statistical Law that stipulates confidentiality requirements

Table 2: Data processing tasks for statistical and administrative purposes

	Question 4	Comments
	Is the statistical office (NSO) in charge of data processing tasks for administrative purposes or responsible for managing administrative registers? How strictly are such activities separated from the statistical activities in terms of organisational structure and IT? What are the repercussions of such non-statistical tasks on the core task of official statistics (including the ability to set up and manage statistical registers), and on the trust of respondents in the statistical surveys?	
Albania	no tasks outside statistics	
Armenia	no tasks outside statistics	
Azerbaijan	yes (no further specification given)	
Bulgaria	yes	The statistical office (NSO) establishes and maintains also an administrative register, this operation is done in a separate department. For NSO's statistical activities, a statistical register is maintained; statistical and administrative activities are separated; information from the administrative register can be used for statistical purposes.
Croatia	yes	The NSO manages a business register (administrative and statistical register) in one department; administrative register data are public while data from the statistical register are treated as confidential.
Czech Republic	yes	Processing of some data from other government departments; use of administrative registers managed in other government departments for statistical purposes
Estonia	no tasks outside statistics	
Georgia	yes	The NSO manages an administrative register; activities are not fully separated from statistical activities (organisation, IT)

Table 2: Data processing tasks for statistical and administrative purposes (cont.)

	Question 4	Comments
Hungary	no tasks outside statistics	
Kazakhstan	no tasks outside statistics	
Kyrgyzstan	no tasks outside statistics	
Latvia	no tasks outside statistics	
Lithuania	no tasks outside statistics	
Poland	yes	The NSO is responsible for maintaining two official registers; thes e activities are regulated in the Law on Official Statistics; work on the administrative registers is strictly separated from the primary statistical tasks
Republic of Moldova	yes	The NSO also maintains administrative and statistical registers; the activities are not separated (organisation, IT); other activities: supplying information from the administrative register
Romania	no tasks outside statistics	
Russian Federation	yes	The NSO develops and adopts forms of primary statistical reporting, i.e. common requirements for reporting on financial, investment and other types of economic activities; the NSO and regional offices maintain an administrative register of enterprises; the NSO ensures that standard economic and Social classifications are used in preparing new legal documents
Serbia and Montenegro	yes	The NSO maintains administrative registers
Slovakia	no tasks outside statistics	The NSO maintains only statistical registers while other government departments maintain administrative registers and have to supply data to the NSO
Slovenia	no tasks outside statistics	

Table 2: Data processing tasks for statistical and administrative purposes (cont.)

	Question 4	Comments
The former Yugoslav Republic of Macedonia	yes	The NSO maintains administrative registers; databases are separated, and microdata from administrative registers can be used for statistical purposes; statistical microdata cannot be used for administrative purposes; new legislation to move the administrative registers out of the NSO is under preparation
Turkmenistan	no tasks outside statistics	
Ukraine	yes	The NSO maintains an administrative register; the activity is under the Law on Statistics; the statistical register is created on the basis of the administrative register
Uzbekistan	yes	The NSO maintains also an administrative register; this activity is covered by the legal acts

Table 3: Confidentiality issues related to dissemination

	Question 5	Comments	Question 6	Comments
	Are there obligations for the statistical office to disseminate (either generally to the public or limited to specific users) disaggregated results that allow inference about single economic units to which statistical confidentiality would be applicable?		Are there other problems related to the protection of individual units in disseminating results?	
Albania	no		no	
Armenia	no		no	
Azerbaijan	no		no	
Bulgaria	no		no	Law on Statistics: individual and personal data cannot be provided, as well as data which summarise information for less than 3 units or in which the relative part per unit is over 85% of the total volume
Croatia	no		no	
Czech Republic	generally no	Special case co-operation with other government bodies which comply with the Statistical Act including confidentiality attachment	yes	Problem with regional data - easy disclosure; also - national level/sole producers (e.g. Skoda Auto)
Estonia	yes	Data which permit identification are only transmitted/dissemin - ated with written consent of respondent; dissemination without consent for scientific research in line with the legislation	yes	Problem: small country
Georgia	no		no	

Table 3: Confidentiality issues related to dissemination (cont.)

	Question 5	Comments	Question 6	Comments
Hungary	no		yes	Problem: dissemination of small area data; some data need to be excluded from dissemination so the full scope dissemination of data is inconsistent due to the confidentiality protection
Kazakhstan	yes	Some data (biggest enterprises, monopoly enterprises) are disseminated to a limited number of persons in government bodies; to prosecutor's office on request in criminal cases	no	
Kyrgyzstan	yes		yes	The Programme of Statistical Work is approved by the Government every year
Latvia	no		no	
Lithuania	no	Law on Statistics / article on confidentiality	yes	Small country, sometimes only one enterprise generating big share of the production
Poland	no		no	
Republic of Moldova	no		yes	Many requests from some Ministries and other government agencies to provide data on economic units; at the same time, data that Ministries have are not used by other government institutions
Romania	no		yes	Regional data, data where one entity has special activity; in these cases, the confidentiality protection is provided by law

Table 3: Confidentiality issues related to dissemination (cont.)

	Question 5	Comments	Question 6	Comments
Russian Federation	yes	More than 20 state bodies have legal rights to request and obtain statistical information	yes	Some representatives of state, regional and local authorities do not seem to understand the principle of statistical confidentiality. They regard statistical data as information means for managing regions or even enterprises. Such attitudes are particularly strong at the regional level.
Serbia and Montenegro	yes	Under the Law on Information System of Bodies and Organisations and only upon request of the government bodies	no	
Slovakia	no		yes	Dissemination of regional data, especially for enterprises/sectoral structure
Slovenia	yes	Small country - dissemination at the micro level vs. confidentiality	yes	Small country - dissemination at the micro level vs. confidentiality
The former Yugoslav Republic of Macedonia	yes	Only for research institutions under the State Statistical Law	yes	Small country - few producers in a specific branch of economic activity; regional statistics
Turkmenistan	no		no	
Ukraine	no	There are legal provisions in the Laws on the Directorate of Public Prosecution, on Internal Affaires Organs and on Security Services giving these institutions the right to request any statistical information they may need.	no	
Uzbekistan	no		no	

Table 4: Access to microdata

* no microdata access provided
** only for legal public information

	Question 7	Question 8
	Do researchers have access, under certain conditions, to microdata of the statistical office for their own statistical purposes?	Have you been confronted with requests from researchers for microdata and, if so, how have you responded?
Albania	legally possible	few requests
Armenia	legally possible	no requests
Azerbaijan	legally possible	few requests
Bulgaria	legally possible	few requests
Croatia	legally possible	few requests
Czech Republic	legally possible	few requests
Estonia	legally possible	few requests
Georgia	**legally excluded***	few requests
Hungary	**legally excluded***	many requests
Kazakhstan	**unclear legislation***	few requests
Kyrgyzstan	legally possible	many requests
Latvia	legally possible	few requests
Lithuania	legally possible	few requests
Poland	legally possible	many requests
Republic of Moldova	**legally excluded***	few requests
Romania	legally possible**	few requests
Russian Federation	**no legislation**	no requests
Serbia and Montenegro	**unclear legislation***	few requests
Slovakia	legally possible	few requests
Slovenia	legally possible	few requests
The former Yugoslav Republic of Macedonia	legally possible	few requests
Turkmenistan	**legally excluded***	few requests
Ukraine	**legally excluded***	many requests
Uzbekistan	**legally excluded***	few requests

Table 5: Respondents' perception of confidentiality protection

	Question 10	comments
	What is the perception of respondents about the statistical office's guarantee to keep their information confidential and to use it only for statistical purposes?	
Albania	Perception of respondents good, they are aware that INSTAT protects confidentiality	positive
Armenia	Results from a survey of respondents showed positive attitude	positive
Azerbaijan	Perception of respondents: they are aware of confidentiality protection	positive
Bulgaria	Survey of respondents (firms) - positive about confidentiality protection by NSO	positive
Croatia	High response rates in surveys - assume that respondents trust in confidentiality protection	positive
Czech Republic	Most respondents trust in confidentiality, few complaints which can be answered by NSO satisfactorily	positive
Estonia	Respondents are notified on confidentiality in questionnaires; respondents accept this	positive
Georgia	Trust expressed by good cooperation, with some exceptions	positive
Hungary	If confidentiality is offended, criminal procedures can be initiated; this never happened; the respondents seem satisfied	positive
Kazakhstan	Not all respondents trust NSO; respondents not even aware of confidentiality protection	some problems
Kyrgyzstan	The NSO does not think that perception of confidentiality protection is a reason for non-response	positive
Latvia	Some public opinion polls show that NSO ranks high in confidence level	positive
Lithuania	Respondents not yet surveyed; events like round tables, workshops indicate that respondents seem to be satisfied	positive
Poland	Respondents perceive that statistical confidentiality is fully respected	positive
Republic of Moldova	General positive perception but respondents are not always fully convinced that confidentiality is kept/disclosure of individual data to other government bodies	some problems
Romania	Respondents are aware and understand confidentiality protection; confidentiality protection is promoted starting with data collection/questionnaire	positive
Russian Federation	Positive	positive
Serbia and Montenegro	There is a certain level of mistrust	some problems

Table 5: Respondents' perception of confidentiality protection (cont.)

	Question 10	Comments
Slovakia	Legal confidentiality protection, positive perception of respondents; still problems with getting information from monopoly enterprises	some problems
Slovenia	General trust, good practice	positive
The former Yugoslav Republic of Macedonia	NSO did not have any comments till now	positive
Turkmenistan	The respondents know that confidentiality is protected by law and accept this	positive
Ukraine	In order to improve respondents trust, confidentiality guarantees are explained on the forms of statistical surveys	some problems
Uzbekistan	Respondents trust the NSO	positive

* no microdata access provided

** only for legal public information

Table 6: Conclusions

Priority issues named	pre-accession countries*	CIS + others**
Issues related to access to microdata	7	8
The legal implementation of the principle of statistical confidentiality (legislation, ensure protection of all individual units, implement one-flow principle, ensure that all producers of official statistics apply the principle of confidentiality)	3	11
The need for methodological rules/unified procedures/technical norms for observation of confidentiality issues/methods for data disclosure	5	1
Issues related to administrative registers	0	5
The relationship with users/respondents	2	4
Dissemination issues	2	1

* Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia

** Albania, Armenia, Azerbaijan, Croatia, Georgia, Kazakhstan, Kyrgyzstan, Republic of Moldova, Russian Federation, Serbia and Montenegro, The former Yugoslav Republic of Macedonia, Turkmenistan, Ukraine, Uzbekistan

II.3 CONFIDENTIALITY OF STATISTICAL DATA - THE RUSSIAN FEDERATION

Invited paper by Vladimir L. Sokolin, State Committee of the Russian Federation on Statistics (Goskomstat)

General

Confidentiality of statistical data is one of the fundamental principles of official statistics. It presupposes that individual data collected by the Statistical Office are confidential and can be used only for statistical purposes, whether the data concern physical or legal persons.

Implementation of the confidentiality principle aims at protecting immunity of private life and at improving trust between the Statistical Office and respondents.

The quality of official statistics greatly depends on good collaboration between the Statistical Office and respondents. Protected by confidentiality, the latter are encouraged not to conceal or distort the information requested from them in statistical surveys.

Legal acts regulate official statistical activities in many countries. Some countries have a special law on statistics, others make use of legal acts relating to individual areas of statistical activity. Usually such legal acts contain norms ensuring confidentiality of primary statistical data.

Currently, there is no specific law on official statistics in the Russian Federation. However, official statistics are part of the state information resources, the creation, maintenance and use of which, including provisions for confidentiality, are governed by a number of federal laws.

One of the instruments for ensuring confidentiality is provided by the federal law "On principles of the state service", according to which persons entering state service take upon themselves an obligation not to disclose statistical data on legal or physical persons which may become known to them through the service they perform. Non-compliance with this obligation may have disciplinary consequences for the person concerned including dismissal from office.

Confidentiality of data on physical persons

The 1993 Constitution of the Russian Federation proclaimed the protection of privacy and provided for protection of personal and family life. Collection, storage, use and dissemination of information on private life without the concerned person's agreement are forbidden.

This norm was further developed in the federal law "On information, informatics and information protection", which contains a special clause on personal data, i.e. information about facts, events and circumstances permitting identification of the person. Such data are defined by the law as confidential.

The application of the confidentiality principle to personal data aims at protecting persons' private life from illegal use of data from files with personal information. This is particularly important in conditions of high criminality.

Goskomstat of Russia follows these legal norms in carrying out surveys of physical persons.

Household surveys and labour force surveys cover only those persons who have agreed to participate in them.

The population census is subject to a special legal act.

According to the federal law "On the population census", all data from the census forms are confidential and cannot be disseminated. They are used to form various federal information resources.

Special organizational and technical procedures and methods were developed to process confidential data of the 2002 population census.

Confidentiality of the population census data is ensured by using an independent local computer network, by applying a system of passwords, by introducing electronic signature to confirm the accuracy of data transferred over telecommunication channels and by certain other techniques.

Anonymisation of data records precedes data aggregation. Aggregate data do not have confidentiality status.

These measures have proven effective in ensuring confidentiality of the census microdata.

Confidentiality of data on legal persons

Goskomstat of Russia produces economic statistics on the basis of source data collected from legal persons and other economic units.

The federal law "On information, informatics and information protection", the Civil Code of the Russian Federation, the Statute of Goskomstat and the Fundamental Principles of Official Statistics adopted by the UN Statistical Commission in 1993 form the legal basis for Goskomstat to ensure confidentiality with regard to the data provided by respondents in the federal statistical reporting forms. These data are used exclusively to compute statistical aggregates for the country as a whole, for its regions, by branches and sectors of the economy and for the social sphere. The reporting forms contain a note informing respondents about the confidentiality guarantee.

Annual balance sheets of enterprises have a different status. According to the federal law "On accounting", annual balance sheets of practically all financial and non-financial enterprises are

non-confidential and can be made public. They are transmitted to territorial statistical organs who make them available to all interested users. This has been Goskomstat's practice since the adoption of the law in 1996.

Possibilities of using legislation on commercial secrets

Legislation on commercial secrets can be used to promote confidentiality of individual data before a special law on statistics is adopted.

The Government has recently submitted a draft law "On commercial secret" to the Parliament. The law stipulates that legal or physical persons undertaking entrepreneurial activities have the right to decide themselves which part of the information in their possession should be treated as forming a commercial secret.

The owner of a commercial secret is obliged to provide information forming the commercial secret to government bodies in cases stipulated by law.

This means that respondents can define themselves what information should be considered as confidential and that statistical organs will then have to ensure confidentiality of that information.

Access to microdata by government bodies

Provision of statistical microdata on enterprises to users poses a number of problems. Goskomstat carried out a special survey of large- and medium-sized enterprises in 2001-2002 asking one question: would they agree that their individual data collected by regular statistical surveys be available to any interested user without any limitation? About 22 percent of surveyed enterprises gave a positive answer.

Access to microdata is presently governed by the procedure established by Goskomstat, according to which reported microdata can be provided to interested users only if there is a written agreement from the enterprise concerned. There are exceptions to this rule resulting from specific provisions in various legislative acts.

There is strong pressure on Goskomstat, particularly from local government bodies and from tax authorities, to provide access to microdata on individual enterprises.

Certain state institutions, like Courts of Justice, Police, Office of Public Prosecutor and some others, do have the legal right to access microdata. Goskomstat has established a separate procedure for these institutions.

In order to obtain confidential data, these institutions must submit a written request to the respective territorial statistical body. The request should be written on a special form and signed by

a competent person. Data are provided only after the request is cleared by the Information Security Unit of Goskomstat regarding its compliance with law.

Access to microdata by researchers

There is great interest in microdata from researchers.

When nominative information on enterprises is not important for the purposes of research, files with microdata precluding the possibility of identifying individual enterprises can be disseminated for research work.

Such an approach is used for dissemination of results of the national survey of welfare of households and of their participation in social programmes. Disseminated information includes documentation of the project, aggregate data and processed microdata, which preclude identification of respondents.

Similarly, beginning in 2005 it is planned to provide researchers with anonymous microdata of household budget surveys.

Methods of providing microdata to users are also considered within the framework of the implementation of the project on measurement, monitoring and analysis of poverty carried out by Goskomstat and the Ministry of Labour, supported by the World Bank and by the Department for International Development of the UK.

Ensuring confidentiality of statistical data at both aggregate and microdata levels implies the implementation of a number of technical and programming measures for information protection.

Data processing and confidentiality

Statistical methods for ensuring confidentiality include aggregation of microdata, adjustment of the levels of detail in the classification structure of data, cancellation of individual data in statistical tables and modification of data. An optimal combination of different methods depends on specific characteristics of the source information and on conditions of data processing.

It is difficult to ensure confidentiality of source data in small groupings, which is often the case with regional data.

Excluding nominative information from microdata is effective, in terms of achieving confidentiality, when data are aggregated over a relatively large number of enterprises of comparable size. However, when there is one large, for example metallurgical, enterprise in the region, concealing its nominative information does not help in preventing identification of this enterprise and of the information related thereto.

This problem should be taken into consideration when transmitting microdata to users. An effective solution might be the application of special rules for aggregation of microdata, such as the "threshold rule" and the "rule of domination".

Special computer software is often needed to facilitate complex and labour intensive methods for ensuring confidentiality of statistical data.

Main problems in ensuring confidentiality

The following can be referred to as main problems in ensuring confidentiality of statistical data in the Russian Federation:

- absence of a special law on statistics, which would provide a legal basis for the confidentiality principle;
- existence of clauses in various legislative acts, according to which certain government bodies have the right to request information. In the absence of a law on statistics, references to the principle of confidentiality of microdata are not accepted by these bodies;
- the role of statistics is not fully understood by some representatives of the executive branches of power, particularly at the local level, who tend to regard statistical data as an information means for administering the territory or even for regulating activities of individual enterprises;
- unresolved technical aspects of data transmission over telecommunication lines.

II.4 STATISTICAL DATA CONFIDENTIALITY - GEORGIA

Supporting paper by Teimuraz A. Beridze, State Department for Statistics of Georgia

Personal data confidentiality is a vital element of a statistical system and is one of the main principles of official statistics reflected in the Resolution of UNECE "Main Principles of Official Statistics in the ECE Region" (15 April 1992, 47th session of UNECE).

In a democratic country and civil society, statistical data confidentiality is considered to be a fundamental principle of relations among producers (Statistical Offices) and respondents (suppliers of personal data). The principle aims are:

- inaccessibility of personal data and protection of personal, commercial and state secrets;
- strengthening of respondents' trust in the statistical system.

The introduction of the confidentiality principle in a country is characterized by various peculiarities specific to that country. Firstly, it is necessary to prepare strict legal grounds for the implementation of such a principle.

The Georgian legislative basis for official statistics regulates all relations of the SDS (State Department for Statistics of Georgia) with providers of personal data and also with users of statistical data. Georgian Law 'on statistics' constitutes rights and obligations not only for producers of statistical data, but for suppliers of personal data (primary statistical data) and for users of statistical data. This law also guarantees the protection of personal, commercial and state secrets reflected in the primary statistical data provided by legal and physical persons.

Personal data confidentiality is regulated also by the General Administrative Code of Georgia and partly by the Georgian Law 'on population census'. This law ensures protection of the data obtained through population censuses. Publication and dissemination of only summary data on population censuses guarantees the protection of the constitutional rights and freedom of citizens.

In order to strengthen trust in primary statistical data suppliers, the public are informed about the confidentiality of the personal data provided. All questionnaires for official statistical observations include a special note on the twelfth item of the Georgian Law 'on statistics' entitled 'Statistical Data Confidentiality'.

Meanwhile, some difficulties related to the implementation of the confidentiality principle should be noted. The first issue is statistical data accessibility for users. Some users (mainly state authorities) consider that data accessibility means personal data accessibility. Therefore, they often require primary statistical data on various legal and physical persons.

Is it possible to protect personal data confidentiality in all cases? Over the last four years, there was serious discussion between the SDS and the State Antimonopoly Organ regarding the contradiction that existed between the Antimonopoly Legislation and the Georgian Law 'on

statistics'. According to the Antimonopoly Legislation, the State Antimonopoly Organ had the right to obtain confidential data from all official organs, while the Georgian Law 'on statistics' constituted a guarantee on primary statistical data confidentiality, but there were fixed legislative exceptions. This contradiction no longer exists. On the legislative initiative of the SDS, the Parliament of Georgia approved and the President of Georgia signed the Georgian Law 'on amendment to the Georgian Law 'on statistics'' (see Box).

The protection of personal data confidentiality is a problematic issue for local statistical offices of the SDS, because some heads of local official authorities often require primary statistical data. As we know, the national legislation in European Countries constitutes two types of sanctions for offences related to personal data confidentiality: fine and imprisonment.

Usually, the SDS conducts various working meetings and seminars on confidentiality for users of statistical data (representatives of legislative and executive powers, mass media, non-commercial and commercial legal persons, etc.). The last press conference held for mass media in May 2002 on the theme 'General Census of Population, 2002' dealt with the confidentiality issue too. Poor means of communication and a low level of technological equipment in local statistical offices can result in imperfect protection of primary statistical data confidentiality.

As for technical and technological means in the central statistical office, the SDS regularly conducts measures for personal data protection. The questionnaires for statistical observations are anonymous in the stage of processing; data are located in Intranet, where internal as well as external access and movement are strictly regulated.

There are some cases in the Georgian economy related to scarcity of economic entities in a concrete type of economic activity (1-3 units), when the SDS does not publish and disseminate statistical data for total type of such activity because of data confidentiality protection. The only exception is made with official consent from the economic entity concerned on the publication and dissemination of its personal data. Altogether, we have 121 such cases (35 percent of the total number of economic activities) with no official permission and therefore data on those types of activity are not published.

Today, the SDS continues activities to ensure personal data confidentiality protection. The following acts (legal acts and instructions) are being drafted:

- on duties of statistical offices regarding primary statistical data confidentiality protection;
- on personal responsibilities of the staff;
- the list of SDS staff with the right to access confidential data;
- instruction for activities in the Intranet of the SDS;
- instruction for activities on documents with confidential data;
- instruction for the creation of reserve copies and maintenance of electronic archives;
- on responsibilities of the Intranet administrator;
- on responsibilities of database administrators;

- on confidentiality protection at the time of statistical data collection, processing and dissemination.

An interesting issue is connected to access to primary statistical data for scientists and researchers. The experience of European countries in this regard will be interesting and useful for us.

The SDS is drafting the 'Oath of Statistician', which also includes protection of primary data confidentiality by staff, a very important issue.

BOX

GEORGIAN LAW

on amendment to the Georgian Law
'On Statistics'

Item 1. 12th item of Georgian Law 'On Statistics' [Parliament acts, N46, 03.12.1997] is edited as follows:

"Item 12. Statistical data confidentiality
1. Data acquired for statistical purposes are confidential, if they enable the identification of the respondent.
2. Confidential data may be used only for summary statistical data preparation, excluding such cases where a respondent assents to use of his/her data for other purposes.
3. Only staff of official statistics are allowed to work on confidential data.
4. Dissemination of official statistical materials, which contain confidential data or enable the definition of such data, is prohibited."

Item 2. This law enters into force since its promulgation.

President of Georgia
Eduard Shevardnadze

Tbilisi
December 25, 2002

II.5 PROTECTION OF CONFIDENTIAL DATA IN PRACTICAL WORK OF STATE STATISTICAL BODIES OF UKRAINE

Supporting paper by Olexander Osaulenko, State Statistics Committee, Ukraine

Introduction

Over the last several years the state statistical bodies of Ukraine have made considerable progress in legal, organizational and technical protection of confidential data. This report presents a summary of the main results achieved, the problems that have arisen and plans for the future.

Legislative support of protection of confidentiality of data

The consecutive integration of Ukraine into the economic, cultural and information environment of the European Union implies the adaptation of national legislation to the modern European legal system. Harmonization of statistical legislation with international legal norms and standards is an integral part of this process. Confidentiality of statistical information relates to basic regulations of legal documents governing state statistical activity in Ukraine.

Standards of statistical data confidentiality within the national legislation

In national legislation, a definition of confidential data was stipulated, first of all in the Law of Ukraine 'On information' adopted in 1992. According to this definition, confidential information is 'information which is in possession, use and disposal of separate natural and legal persons, and disseminated according to their desire and conditions established by them'. In line with the Regulation on technical protection of information in Ukraine, adopted by the Decree of the President of Ukraine in 1999, 'confidentiality is a feature of information to be protected against unsanctioned familiarization'.

Subsequently, national statistical legislation has further strengthened the principle of confidentiality in relation to statistical data. Thus, the Law of Ukraine 'On State Statistics' adopted in 2000 stipulates that primary data obtained by state statistical bodies from respondents when carrying out statistical observations is the confidential information that is protected by Law and used only for statistical purposes in depersonalized form. Dissemination of statistical information, which could allow access to confidential statistical data concerning a particular respondent, is prohibited. Statistical information obtained by state statistical bodies in the process of statistical observations cannot be requested by the government authorities, local executive authorities, other legal persons, public associations, officials and other persons in order to be used for making decisions with respect to particular respondents.

The Law of Ukraine 'On State Statistics' also specifies statistical information that could be disseminated by state statistical bodies. In line with international standards, depersonalized statistical information in a disaggregated format is considered to be information that does not permit the identification of confidential information about a respondent.

In particular, such information as names, addresses, telephone numbers, type of activity of enterprises and organizations are not considered as confidential. In line with this, as well as according to the Regulation on Unified State Register of Enterprises and Organizations of Ukraine (USREOU), maintained by state statistical bodies, users are provided with restriction-free information on the name of an enterprise, its location, post code, area code and locality, telephone and fax numbers. All other information not covered by this list is considered as confidential data of the USREOU.

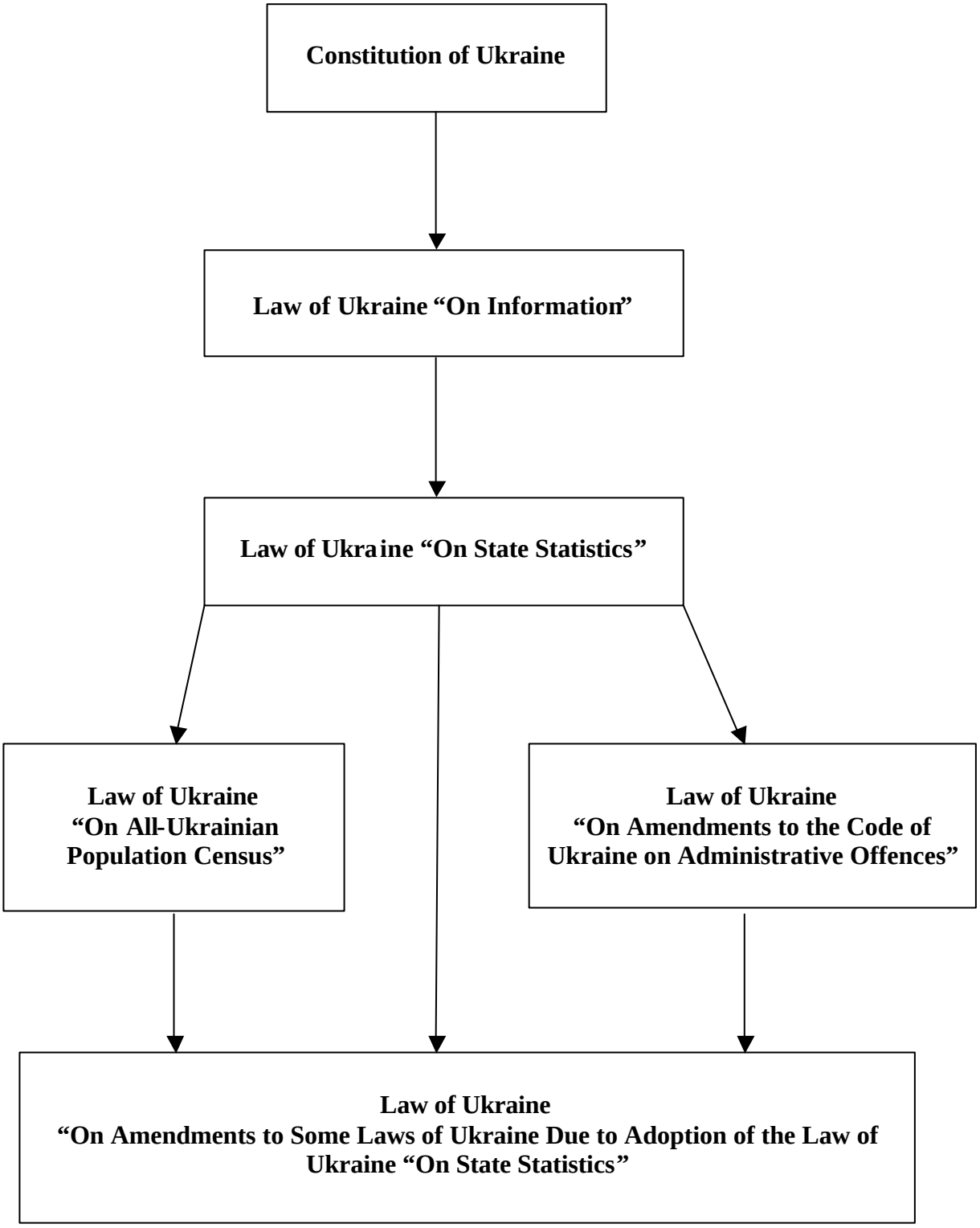
The organization of a number of statistical observations with regard to respondents – natural persons (population census, household living conditions survey, labour force survey, household survey on agricultural activity, etc.) is regulated, in the first place, by the Constitution of Ukraine adopted in 1996, which set up strict restrictions in terms of the possibility to access identified data on natural persons. Article 32 of the Constitution of Ukraine says that 'it is prohibited to collect, store, use and disseminate confidential information about a person without his/her agreement, except in cases specified by the law, and only for the interest of national security, economic welfare and human rights'.

The Law of Ukraine 'On State Statistics' indicates that employees of the state statistical bodies are obliged to observe the requirements for protection of confidential information about physical persons collected in the course of statistical surveys.

The guarantees for protection of confidential information about an individual are also anticipated in the Law of Ukraine 'On All-Ukrainian Population Census', which was adopted in 2000. According to this Law, primary data and other information obtained during the Census taking are confidential and are protected by the Law. Records in census questionnaires concerning respondents are not subject to dissemination without the respondent's consent and are used in aggregated depersonalized form only for statistical purposes. The primary census data cannot be requested by the court, public prosecutor's offices or other authorities to be examined and used as evidence in civil and criminal cases. Completed census questionnaires and other census documentation comprising primary data are subject to archive storage on premises inaccessible to unauthorized persons. After the term for storage of census documents is expired, this information is destroyed in accordance with established procedure.

It should be mentioned that the norms listed above not only meet the requirements of international law in the field of personal data protection, but also create a more trustful attitude on the part of respondents to statistical observations. Such an approach is a key one in establishing transparent and constructive cooperation of all stockholders in implementing state statistical activity. Thus, it is well known that even in cases where the law stipulates an obligatory participation of the population in a statistical observation (e.g. in Ukraine, participation in the Census is obligatory for respondents), reliable data can be expected only if the citizens have confidence in the observation. Such an attitude should be based on respondents' confidence in the

Legislative Support System for Protection of Confidential Statistical Information in Ukraine



fact that confidentiality of information provided by them would be ensured. In this case, as practice shows, a wide-scale informing of respondents about the responsibility for inobservance of requirements for statistical data confidentiality set by legislation is also a very convincing argument.

Liabilities for violating the standards of confidentiality

The Code of Ukraine on administrative offences stipulates the following liability for violation of the procedure for using confidential statistical data:

- for citizens: in the form of a fine amounting to five to ten untaxable minimum incomes of citizens; for officials, including employees of state statistical bodies: up to the amount of ten to twenty untaxable minimum incomes of citizens.

Stages for implementing the legislative standards for confidentiality

Regarding the work undertaken by the state statistical bodies of Ukraine in the sphere of legal protection of statistical data confidentiality, the main stages as described below can be specified.

The first stage covers the period during which the legislative basis providing legal guarantee for securing confidential statistical information provided by the respondents to state statistical bodies has been developed. At this stage, general laws regulating the implementation of state statistical activity in Ukraine have been drafted and then adopted by the Parliament of Ukraine. These laws include the Law 'On state statistics', the Law 'On All-Ukrainian Population Census' and the Law 'On Amendments to the Code of Ukraine on Administrative Offences'.

During the second stage, work was carried out to bring some legislative acts in line with the above-mentioned laws. Thus, the Law of Ukraine 'On Amendments to Some Legislative Acts of Ukraine Due to Adoption of the Law of Ukraine 'On State Statistics'' has made amendments to 16 effective laws which, in one way or another, regulate activities related to statistical activity and information relations arising from this process.

Today, the state statistical bodies of Ukraine are in the third stage of development of legal protection of confidential data, which obviously is the most complicated stage and requires long painstaking work. The efforts of state statistical bodies are now focused primarily on organizing relevant training for both respondents and data users, in terms of their perception of legislative changes made, as well as on issues of practical implementation of these innovations.

Ways and problems to implement transformations

Attention must be paid to the fact that the reform of legislation regulating state statistical activity in transition countries can be undertaken in two ways. The first is a method of radical changes in statistical legislation and a rapid and complete adaptation of national legislation to

corresponding international standards. The second method is evolutionary, with a gradual approaching of national statistical legislation to international standards in combination with gradual adaptation of all other legislative acts, regulating activities related to official statistics, to statistical legislation.

In Ukraine the first approach involving radical changes, which looks more effective, has been chosen. However, as experience has shown, this approach contains a number of drawbacks which could seriously affect the quality of results to be achieved.

Thus, precipitate amendments in national legislation could lead to certain legal collisions. In Ukraine, for example, one obvious legal contradiction is observed, i.e. on the one hand, state statistical bodies according to effective legislation are obliged to observe the confidentiality of statistical information, but on the other hand, in order to observe such laws of Ukraine as 'On Prosecutor's Office', 'On Police', 'On investigating activity', 'On Organizational and Legal Basis to Combat Crime', 'On Security Service of Ukraine', statistical bodies have to provide law enforcement authorities, based on official written requests from them, with all data from statistical returns needed for the work of those authorities (including confidential information). Such a situation is explained by the fact that amendments to legislative acts regulating the activity of law enforcement authorities, especially changes of a restrictive nature, are quite difficult to pass due to the well-known specificity of this area of government regulation.

Apart from this, there are serious difficulties in the practical implementation of legislative changes in a statistical area if such implementation is precipitated. Both respondents and primary data users cannot adequately comprehend the essence of change, i.e. they find themselves unready for such radical changes, especially where strict observance of statistical data confidentiality is concerned.

The state statistical bodies of Ukraine also face other problems in ensuring data confidentiality in terms of legislation. In particular, one such problem is the absence of normative and legal regulations to define the types of aggregated indicators not subject to dissemination due to the confidentiality issue. It is also necessary to identify criteria for dissemination of data on small administrative areas (rayon level, settlements) with respect to those features that are presented in the observed population by only one respondent.

Organizational and technical support for protection of data confidentiality

The guarantee of statistical data confidentiality, apart from the creation of a legal basis, also implies a number of organizational and technical measures, which should cover the whole process from data collection to their destruction.

Arrangements

While preparing the staff to carry out statistical surveys, each employee of statistical bodies of Ukraine dealing with personal data studies his/her duties and responsibility established for non-securing the confidentiality of information obtained, as well as the rules for the protection of this data against inadvertent disclosure. One such rule is a ban on providing any other person with documents including confidential information or on discussion of the information provided by respondents with unauthorized persons, or leaving completed questionnaires lying around, and so on. In this connection, one important task is the preparation, adoption and introduction in statistical practice of normative and methodological documentation on techniques to be used to protect statistical data against disclosure.

For example, statistical tools for household and natural persons surveys are developed in such a way that information on their identification characteristics in the course of data processing is presented in the form of a special system of codes within a primary area survey unit, which effectively ensures depersonalization of personal data. This avoids the unauthorized use of confidential information from magnetic tapes and in electronic format.

Special software and hardware

Modern systems of automatic processing and dissemination of statistical data, based on the use of local and global networks, Internet and distributed and local databases, considerably contribute to the risk of loss of data or data disclosure. Therefore, the issue of program and technical protection of information circulating within the information system of statistical bodies requires permanent attention and resolution so as to ensure effective protection of confidential data. The measures to be taken for this purpose are complex ones and should foresee the necessary organizational actions and application of relevant programs and technical tools.

In line with the concept of informatization of state statistical bodies adopted in 2000, the State Statistics Committee of Ukraine conducts purposeful work on the application of programming and technical tools for data protection, ensuring data protection against distortion and destruction, as well as against unauthorized use of data.

Autonomous local networks providing access to databases through separate servers using a relevant system of passwords, thus eliminating access to the data from the outside, are developed for the purpose of automated processing of confidential information. In order to ensure sound data protection, special system tools and programs are used, in particular:

- establishment of service units for administration of networks, databases and data security with clear identification of their functions on the basis of relevant regulations on these units and job descriptions;
- classification of users of the automated system by set of information, which prevents different users from working with the same set of data;

- allocation of a unique code to each user and provision of passwords for authorized access to users, coding of users' passwords (number of symbols in password code and password itself could be different, thus preventing anyone who may possess the password code, even a system administrator, from identifying the original password);
- restriction of users' possibilities in employing only those technical operations which are specified for the corresponding user category (e.g. only users from the group "scanning operator" could start modules for scanning and recording of document images);
- registration of all users' requests to access databases;
- central storage of copies of primary data at a separate server, putting a ban on downloading information from databases into PCs, excluding cases requesting implementation of concrete tasks linked with the technological process.

The main focus is on use of tools for data protection, which are an integral part of the network and local operation systems, and systems for database management; programming and equipment components of networks that allow to differentiate users' rights and control access to the data. Antivirus software is widely used to protect servers, working stations and mail systems. Separate local computer networks, which are located in protected premises with restricted access, are used for processing and storage of information requiring restricted access.

In order to protect data from physical damage, first of all, databases and other more valuable information are copied onto removable magnetic tapes that are stored in separate premises protected from unauthorized access. When the required information is restored, this could lead to the loss of information due to differences in handling procedures. In the state statistical bodies of Ukraine, it is planned to use in future specialized data storage systems based on Storage Area Network technologies (SAN), which minimize the loss of information and time involved in the process of data restoration.

The key objectives for the near future

The following key objectives are set for the state statistical bodies of Ukraine to ensure the protection of data confidentiality in the near future:

- Legislative support:

  - to ensure the synchronization of standards for confidentiality laid down within the national legislation. This primarily concerns the coordination of legal acts that regulate the activity of law-enforcement bodies on the one hand, and the state statistical activity on the other hand;

  - to develop and implement the measures aimed at increasing the confidence of respondents in state statistics with regard to confidentiality standards, so that users can also understand the need to implement these standards and the corresponding adjustment of these approaches to the work with statistical data.

- Organizational and technical support:

- the preparation, approval and implementation of standard and methodological materials (rules, instructions, documentation, etc.) that regulate the practical work of the employees from the state statistical bodies to ensure the protection of data confidentiality;

- the implementation and exploitation of the specialized software and hardware tools that could identify and eliminate the consequences of network attacks, identification of the vulnerability of operational systems and database management systems; data coding tools for exchange of information, tools for monitoring the state of information resources, etc.

References

Baranov, Bryzhko, Bazanov, "Human Rights and Personal Data Protection", State Committee of Communication and Informatization of Ukraine, Kyiv, 2000

"Improvement of Statistical Legislation of Ukraine", Information and Publishing Center of the State Statistics Committee of Ukraine, Kyiv, 2002

Statistical Data Confidentiality. Papers to UNECE/Eurostat Joint Work Session, 14 – 16 March 2001

II.6 STATISTICAL CONFIDENTIALITY - POLAND

Supporting paper by Tadeusz Toczynski, Central Statistical Office, Poland

Historical outline and current legal regulations

The year 2003 is a special year for Polish statistics. In the middle of the year the Central Statistical Office (CSO) as a body of official statistics will celebrate the 85th anniversary of its establishment.

As is well known, one of the basic canons of statistics is statistical confidentiality.

I have to note, to my satisfaction, that the problems of confidentiality, and the associated problems of protection of statistical data, are emphasized in all legal regulations with the status of a legal act and which define the functioning of statistics in Poland.

The first law, dated 21 October 1919, on the organization of administrative statistics, Article 4, determines that the data provided according to the obligation imposed by the law shall be used only for statistical purposes and shall not be available for other purposes, either to public authorities or to private persons. Officials participating in statistical surveys were personally responsible for the strict implementation of this provision.

Article 12 of the Decree on the organization of state statistics and on the Central Statistical Office dated 31 July 1946 also contains a provision determining that individual data obtained from housing and population censuses, as well as from other statistical surveys according to the obligation imposed on private persons and institutions, may be used exclusively for statistical compilations. These data shall not be available for other purposes either to public authorities or to private persons. Officials participating in statistical surveys are personally responsible for the strict implementation of this provision.

In the subsequent law dated 15 February 1962 on the organization of state statistics, the provision pertaining to statistical confidentiality was strengthened, stipulating that any use of information and individual data obtained from population and housing censuses, as well as from other statistical surveys for purposes other than statistical compilations, is forbidden on pain of criminal responsibility.

Similarly, in the next law dated 26 February 1982 on state statistics, Article 24 stipulates that it is forbidden to use individual information obtained from population and housing censuses, as well as from other statistical surveys, for purposes other than statistical compilations and analyses.

It should be emphasized that in the period from 1946 to 1989, such provisions of law covered survey data provided by citizens and private enterprises including individual agricultural holdings, whereas individual data from the public sector enterprises were published and made available through statistical services. This concerned the state sector exclusively.

The situation has changed substantially in connection with obligations adopted under Art. 92 of the Europe Agreement, establishing an association between the Republic of Poland on the one hand, and the European Communities and their Member States on the other, (signed on 16 December 1991), and under which Article the Government adopted obligations to ensure the development of the effective statistical system, as well as to ensure data confidentiality. The principle of individual data confidentiality and protection as provided for in the Resolution of the United Nations Organization, also called the Statistician's Decalogue, is fully reflected in the binding Law on Official Statistics of 29 June 1995. This Law provides for a number of provisions ensuring the protection of individual and personal data collected in statistical surveys where personal data are defined as data pertaining to particular natural persons including various aspects of their lives, and individual data as data coming from or pertaining to particular business entities. At the same time, the Law imposes on statistical services an obligation to inform business entities about surveys to be conducted, as well as an obligation to observe the principles of statistical data confidentiality.

Among the provisions of the Law in force that relate to the confidentiality of individual data and to the obligation of their protection, special attention should be paid to the following articles:

- Article 10, stipulating that individual and personal data collected in statistical surveys of official statistics shall be confidential and subject to special protection; the data shall be used exclusively for statistical calculations, compilations and analyses, as well as for the creation by official statistics services of sampling frames for statistical surveys conducted by the services; making available or using individual and personal data for purposes other than those specified above shall be prohibited (statistical confidentiality);
- Article 38, stipulating that any individual data obtained from official statistics surveys shall not be published or made available. Statistical information which is obtained from official statistics surveys and which can be linked with or can identify particular persons, as well as individual data characterizing economic activity of business entities, shall not be published or made available. This refers in particular to data aggregations that consist of fewer than three entities or to those in which the share of one entity is higher than three-fourths of the total;
- Article 39, obliging the CSO President to ensure that the method used to store collected statistical data guarantees observance of the principles of statistical confidentiality;
- Article 54, stipulating that anyone who violates principles of statistical confidentiality shall be subject to imprisonment of up to 3 years;
- Article 55, stipulating that anyone who, in order to obtain material or personal profits, uses statistical data he/she has been acquainted with while performing his/her work or tasks to order of the organizer of an official statistics survey, is subject to imprisonment of up to 5 years.

Apart from the above-mentioned provisions of the Law directly pertaining to the subject in question, attention should also be paid to other provisions of the Law that, according to the intention of a legislative body, have been supposed to create appropriate mechanisms to serve proper implementation of principles of statistical data confidentiality. They include the following:

- Article 2, subparagraph 6, stipulating that working on statistical data consists of creating data files, after separating or coding the information which would allow identification of particular business entities or natural persons, and then, on the basis of these files, making calculations, compilations and analyses;
- Article 11, imposing on the official statistics services which conduct statistical surveys a duty to inform data respondents about the rights and obligations of entities providing data for statistical surveys, and also about the guarantee that the principles of statistical data confidentiality shall be observed;
- Article 12, stipulating that the staff of the official statistics services, census enumerators, statistical interviewers, and other persons having direct access to individual and personal data shall observe without exception the principles of statistical confidentiality and shall be allowed to perform their tasks only after delivering a written oath worded as follows: "I hereby take the oath that I shall perform my tasks on behalf of official statistics with due care and diligence, in accordance with the professional ethics of a statistician and I shall not reveal to third parties individual data I have obtained during the performing of those tasks";
- Article 13, establishing the principle of one-way flow from state administration bodies, local authorities and other governmental agencies, as well as bodies maintaining official registers of administrative data containing individual data in the form of, among others, extracts from registers and copies of data files, collected declarations, registration documents and other official forms, and data from computerized systems databases. This principle has been established and strictly obeyed. The reverse situation shall not occur under any circumstances, i.e. requests for making available individual data collected from surveys and in the possession of the official statistics services. This limitation does not apply to two separate official registers that are kept by the services: the register of business entities and the register of territorial division of the country;
- Article 14, stipulating that data collected from official statistics surveys shall be made available exclusively in the form of compilations and analyses made on the basis of collected individual data;
- Article 35 which authorizes the official statistics services to collect for statistical purposes and for the preparation of demographic projections the following strictly defined data on natural persons residing within the territory of the Republic of Poland: first name and surname, gender, date and place of birth, citizenship, marital status, place of residence.

These data shall be collected in the form of responses provided directly to the official statistics services by the person concerned or an adult member of the household, or can be derived from administrative records; the scope and form of collecting personal data necessary for a given survey are defined each time by the program of statistical surveys of the official statistics and, in the case of population and housing censuses, by a separate law.

Data other than those defined above which would allow identification of a particular natural person they refer to may be collected by official statistics services exclusively:

- when the program of statistical surveys indicates as a source of statistical data a document containing personal data that cannot be separated in a simple way for the transmission for statistical purposes; or
- when identification is necessary for generalization of the statistical surveys results.

The storage in one data file or database of all personal data collected by official statistics services from various statistical surveys that concern a particular natural person and that, when combined, could be used for characterizing and evaluating that person shall be prohibited.

The name, surname and place of residence of a given natural person shall be excluded from the generalization of personal data when entering them into the computerized systems databases of official statistical services. These data may be entered only into a sampling frame used for statistical surveys conducted by the official statistics services. This regulation establishes the limits protecting the data from an excessive interest of statistics in collecting and storing personal data.

Current Regulations of the Law on official statistics do not provide for:

- any exceptions for making individual data available. This also applies to the requests made by the court in the course of penal proceedings;
- making statistical data concerning a given entity available with that entity's consent;
- special principles of handling protected statistical data for scientific purposes;
- possibility to protect available data through statistical confidentiality after a fixed period of time.

Principles and procedures for handling statistical data

In order to implement the guarantee of statistical confidentiality and protection as discussed in the first part of this paper, it is necessary to undertake a very wide range of practical activities which shall ensure real protection of personal, as well as individual, data.

For the CSO, the activities were defined in detail by the CSO President in "Principles and Procedures of Handling Statistical Data", to be applied by all agencies of the official statistics services.

This regulation defines in particular:

- principles and procedures of handling statistical data at the stage of data collection by statistical interviewers, via mail, fax; the method of data reception by statistical agencies; transmission of reports within statistical agencies and for registration;
- methods of registering, editing and creating statistical data files in statistical offices including the principles and the level of data protection against unauthorized access, loss, destruction or falsification, as well as handling of reports after data have been edited;

- protection of places of data storage and processing including the separation of protected areas and safety zones with appropriate protection and control of entries and exits;
- principles of creating, maintaining, storing and updating of national files and databases, including the protection of national files and databases on servers, access to data in the network of a given unit, as well as statistical units' wide area network (WAN);
- obligation to label data while creating the files for storing aggregated data which must remain inaccessible because of the risk of indirect identification;
- principles of establishing levels of aggregates in working and result tables, including the creation of the working tables for analytical purposes, and the result tables and tables for publication;
- principles of protecting the results and the level of detail preventing indirect identification, as well as principles of labelling the protected data in cases where there are fewer than three entities in the aggregate and when one entity is dominant;
- procedures in case of the possibility of indirect identification when there is only one protected aggregate at a given level among all aggregates that make up a higher level aggregate, which means that the unit, conducting a survey and making assumptions for reckoning result tables to be widely available, assumes that the second aggregate of the lowest value out of component aggregates of a higher level shall be hidden, or the aggregate data are combined with the preceding or next aggregate, or all aggregates making up the aggregate of a higher classification level shall be hidden;
- adopted principles of making unidentifiable statistical data available to scientific centers (excluding data describing the economic position);
- making generalizations and analyses to individual orders to which all above-mentioned regulations limiting access to individual data shall apply;
- storing and keeping the data, as well as the safety copies on computer carriers, protection of data carriers, computers, and premises.

These principles are regularly updated and their application is subject to constant control due to the supervision of the CSO President over subordinate units of the official statistics services.

In cases of doubt regarding the possibility of making available statistical data both collected in surveys covered by the program of statistical surveys and those prepared upon request, the decisions of the CSO President are supported by the Commission of Statistical Confidentiality headed by the CSO Vice-President and consisting of experts who are also employed in statistics.

The decisions adopted by this Commission apply to specific cases and, at the same time, serve as precedents helpful in defining the possibility of making available data in cases analogous or similar to those already examined by the Commission.

Planned supplementary regulations on data confidentiality and protection

Currently, draft amendments to the binding Law on Official Statistics are being prepared. A number of issues have already been discussed. The amendments shall refer to a wider range of

problems than data confidentiality and protection. Nevertheless, new solutions will concern this subject as well. So far, the article regulating access to anonymised individual data for scientific purposes has already been introduced into the draft in question. In this field, solutions have been adopted similar to those provided for in the Regulation Commission (EC) No 831/2002 dated 17 May 2002. In particular, it is possible to make individual data from the following surveys available to higher schools and scientific and research institutes:

- households' budgets;
- living conditions of the population;
- economic activity of the population and labour force;
- continuing education;
- innovations.

The data shall be available upon written request defining the scientific purpose of their use. An additional provision has been introduced stipulating that the scope of availability for national scientific purposes shall not be narrower than the scope of the availability planned in Eurostat.

We are also considering the possibility of adding a provision ensuring that government administration bodies that conduct analyses and which, in line with the Law on Official Statistics, are also authorized to independently conduct statistical surveys, have free access to unidentifiable individual data from social statistics surveys. This pertains to absolutely anonymous series of records from surveys on living conditions and households' budgets, i.e. individual records that are properly modified and that cannot in any way be linked to a particular person. This issue has not been settled yet.

The next issue to be worked out is a possible introduction to the Law of regulations concerning the rights of the exporter or importer to submit to the CSO a motion for a data compilation that would make indirect identification impossible, and the data dissemination would not violate the principles of data confidentiality in the movers' own interest (passive confidentiality).

There are no plans for extending the scope of regulations stipulating that certain data may be provided by official statistics services with the consent of the entity from which they are coming or to which they are referring.

The intended amendments to the Law on Official Statistics have been currently under consideration. Further decisions and social consultations will finally define the scope of these amendments.

II.7 STATISTICAL DATA CONFIDENTIALITY - KYRGYZSTAN

Supporting paper by Zarylbek Kudabaev and Nataliya Gudkova, National Statistical Committee of the Kyrgyz Republic

Importance of statistical data confidentiality for statistical offices

All statistical offices of transition countries admit that the protection of statistical data is important. Ensuring primary data confidentiality is one of the Fundamental Principles of Official Statistics adopted by the Statistical Commission of UN in 1994. Under the democratization of society, statistical data confidentiality becomes a basic principle of cooperation between statisticians and the public, i.e. statistical data providers.

There are two main objectives in ensuring primary data confidentiality: firstly, protection of privacy and non-disclosure of state and commercial secrecy; and secondly, strengthening the confidence of users in official statistics.

Kyrgyzstan was confronted with the problem of data confidentiality only in the '90s, unlike western countries who have been dealing with the problem since the '80s and even the '70s. For Kyrgyzstan, on the one hand this was due to economic, social and legislative changes, and on the other hand it was due to the development of information technologies and communication, increased use of personal computers, databases and networks. At present, statistical offices are not under the same pressure to present microdata as is the case in western countries. Nevertheless, one can suppose that as the technologies and markets develop, the demand for more detailed data will also increase.

The Population Census of 1999 and the Agricultural Census, which is being conducted at the present time, may draw increased attention to the problem of data confidentiality. Such vast information collection as takes place during the census inevitably causes public concern about confidentiality. At the same time, the census is a unique opportunity to discuss the issues of confidentiality at the national level, to reconsider the principles of data confidentiality protection, to evaluate public opinion about data confidentiality and to prove that statistical offices fulfil a promise to protect confidential data.

Each country solves the problem of applying the principle of data confidentiality in its own way. This paper outlines the main aspects of application of the confidentiality principle and problems deriving from it.

Legal setting

In recent years, more attention has been paid to all aspects of confidentiality, perhaps primarily to the legal and administrative aspects. Obviously, the principle of confidentiality can be implemented through the strengthening of the legal basis. In accordance with the law, the State Statistical Service of the Kyrgyz Republic cooperates with data providers and data users.

The Law of the Kyrgyz Republic "On State Statistics" regulates rights and liabilities not only of those who collect, process and publish statistical data, but also of those who submit and use statistical data.

Thus, the Law guarantees the confidentiality of data provided by legal and physical entities. In accordance with the Law, statistical offices at all levels of the Republic are responsible for the disclosure of commercial and state secrecy, as well as individual information. At the same time, the Law determines rights and duties of persons submitting primary information to statistical services.

Another legislative act of the Kyrgyz Republic, i.e. the Code on Administrative Responsibility, regulates rights and duties of those who are involved in the production of statistical information. Observance of data confidentiality is one of the fundamental principles of the Code on Professional Ethics of Civil Servants adopted in the system of state statistics of the Republic.

Other legislative acts in the field of official statistics also contain the principle of confidentiality. These legislative acts are the Laws of the Kyrgyz Republic "On Population Census" and "On Agricultural Census". These laws guarantee the confidentiality of individual data obtained from censuses. Publication of summary results only in an aggregated type avoids the violation of constitutional rights and freedom of individuals.

General policy of statistical offices in the field of confidentiality

Some special issues of general policy in the field of data confidentiality of state statistical services are being studied. These issues are access to primary data, and study of public opinion. Statistical offices better understand the importance of the position of the public and the respondents in regard to statistical data confidentiality. To strengthen confidence in state statistics, individuals have been informed about the confidentiality of primary data. All types of statistical reporting documents and questionnaires contain a reference to the corresponding article of the Law on State Statistics.

To encourage the participation of households in surveys and increase their interest in providing reliable information, the National Statistical Committee uses material incentives and guarantees the confidentiality of the information obtained. At the same time, the National Statistical Committee carries out information exercises with the public, providing information on the objectives and importance of conducting surveys. This results in a high response rate (the share of those who refuse to respond to surveys is valued at nearly 1-1,5% in a year). The statistical data collected on citizens are used only in aggregate or anonymous form, with no indication that could be used to identify individuals. Primary data on private individuals and families may not be published without their consent.

Access to impersonal primary data can be granted to research centers only for research purposes, as well as to Government institutions, and fiscal and other state authorities in particular cases. Consideration of any data on enterprises as exclusively confidential may result in problems

when we distribute such data. A certain company is often the only, or one of the few, producers of a certain product. This problem exists in the Kyrgyz Republic because selected sectors of the economy are represented only by a few enterprises.

Due to the lack of special software to protect data confidentiality, we use our own software or protect data manually. We recode data into wide categories, displace them and implement micro-aggregation. Special measures on data protection must be taken in the transmission of primary data through software networks. In recent years, on-line data collection has increased. Distribution of data through software networks is increasing.

Within the framework of technical and technological policy, the National Statistical Committee ensures the protection of statistical data from unauthorized access. The data on local networks can only be accessed by personal passwords. The software PROXY-Server was installed to protect the local (internal) network from external users.

Measures to protect confidential data

The most simple way to avoid the violation of data confidentiality is not to give microdata to users. Administrative and organizational measures are the main means of confidentiality protection, together with general measures and only special staff have access to confidential data.

Other measures frequently taken are access under controlled conditions, and access to data only by researchers whose names appear in contracts. The person storing the data is responsible for the provision of its confidentiality. In many cases, third parties must define the reason for wishing to access data, and statistical offices may refuse to grant access to primary data.

Main problems of observing the confidentiality principle

There are certain problems attached to the application of the principle of confidentiality.

The Law on Statistics guarantees the principle of confidentiality, but legislative acts of other institutions contradict the Law on Statistics. Such legislative acts oblige statistical offices to provide individual and confidential data to other institutions. That is why it is extremely important to coordinate the law-making process in different fields and institutions. Such a problem exists connected with the Antimonopoly Service, which requests data on monopoly enterprises with some statistical indicators explaining that it is needed for the control of activities of monopoly enterprises. Thus, article 13 of the Law of the Kyrgyz Republic 'On Monopoly Restriction, Protection and Development of Competition' obliges enterprises and state authorities including the National Statistical Committee to provide reliable information to Anti-Monopoly organs of the Kyrgyz Republic so that they can carry out the control of monopoly enterprises. At the same time, statistical offices of the country are responsible for revealing the confidentiality of data provided by legal and physical entities in accordance with the Law "On State Statistics".

Another field requiring special measures to be taken to protect the data is data transmission through software networks, since the number of collected and distributed data through software networks is increasing. Most users, especially from governmental and political authorities, understand 'access to information' to mean access to individual data as well. The reason for this is that users do not correctly understand the principle of confidentiality in statistics and why one should observe it. As a result, they request microdata on enterprises from statistical offices.

It is important that users, especially civil servants, understand clearly what "confidentiality" means. In this context, the organization of an explanatory campaign of statistical data confidentiality through different seminars and conferences can be an important step in solving this problem.

Data storage and transmission problems are also important. Nowadays, with the growing use of the internet, the problem of protection of data from unauthorized access and provision of data confidentiality is increasingly valid. The situation is worse in regions of the country where the available premises and technical tools cannot provide proper storage of confidential data. Accordingly, this also concerns electronic data transmission from regional statistical offices to the central statistical office.

The decision of some international organizations to grant access to microdata for research purposes creates big problems for national statistical offices. In general, this good idea is difficult to realize and may have negative consequences, due to the lack of the mechanisms for its realization, because it allows unfair users to access microdata.

Even if all those who obtain access to microdata are fair-minded, they are still numerous and the concept of confidentiality would lose its importance. The situation would resemble that described by one Russian satirist: "more and more people keep our secrets", i.e. the secret is not what it used to be.

Challenges for the future

Work carried out to ensure data confidentiality cannot be immediate. With the increase in demand for information, and new technologies of data transmission and reception, the necessity to protect individual statistical data will increase and that is why the principles and methods of data confidentiality provision must be improved on a continuous basis.

One of the important issues that must be resolved is staff training on confidentiality and conduction of a single technical policy, organization and coordination of work regarding statistical data collection. Technical assistance is needed in all spheres (methodology, organization, software provision, training), but the most frequently noted problems are software provision and staff training.

Training and re-training of staff is a general problem in all fields of statistics. Salaries in statistical offices are not comparable to wages of employees working in the private sector and that is why statistical offices have to look for other methods of keeping highly qualified staff (promotion, interesting work, etc.).

As for confidentiality, especially mathematical methods and software provision for these purposes, we can say that it is a relatively new direction of work in transition countries. An additional difficulty is the lack of necessary knowledge in the country and the only way to overcome this difficulty is to recruit foreign experts to train our staff.

Attempts to hack the networks of the National Statistical Committee or to intercept data flows between statistical offices and data providers are not a big problem in the Kyrgyz Republic. This situation will probably change and that is why more attention should be paid to data protection from hackers. We need mathematical methods of data protection when we transmit the data through software networks and better software to identify attempts to hack and to protect stored data. To solve the above-mentioned problems, we are especially interested in recommendations on and appropriate instruction in practical aspects of data confidentiality protection.

II.8 STATISTICAL DATA CONFIDENTIALITY AND MICRODATA - LITHUANIA

Supporting paper by Sigitas Biciunas, Statistics Lithuania

Introduction

This paper describes the situation of statistical data confidentiality and the use of microdata in Lithuania. The legal situation, theoretical and practical confidentiality experience and problems, access to microdata and thoughts about future developments are examined.

Legal situation

Statistical data confidentiality in Lithuania is regulated by several acts. The main act is The Law on Statistics of 12 October 1993 (revival of the Law 1999-12-23) adopted by the Seimas of the Republic of Lithuania. This law describes the principle of statistical confidentiality, the definition of confidential statistical data, the purposes for the use of statistical data and obligations to protect confidential data. Confidentiality of statistical data - use of data received from statistical surveys or by other methods for statistical purposes in such a way that no concrete respondent or results of its activity could be identified. According to this law, all statistical data collected for official statistical uses has to be applied only for the preparation of statistical information. Statistics Lithuania or any other institution or agency involved in producing statistics shall take organizational and technical measures to ensure the protection of the data submitted by a respondent, and introduce computer technologies to prevent illegal usage, dissemination and destruction of the data.

Another important act is The Law on Legal Protection of Personal Data of 11 July 1996 (revival of the Law 2003-01-21) that was adopted by the Seimas of the Republic of Lithuania. This act regulates the management and the protection of data on natural persons. The usage and protection of microdata is not excluded in any Act in Lithuania, so the microdata confidentiality concept falls under the general definition of confidential statistical data.

Statistical data confidentiality

Official statistical data shall be considered confidential and protected in accordance with the procedure established by the law, if the respondent on whom or on whose activity results the primary information has been collected may be directly or indirectly identified from that official statistical data. According to this definition, microdata are confidential and have to be protected. This means that Statistics Lithuania has no right to disseminate microdata with direct or indirect identification of statistical unit and has to apply organizational and technical measures to ensure protection of these data.

The gist of the principles is that different levels of security and confidentiality have been developed: legal, organizational, methodological and technological. In these levels there is a whole complex of means and documents that ensure data security and statistical confidentiality.

A data security and confidentiality service has been established in Statistics Lithuania. This service is responsible for data security at all levels: physical, legal and technological.

There are people responsible for confidential data protection in Statistics Lithuania, who prepare the plan of data protection and control its implementation. To implement the principle of confidentiality, instructions and measures ensuring data security and confidentiality have been worked out. There are internal documents that regulate data protection in our office:

- the plan of confidential statistical data protection means;
- data protection on the networks;
- confidential statistical data protection order on physical level;
- permittance routine rules;
- the rules for internal network;
- special Confidentiality rules for Population and Housing Census;
- specific regulations on statistical confidentiality.

All questionnaires enclose a note on data security and a confidentiality guarantee. All employees who work with confidential data sign a deed of covenant to the effect that they will not disclose confidential statistical data during their lifetime.

Specific regulations on statistical confidentiality define principles of confidentiality more exactly than do our laws. Organizational confidentiality measures, methodology protecting confidential data from disclosure and conditions under access to confidential data are described therein.

Access to microdata

From 1990 to 2000, the demand for microdata was not great, but in recent years it has significantly increased. Statistics Lithuania is not allowed to hand out microdata because of the Law on Statistics and the specific regulations on statistical confidentiality laid down by the Management of Statistics Lithuania.

The Law on Statistics anticipates the possibility of releasing confidential statistics for scientific purposes. Microdata may be presented for scientific purposes if scientific institutions ensure protection of the data in such a way that it is not possible to identify respondents directly.

Access to microdata could be granted in two ways: releasing an anonymised microdata file or on the premises of Statistics Lithuania. Both methods involve a signed contract with the scientific organization concerned and all data users undertake confidentiality obligations.

If not used for a scientific purpose, data should be changed in such a way as to render it impossible to identify the statistical unit either directly or indirectly. This means that if Statistics

Lithuania receives a request for microdata other than for scientific purposes, the data will be changed using special confidentiality methods preventing disclosure.

To avoid disclosure of confidential statistical data, Statistics Lithuania uses various confidentiality methods. Different methods are used for tabular data and for microdata.
The methods used for microdata are:

- anonymisation – deleting direct identification data;
- top and bottom coding – setting top-codes and/or bottom-codes on continuous variables. A top-code for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is not published on the microdata file. Similarly, a bottom-code is a lower limit on all published values for a variable;
- sampling – releasing only a small proportion of the original data as a microdata file;
- global recoding – recoding variables into broader categories to reduce detail;
- local suppressions - deletion of especially sensitive records or items.

We also expect to start using new methods such as data perturbation or swapping. The problem is that confidentiality methods are used manually. We have tested some packages of software that help to avoid disclosure of confidential data and are preparing for their implementation.

Future developments

As for the future, Statistics Lithuania should think about remote access to microdata. In this case, there is a lot to think about and plan in order to implement an effective, reliable, secure and attractive system. The ideal for every remote access system should be the provision of an environment that allows a user to feel as much as possible as if they are working on their own PC. The major factors that should be an assurance of a good system are:

- speed: the results of analyses must be returned as quickly as possible;
- familiarity: it should not be necessary for researchers to learn new software and new programming techniques to access data;
- flexibility: the restrictions on data manipulation must be kept to a minimum;
- security: the prevention of unauthorized users;
- confidentiality: the prevention of confidential data release.

The sense of community can be encouraged through steering committees, user groups, seminar series and workshops. Good communication between all parties has a positive impact on the effective allocation of resources, research quality, levels of use and data security.

One of the main problems that we encounter is the small country problem. There are just one or two monopolistic enterprises in some economic activities. Keeping to the main principle of confidentiality is indispensable. Every time we release statistical information, we ask permission

from those enterprises to release information on their activity, but we do not obtain such authorization every time. If we do not obtain permission then we will not release any information on these economic activities, and statistics will not meet its goals. In such cases, we are thinking about a review of the exceptions on statistical confidentiality in order to change them. As this is a really sensitive solution and it could alarm respondents in regard to furnishing the correct data, the problem of the exceptions of confidentiality could be discussed during the meeting.

CHAPTER III: LEGAL ASPECTS OF MICRODATA

III.1 SUMMARY OF DISCUSSANTS' MAIN POINTS

by Katherine K. Wallman and Brian Harris-Kojetin, United States Office of Management and Budget

Introduction

Good afternoon. As we continue our seminar, we recognize as heads of National Statistical Institutes that the willingness of our respondents to provide data is inexorably linked to our ability to guarantee that we will keep their data confidential. At the same time, we are all under increasing pressure to provide the detailed microdata that are so valuable in addressing complex issues in our economy and society. Thus, a considerable effort must be made to carefully balance legitimate concerns and interests in formulating and implementing legislation in this arena. While the previous session focused on basic confidentiality legislation in Russia and the transition economies, this session examines the legal foundations for the Nordic countries and new regulations from Eurostat. Historically, confidentiality protection has to a large extent been a national issue, but in the EU-context, it has become a supranational issue for the EU institutions as well.

As we have been discussing today, the legal aspects of confidentiality and microdata access lay the groundwork for what is possible and really provide the basis for all of the issues that we are discussing in this seminar. Legislation can enable us to fully protect the confidentiality of our data, and it provides the penalty structure for violations. Legislation can also be a barrier, preventing us from being able to protect the data fully or to provide access to others who have legitimate statistical research objectives.

My plan for this discussion is to ask a number of basic questions to ascertain where we are with confidentiality legislation, and I will draw upon - and in fact, quote quite liberally from - the papers for this session on the "Legislation in the Nordic Countries" and the "Recent EU Legislation for Research Access to Confidential Data" to try to answer these questions. In so doing, I caution that any misinterpretations are mine, and I expect any errors will be corrected by the authors. Along the way, I will also interject some of our experiences in the United States. Finally, I will conclude with some implications for researchers and respondents, and with a few opportunities and challenges we face in the legal arena.

What is the scope of legislative coverage?

As the paper from Eurostat notes, from the formal legal point of view, most of the European countries 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 states 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 Nordic countries have specific Statistics and Data Protection Acts regulating the use of statistical information. According to these acts, data collected for statistical purposes, whether the information being provided is prescribed by law or is given voluntarily, may in principle only be used for the production of statistics. In Finland and Norway, the provisions for confidentiality are regulated in the Statistics Acts. In Sweden, confidentiality of data is prescribed in a special statute, the Secrecy Act.

From the supranational view, data that are received, held, used and disseminated by Eurostat are controlled by a set of laws that have developed since the Treaty founding the European Community. Basic rules and safeguards for the handling of confidential data were set out in 1990, and expanded on in 1997 in the Statistical Law (EU regulation 322/1997). These regulations represent agreements between the Commission and the Member States on the purposes for which data are provided and the conditions under which such data are provided - in essence, statements of what can and cannot be done with the data.

I gathered from reading the two papers that these laws cover all data acquired for statistical purposes, and that in the Nordic countries and at Eurostat there is broad coverage for any kind of statistical data. In the U.S., we recently passed a law, the Confidential Information Protection and Statistical Efficiency Act of 2002, which we refer to as CIPSEA. This new legislation provides one uniform set of protections to replace the earlier patchwork of legal provisions across our many government agencies that produce statistical data - currently more than 70 - and extends these protections to all individually-identifiable data collected for exclusively statistical purposes under a pledge of confidentiality. Thus, the protection is focused on the use of the data and the pledge of confidentiality, rather than being based on the legislative authority of the agency that is collecting the information. Our legislation does not restrict or diminish any confidentiality protections or penalties for unauthorized disclosure that currently exist, but does offer a uniformly high standard of protection with severe penalties (up to $250,000 fine and/or five years in prison) for Federal employees or their agents who wilfully disclose confidential statistical information.

Who may access the data?

The Eurostat Regulation 831/2002 offers a fairly straightforward and simple request process for researchers from EU member countries who are in two categories of organizations - universities and other higher education organizations established by Community law or by the law of a Member State, or organizations or institutions for scientific research established under Community law or under the law of a Member State. Other researchers who might want to access data must go through a lengthy process.

In the Nordic countries, microdata access is granted only to an officially approved institution or an individual "bona fide researcher," and signed confidentiality statements and legal contracts are

required. Researchers outside the Nordic countries and the EU who wish to access data from the National Statistical Institutes face a number of legislative restrictions. Denmark and Iceland do not provide any access to microdata to researchers in other countries. In Sweden release of microdata to an authority in another country for research is possible only if the release is compatible with Swedish interests; release of microdata to researchers in other countries is thus very restricted. Statistics Norway releases only anonymous microdata to researchers outside Norway, and the researcher must fulfil other conditions for release of data. However, there are also "safe harbour" relationships with Switzerland, Hungary and the U.S. Department of Commerce that are recognized as providing adequate protection, and may allow greater access for researchers from those National Statistical Institutes.

For what purposes can the data be accessed?

According to the main principle in the Nordic countries, confidential data may be released to a third party only for the purpose of statistical surveys and research. Under the main principle, access may be granted in forms which do not allow direct or indirect identification of individuals or of other data subjects such as enterprises. It was noted in the paper on the Nordic countries that, in practice, the Nordic National Statistical Institutes only provide access to anonymous data or microdata without name, address and identification number.

Article 15 of the Eurostat 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. The Eurostat Statistical law also makes provision for access to confidential data for scientific purposes (article 17).

Where can the data be accessed?

The legislation in the Nordic countries does not contain any rule that restricts the method used to release microdata. In Norway and Sweden, datasets on individuals or enterprises are delivered to researchers working outside the statistical office. This approach is also widely used in Finland, but mostly with personal datasets. Business datasets, however, are infrequently delivered to external users, and only after very careful data inspection and protection that involves removing large firms and adding random noise to variables.

Statistics Denmark has another practice. Since its overriding principle is not to release data outside the agency, Statistics Denmark has set up a scheme with an on-site arrangement for external researchers. Statistics Denmark also provides access to microdata from a special computer at Statistics Denmark, and the user has the possibility to manage this computer from his own office over an encrypted Internet communication. However, access is not granted for all datasets.

The new Eurostat 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 Member States if the level of the security and checking facilities is the

same as the level at Eurostat. Access of this type is often referred to as controlled access or access through a "Safe Centre".

As my colleagues from the Census Bureau point out in their contributed paper (Chapter 4), there are several modes for providing restricted access to confidential data while limiting the risk of their disclosure. The Census Bureau pioneered Research Data Centers, or RDCs. The RDCs permit restricted use of confidential files at secure sites under Census Bureau control, using limited access to dedicated computing equipment and enhanced physical and computer security. Access to an RDC facility is given only to Census Bureau employees or other persons with Special Sworn Status (SSS) who are approved to use the facility - including researchers carrying out active, approved projects at the RDC. To be granted SSS, a researcher must have an approved project, obtain a security clearance, and sign the Census Bureau's standard sworn agreement to preserve the confidentiality of the data. Persons with SSS are subject to the same legal penalties for revealing confidential information as are regular Census Bureau employees - up to a $250,000 fine and/or five years in prison.

However, for some, the concept of a physical "safe centre" has already been overtaken by events - in particular by new technology. There have been several successful "remote access" facilities for confidential data. The leading example of this cited in the Eurostat paper is the Luxembourg Income Study (LIS). It seems likely that improvements in technology will lead to greater use of this alternative. However, a non-technological issue is whether the Eurostat regulation can be interpreted as permitting this type of access.

This option is currently being offered by the U.S. National Center for Health Statistics. Researchers can remotely submit statistical programs to a centralized site, which runs the programs on confidential data, and the results are sent back to the researcher via e-mail. Certain procedures are not permitted, such as those that would allow an investigator to print out individual cases, and output is scanned before being forwarded to the researcher. These restrictions are designed to reduce to an acceptable minimum the risk of identification or of disclosure.

Another avenue that has been used by the U.S. National Center for Education Statistics (NCES) is to license researchers to allow them access to confidential data. Researchers must submit a formal proposal and include a security plan that meets the requirements of NCES. The researcher must sign an affidavit of disclosure, and the license itself must be signed by the researcher and a ranking individual who can legally bind the institution to the agreement, which in the case of a university is someone at the Dean's level or higher. On-site inspections are also conducted by NCES to insure that the proper security procedures are being followed. Finally, researchers must send copies of papers based on such data to NCES for disclosure review prior to publication.

Under what conditions can the data be accessed?

Before providing access to microdata, all of the Nordic countries follow some kind of screening procedure requiring written confirmation that the researcher has signed a general confidentiality statement. Legal contracts are made that include various limitations to access to the microdata by specifying the people, projects, variables and periods during which data can be used in the research. The obligation of confidentiality will also – according to the law or by imposition of a duty of non-disclosure – apply to the recipient of the data. The National Statistical Institute may also impose a restriction limiting the researcher's right to re-communicate or use the information. Breach of confidentiality restrictions is punishable.

Prior to the new Eurostat regulation, researchers had to contact the national statistical authority in each Member State to request permission to access the data of that state from a particular survey. Eurostat was then authorized to provide access to data of those Member States that agreed. The new regulation sets out simplified procedures under which access to confidential data for scientific purposes may be granted. For many researchers, the regulation 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 Member State and to various conditions. Although the regulation currently refers to four important datasets [European Community Household Panel (ECHP), Labour Force Survey (LFS), Community Innovation Survey (CIS), and the Continuing Vocational Training Survey (CVTS)], more datasets may be included if this approach proves to be successful.

Summary

To summarize, current legislation and regulations in the Nordic countries and for Eurostat provide broad confidentiality protection for data gathered for statistical purposes. Researchers, primarily from universities or other scientific organizations, may access microdata strictly for statistical or other scientific purposes. They are often restricted in accessing the data physically only at Eurostat or a National Statistical Institute, though the technology (and convenience) of remote access seems likely to spur greater use of this form of access. What researchers must do to access confidential microdata varies according to the country that has the data, and the researcher's country of origin. The Eurostat regulations provide a standardized means of accessing a number of important datasets for researchers from EU Member States.

What are the implications of this legislative slate for researchers and for respondents?

Implications for researchers (data users)

On the one hand, there are new opportunities for access to microdata, and on the other hand, there are tight disciplines and limitations that are imposed as the price of these opportunities. The research community must accept that there is no "right of access" and that researchers will have to share a responsibility to maintain and uphold the confidentiality of data they access. Although the

limitations and safeguards may be more restrictive than those prevailing in the researchers' universities and those they have encountered in accessing other datasets, the limitations and safeguards still must be honoured.

Some U.S. agencies have approached this challenge by using license agreements that legally bind the individual's institution to maintain the confidentiality of the data. Still, the existing mechanisms apply to those in the institutions who use the data, rather than to the data. One of my colleagues recently suggested that it would be helpful if statistical datasets could carry the same warning about disclosing identifiable information as videotapes in the U.S. carry for copyright infringement. We need to make the penalties for breaches of confidentiality as obvious to data users as these warnings are to videotape users.

Implications for respondents (data subjects)

There are also implications for our respondents. The principle underlying statistical data collection is that of informed consent. Our respondents have a right to know what their information will be used for and who will see their information. But the question arises, "how much do we need to tell respondents about who will have access to the data for research and statistical purposes?" For example, do we need to specify that researchers within the country or even perhaps outside the country, who may be accessing the data from Eurostat, will be able to use the data for statistical and scientific purposes? How much information should we share to adequately inform respondents and what are the implications for their cooperation with our requests for information?

In the Eurostat paper, mention was also made of necessary field research to understand respondents' perceptions. Work by Eleanor Singer in the U.S. along these lines has shown that the more we emphasize the confidentiality protections for the data, the more respondents become concerned about these issues and the more reluctant they are to provide information. What is the optimum balance between providing adequate information about the possible uses of the data and creating unwarranted perceptions with respect to problems of disclosure? How can we reassure respondents without making them more concerned than they were before we attempted to allay their concerns?

Other opportunities and challenges

We face a number of other opportunities and challenges in the legislative arena. Let me just describe a few of these.

Legislation in the face of the war on terrorism

We are now facing a far more complex environment that may well make researcher access a relatively modest matter. In the immediate aftermath of September 11th in the U.S., there was one specific case of the potential reversal of a prior pledge of confidentiality by allowing the Attorney General access to statistical data collected under a pledge of confidentiality in cases of national

terrorism. Fortunately, a more measured stance has prevailed since that time, and in a broader bill on homeland security information, specific exclusions were made for data gathered for exclusively statistical purposes under a pledge of confidentiality. We wonder if there have been any similar efforts to infringe upon confidential data collected for statistical purposes in other countries, particularly in response to more recent threats to national security.

Confidentiality of publicly available information

Another area where we face an interesting challenge and opportunity is the use of publicly available data by National Statistical Institutes. An article in the Eurostat Statistical Law states that "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, namely that information obtained by a National Statistical Institute 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.

This issue has a close parallel in the U.S., where the Bureau of Labour Statistics (BLS) currently treats the North American Industry Classification System (NAICS) codes as confidential because the agency treats all information it collects from business establishments as confidential. However, because of the risk of disclosure, this practice results in BLS publishing tables in reports that are missing two-thirds of the entries when industry codes are crossed with some levels of geography. These NAICS codes reflect the line of business the establishment is in. Such business lines (though not necessarily the codes) are certainly publicly proclaimed by the establishments and are actually published at the county level by the Census Bureau, which does not consider the information confidential.

So, it appears that our challenge is to discern how we can make clearer to establishments that there is some information National Statistical Institutes are collecting that is public information (and will be treated as such), and some information that will be kept confidential. It would certainly be interesting for countries to share their experiences in handling these kinds of situations.

I would like to conclude my remarks by thanking Statistics Sweden and Eurostat for their interesting and informative papers, and give them the opportunity to respond to my comments before we open this up for wider discussion.

III.2 RECENT EU LEGISLATION FOR RESEARCH ACCESS TO CONFIDENTIAL DATA – IMPLEMENTATION AND IMPLICATIONS

Invited paper by John King, Eurostat

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.

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 organizations. Sometimes these analyses are seen as being outside the remit of national statistical organizations (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 organizations 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 - microdata 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 microdata records for computing correlation matrices, estimating models and other analyses, depending on the context of the research topic.

Examples of the microdata 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 microdata" last September 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 on the supply side, statistical organizations, 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 organizations, particularly NSIs, can do and on how they can do it. The role of researchers and research organizations 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.

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 organizations within the MSs (other researchers or organizations 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 organization.

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."

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 organization. This means that there must also be a contractual relationship between the researcher and his or her organization. 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, the dialogue has started with recently contacts with CEIES, ESF (European Science Foundation), and other international research bodies.

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 organizations:

1(a), i.e. universities and other higher education organizations established by Community law or by the law of a Member State; or

1(b), i.e. organizations 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 organizations 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 organization;
- the organizational arrangements for research in the organization;

- the safeguards in place in the organization;
- 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 organization 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 organization;
- brief description of the research project(s);
- organizational and financial arrangements for research within the organization;
- security in place in the organization;
- 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".

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 organizations 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. The problems of disclosiveness in tabular data are understand well - and methodology for this exists and is also still developing - but there is 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. Greater understanding is also needed about the disclosive potential of parameter estimates - particularly when a series of similar models are run and compared.

Implications of Regulation 831/2002 for the research community

The implications of the new Regulation for the research community illustrate the nature of the partnership between statistical organizations 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.

Implications of Regulation 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 organizations 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 organization 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 organizations. 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 organization. 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 organization and will be seen by the latter as a factor affecting response rates to voluntary enquiries.

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?

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.

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.

III.3 LEGAL ASPECTS – LEGISLATION IN THE NORDIC COUNTRIES

Invited paper by Birgitta Pettersson, Statistics Sweden

Introduction

This paper presents some of the legal issues concerning access to microdata and describes the legislation in the Nordic countries. The paper is based on the Nordic Statistical Agencies report "Access to microdata in the Nordic countries".

National Statistical Institutes (NSIs) are dependent on the confidence of the respondents, and they are required to respect confidentiality and protect individuals' integrity. The willingness of respondents to provide data is dependent on the ability of statistics offices to guarantee respondents' anonymity.

Confidentiality protection of individual and business data is one of the main principles of official statistics. The individual is entitled to be protected by society against unacceptable intrusion of personal integrity. At the same time the need of the individual for protection must be balanced against legitimate needs of using information connected to people, for example, for purposes of statistics and research. Microdata collected for statistical purposes is of vital interest for researchers. In recent years the demands to make statistical microdata available for research purposes have increased. Researchers also ask for increasingly detailed data. A considerable amount of balancing is necessary when formulating legislation to protect personal integrity as regards personal data. Furthermore, legislation must not unnecessarily restrict the use of new technology, which brings with it not only risks but also advantages.

The protection measures applied to confidential data obtained for statistical purposes are based on several legal acts and directions. It is important that the regulation concerning statistics and confidentiality is clear and well suited to its purpose. Clear rules are needed to create confidence, especially with the respondents. The legislation concerning confidentiality and protection of individuals' integrity is of importance for the possibility of the NSI to provide access to microdata. The legislation provides the limits for release of data for e.g. research purposes, and constitutes and strengthens administrative and technical safeguards for legal founding. Specific legislation of importance in the Nordic countries is the Statistics Act and the Data Protection Acts. To this specific legislation, the current EU legislation with respect to statistical confidentiality should also be added.

General rules

The use of statistical information is often regulated by legislation or in a code of practice. In the Nordic countries there are specific Statistics Acts regulating the use of statistical information. According to these acts, data collected for statistical purposes, in accordance with any prescribed obligation to provide information or which is given voluntarily, may in principle only be used for the production of statistics. There are exceptions that allow access to data for research purposes and

public planning. However, a condition for their use for research is that there is no incompatibility between the purpose of such processing and the purpose for which the data was collected. The processing of data, which includes release of data, must also be in accordance with the regulation concerning protection of individual integrity.

The general personal data protection acts in the Nordic countries (the Personal Data Acts[1]) also apply to the production of statistics and the release of microdata. The Acts are based on the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. The Acts contain rules about the fundamental requirements concerning the processing of personal data. The Personal Data Acts are similar in the Nordic countries and apply to all forms of processing of personal data, including registration, storing, disclosure, merging, changes, deletion, etc.

According to the Personal Data Acts, data must be:

- processed fairly and lawfully;
- collected for specified explicit and legitimate purposes and not further processed in a way incompatible with those purposes. However, further processing of data for historical, statistical or scientific purposes is not considered as incompatible;
- adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;
- accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, with regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;
- kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Personal data can be stored for longer periods for historical, statistical or scientific use.

If data about a person is collected from the person him/herself, the controller of personal data shall in conjunction therewith voluntarily provide the person in question with information about the processing of the data.

Very stringent rules apply to the processing of sensitive personal data. Sensitive data is personal data that discloses race or ethnic origin, political opinions, religious or philosophical convictions, membership of trade unions and personal data relating to health or sexual life. The main principle is that such data may be processed only with the consent of the person in question. However, sensitive data may be processed for research and statistical purposes without consent, provided the processing is necessary and provided the public interest in the project clearly exceeds the risk of improper violation of personal integrity.

[1] In Denmark – the Act on Processing of Personal Data, Act No. 429 of 31 May 2000, Finland – the Personal Data Act 523/1999, Island – Act on the Protection of Individuals with regard to the Processing of Personal Data 77/2000, Norway – the Personal Data Act (2000), Sweden – the Personal Data Act (1998:204).

Furthermore, in Denmark, Norway and Sweden a scientific project involving processing of sensitive personal data without consent is subject to notification to and approval by the Data Protection Agency before such processing can commence. This applies to all surveys, whether conducted by a public administration, individuals or enterprises. (In Sweden the approval of the National Data Protection Agency is not necessary if a research committee has approved the processing.) If the Data Protection Agency approves the processing, personal data may be released and used in research projects unless otherwise provided by the rules on confidentiality. This means that NSIs may take other issues into consideration even if the Data Protection Agency (or in Sweden a research committee) has approved the processing of data. The Data Protection Agency only considers whether the processing is in accordance with the Personal Data Acts. The NSIs must also consider whether data can be released without disclosing individual information.

Confidentiality

Data, even anonymous data obtained for statistical purposes, are confidential. The statistical data are confidential irrespective of source. Also data taken from public administrative sources are confidential while in the possession of the NSI. According to legislation in the Nordic countries, it is prohibited to disclose confidential data to unauthorized people. In Finland and Norway the provisions for confidentiality are regulated in the Statistics Acts. In Sweden, confidentiality of data shall be stated in a special statute, the Secrecy Act. The Secrecy Act contains provisions on what is to be kept confidential in state and municipal activities. Confidentiality in the Secrecy Act is usually expressed to apply in relation to certain matters, for certain operations and regarding certain public authorities. Confidentiality does not apply to the information released to another authority unless this is provided in the Secrecy Act.

According to the main principle in the Nordic countries, confidential data may be released to a third party only for the purpose of statistical surveys and research. Under the main principle, access may be granted in forms which do not allow direct or indirect identification of individuals or of other data subjects like enterprises. In practice the Nordic NSIs only provide access to anonymous data or microdata without name, address and identification number.

There is often no legal definition in national legislation that aims at whether an individual or enterprise is identifiable. However, the definition in the Council Regulation (EC) No 322/97 of 17 February 1997 on Community Statistics, Article 13 states: 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.

The avoidance of indirect identification is difficult, especially when the data set includes a lot of detailed information that, for example, may have skewed distribution in statistics. This is the case in the so-called ‘linked employer-employee data’. Disclosure of data is also a problem in small area statistics or statistics concerning business data where one entity is the sole producer.

The obligation of confidentiality will also – according to the law or by imposition of a duty of non-disclosure – apply to the recipient of the data. The NSI may also impose a restriction limiting the researchers right to re-communicate or use the information. Breach of confidentiality restrictions is punishable. In Sweden, however, it is not possible to impose restrictions when data are released to other authorities. It is therefore important for Statistics Sweden to also take into consideration whether data will be confidential according to the Swedish Secrecy Act by the authority receiving data. If not, anyone who so desires can have access to the data because of the authorities' obligation under Chapter 2 of the Freedom of the Press Act to provide access to data that are not confidential.

All exceptions from this principle of public access to official information must be stated explicitly in the Swedish Secrecy Act. The Secrecy Act does not contain any general rule concerning the transfer of confidentiality between public authorities. As a rule, that authority with its own confidentiality rule applicable to the information satisfies the need for confidentiality by the recipient authority. However, there are rules ensuring that secrecy accompanies information to another authority in special situations. One of these rules states that if an authority receives data for research purposes from another authority where the data is confidential, the confidentiality will also apply within the receiving authority. In practice, this means that in most cases there is not any problem with providing access to microdata to researchers working at authorities, e.g. universities. However, there are no such rules concerning release of data for statistical purposes or public planning.

In addition to laws and regulations on data confidentiality, all the Nordic countries follow some kind of screening procedure requiring written confirmation that the researcher has signed a general confidentiality statement. Legal contracts are made that include various limitations to the access to microdata by specifying the people, projects, variables and periods during which data can be used in the research. However, as previously mentioned, Sweden does not impose restrictions when data are released to another authority.

In practice, the Nordic NSIs will provide access to statistical microdata only for a specific research purpose. In principle, access to microdata is provided only to an authority, officially approved institution or individual "bona fide researcher".

Making microdata available

The legislation in the Nordic countries does not contain any rule that restricts the method of releasing microdata. As long as the general demands in the legislation are fulfilled, the NSI can choose the method used. How access to confidential data is provided in practice can be divided into several categories: Off-site access, on-site access, off-line access and on-line access.

In Norway and Sweden, data sets on individuals or enterprises are delivered to researchers that are working outside the statistical office. This approach is also widely used in Finland, but mostly with personal data sets. Business data sets, however, are infrequently delivered to external

users, and only after very careful data inspection and protection that involves removing large firms and adding random noise to variables.

Statistics Denmark has another practice. Since its overriding principle is not to release data outside Statistics Denmark, they have set up a scheme with an on-site arrangement for external researchers at Statistics Denmark. Under this scheme, researchers can obtain access to register data that do not have identifying variables or are anonymous from a workstation at the premises of Statistics Denmark. Statistics Denmark also provides the user with access to microdata from a special computer at Statistics Denmark, and the user has the possibility to manage this computer from his own office over an encrypted Internet communication. However, access is not granted for all datasets; particularly sensitive data are excluded from the scheme and data on enterprises are assessed carefully to avoid any problems of confidentiality.

Release of microdata to researchers in other countries

According to Directive 95/46/EC, Member States are required to ensure that the transfer of personal data to a third country (a country outside the EU and EEA) may take place only if the third country in question ensures an adequate level of protection and the Member State laws implementing other provisions of the Directive are respected prior to the transfer. However, the Commission may find that a third country ensures an adequate level of protection. In that case personal data may be transferred from the Member States without additional guarantees being necessary. According to the Directive, the level of data protection should be assessed in the light of all circumstances surrounding a data transfer operation or a set of data transfer operations and in respect of given conditions. The Working Party on Protection of Individuals with regard to the Processing of Personal Data established under that Directive has issued guidance on the making of such assessments.

Switzerland[2], Hungary[3] and the US 'safe harbour' arrangement have been recognized as providing adequate protection. In order for US organizations to comply with the Directive, the US Department of Commerce in consultation with the European Commission developed a safe harbour framework. The EU approved the safe harbour principle in July of 2000[4]. Certifying to the safe harbour will assure that EU organizations know that an enterprise provides ‘adequate’ privacy protection, as defined by the Directive. Safe harbour does not cover all organizations in the US.

[2] Commission Decision of 26 July 2000 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequate protection of personal data provided in Switzerland.
[3] Commission Decision of 26 July 2000 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequate protection of personal data provided in Hungary.
[4] Commission Decision of 26 July 2000 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the safe harbour privacy principles and related frequently asked questions issued by the US Department of Commerce.

The European Commission[5] has also recognized that the Canadian Personal Information Protection and Electronic Documents Act provides adequate protection for certain personal data transferred from the EU to Canada. However, the Canadian Act and the Commission Decision do not cover personal data held by public bodies, both at federal and provincial level, or personal data held by private organizations and used for non-commercial purposes, such as data handled by charities or collected in the context of an employment relationship.

For transfers of data to recipients in organizations not covered by the above-mentioned decisions and other countries, such recipients in the EU will have to enact additional safeguards, such as the standard contractual clauses adopted by the Commission in June 2001[6], before exporting the data.

In the Nordic countries, the same regulation concerning data confidentiality as for release of data outside the Statistics agencies is in principle also valid when data is delivered to other countries. However, there are some restrictions. The Personal Data Acts in the Nordic countries contain similar rules as the Directive that restricts release of data to a third country. According to the Personal Data Acts, it is in principle forbidden to transfer personal data that is being processed to a third country unless the third country in question ensures an adequate level of protection.

In Sweden the Secrecy Act is also of relevance. According to the Act, release of confidential data to an authority or an international organization outside Sweden is not allowed unless i) it is released in accordance with special provisions in legislation, or ii) the data in a corresponding case might be given to a Swedish authority and the authority holding the data deems it evidently compatible with Swedish interest that the information be released.

The EU regulation includes special provisions that make it possible to release microdata to Eurostat. There are no other special provisions concerning statistical microdata. In Sweden release of microdata to an authority in other countries for research is therefore possible only if it is compatible with Swedish interest that the data be released. Microdata may be released to private researchers in other countries if it is clear that the information can be released without the person whom the information concerns suffering loss or being otherwise harmed. In practice, Statistics Sweden is restrictive with release of microdata to researchers in other countries.

Statistics Norway only release anonymous microdata to researchers outside Norway and the researcher must fulfil the other conditions for release of data.

In Finland the same regulations concerning data confidentiality are valid as for the release of data outside Statistics Finland. The only exception is that the delivery of data outside Finland

[5] Commission Decision of 20 December 2001 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequate protection of personal data provided by the Canadian Personal Information Protection and Electronic Documents Act.

[6] Commission Decision (2001/497/EC) providing standard contractual clauses for the transfer of personal data to third countries.

requires the approval of the Director General. However, before the Director General submits an application for the final decision, the application needs to be addressed by the Data Protection Board. An applicant must provide the Board with a description of how the confidentiality of the data will be ensured outside Finland.

Denmark and Iceland do not provide access to microdata to researchers in other countries.

The release of information to Eurostat is regulated in the EU regulations on statistics. Member states are in principle bound by these regulations to release microdata for community statistics.

References

Access to Microdata in the Nordic countries (2003). Statistics Sweden.

CHAPTER IV: ACCESS TO MICRODATA ISSUES, ORGANIZATION AND APPROACHES

IV.1 ACCESS TO MICRODATA – THE SITUATION IN THE AUSTRALIAN BUREAU OF STATISTICS

Invited paper by Dennis Trewin, Australian Bureau of Statistics

Background

Ensuring confidentiality is not only important for legal and ethical reasons, but to maintain public trust. The increasing demand for detailed data, combined with the increasing power and capability of technology, and the availability of potentially matchable data sets, makes the challenge of maintaining the confidentiality of microdata more difficult. It is virtually impossible to release microdata which contains useful information that is unlikely to be unidentifiable. Longitudinal data sets increase the problem. We can no longer rely solely on different forms of data amendments to manage disclosure risks.

On the other hand, there is increasing demand for access to microdata to support a range of research and secondary data analysis. Increased computing power increases the capability of researchers to undertake these types of analysis.

There are several motivations for addressing the issue on how to best provide researcher access to microdata:

- valuable (and high quality) data is underutilised;
- researchers may try to collect substitute data sets in order to obtain microdata, which is a waste of public resources (to obtain what is probably lower quality data);
- government agencies may look to use alternative data providers to obtain survey data for research and analysis purposes, resulting in lower quality data that may not contribute to national statistics.

There is another important element that we need to consider - the incredible potentially valuable analytical power of linked data sets; including links with ABS data sets.

This range of factors has led us to rethink how we provide access to microdata. This is true for many other NSOs, many of whom are in the process of changing their practices. The steps we have taken, or plan to take, are described in this paper. Different strategies may be required for household and business based surveys. This paper only attempts to describe the ABS situation but hopefully this will be relevant to the situation many other NSOs face.

Before moving on, it is worth emphasising that whatever is done must be both legal and publicly acceptable. The law could be changed but this is not a quick or straightforward process and may raise unnecessary concerns. Consequently our approach is to work within existing law.

A brief description of demand

Ideally users would like:

- the ability to work interactively with the data;
- access to ABS experts and good documentation to describe the data;
- an increasing number of data sets available;
- good quality data for populations and variables of interest and some information about the sources of error;
- timely releases; and
- increasingly, access to linked data set (including data linked over time).

There has been a consistent message from researchers that we have taken too conservative an approach to the release of microdata. As a result, a recent focus has been to consider how we can increase access to microdata, while maintaining our high reputation for safeguarding privacy (and staying within the law) which is so important for maintaining a high level of cooperation in our surveys.

The use of linked data sets raises the possibility of ABS acting as a custodian of non-ABS data sets to ensure that there is appropriate confidentiality protection. While this is entirely consistent with National Statistical Service objectives, appropriate policies and operational procedures need to be developed. This is discussed further below.

Means of satisfying the demand

There are a range of options or dissemination streams, which vary in terms of their "safety" from confidentiality breaches. The first listed options tend to rely more on safe data whereas the last listed rely more on a safe environment, including reliance on legally binding undertakings with strong sanctions for breaches.

The accessibility and convenience to researchers will also vary by option.

Release of microdata, which is the specific subject of this session, is a key element of providing access for research purposes. Statistics legislation allows us to release microdata but only "in a manner that is not likely to enable the identification of the particular person or organization to which it relates". Undertakings are also required. Nevertheless, there are several ways of accessing microdata whilst complying with this legal constraint. These are explained below and summarised in Table 1.

A Microdata Review Panel has been established to help us assess whether the disclosure risk is acceptably low (i.e. "not likely to enable the identification of ") for those dissemination streams that involve microdata. They look at two key risk areas:

- prevention of spontaneous identification; and

- prevention of matching risk.

Legal advice is that a legal undertaking preventing certain actions is consistent with "in a manner not likely to" and should be taken into consideration when making these judgements.

The advantages and disadvantages of each stream will be further developed in the following sections.

Table 1: Dissemination Streams to Support Research

Dissemination Stream	Notes
1. Standard Statistical Outputs	Usually in the form of tables. Restricts the type of analysis that users can undertake.
2. Datacubes	Provide more detail and the flexibility of researchers to generate their own tables.
3. Special Data Services	At the request of researchers, usually at marginal cost.
4. Confidentialised Unit Record Files (CURFs)	Data is unidentifiable. Release is on CD ROM. Equivalent to what are generally termed microdata releases.
5. Remote Access Data Laboratory (RADL)	Access to CURFs but more detailed release may be possible because of the greater control over prevention of matching with external databases.
6. ABS Site Data Laboratory	Still only provides access to unidentifiable data.
7. Collaboration	Means working through an ABS officer rather than accessing microdata directly.
8. In-house Analysis	In effect, working as an ABS officer working on ABS premises. This is only possible if the researcher is assisting the ABS with its functions.

Standard Statistical Outputs

- What does it involve? - The release of statistical outputs, usually in the form of tables, in printed and/or electronic form.

- Confidentiality Protection - This is a safe data. Standard ABS Confidentiality Practices are applied.

- Advantages - Convenient and cheap. Provides a good indication of full range of data. Increasing availability of electronic data in downloadable form improves convenience of use for further analysis. Easily accessible to a range of researchers. Low cost to the ABS.

- Disadvantages - Limits the types of analysis that can be undertaken. Not possible to undertake analysis that relies on microdata.

- Current State of Play - Increasing the availability of data in this form on the web site. Improving the availability of supporting metadata.

- When to use? - Should not be underestimated as a convenient means of supporting research. Should be a key consideration of the dissemination strategy for all statistical outputs.

Datacubes

- What does it involve? - The release of detailed statistical matrices that have already been confidentialised. It is a more appropriate form of release when confidentiality protection can be automated, particularly for small cells (e.g. population census). Special confidentiality provisions for trade data also allow data to be released in this form.

- Confidentiality Protection - This is safe data. Standard ABS Confidentiality practices apply (unless there are special provisions which exist for some data e.g. trade).

- Advantages - Reasonably convenient access to more detailed data than standard statistical outputs.

- Disadvantages - Same as for Standard Statistical Outputs. Also, design of good datacubes is not straightforward. Some researchers also find it difficult to use datacubes. Will not be possible to produce confidentialised datacubes for many statistical outputs.

- Current State of Play - We are slowly increasing the availability of datacubes. Increasing the know-how of the designers of datacubes.

- When to use? - Will generally be more useful for personal data than business data. For some statistical outputs, should be considered as part of the dissemination strategy.

Special Data Services

- What does it involve? - The release of statistical outputs, not necessarily tables, at the request of researchers.
- Confidentiality Protection - This is safe data. It will not be provided to the researcher unless confidentiality is already protected.

- Advantages - The data and form of delivery can be tailored to the researchers need.

- Disadvantages - Will be expensive to some researchers (and for the ABS to service). Analysis limited by inability to work interactively. Researcher cannot apply own adjustments (e.g. for outliers) to the microdata. Turnaround to different runs of the data analysis might be slow.

- Current State of Play - Offered as a service but demand is not great. Not trying to develop, except for key clients and selected areas (e.g. regional statistics).

- When to use? - Usually for tabular outputs when not provided through standard outputs and access to microdata is not possible. Other forms of analysis are more likely to be run as a collaborative arrangement (see below).

Confidentialised Unit Record Files (CURFs)

- What does it involve? - The release of microdata files on a CD ROM which have been amended so that the identification of an individual person or organization is unlikely.

- Confidentiality Protection - A Microdata Review Panel advises on the adjustments that are required to protect the confidentiality of the data. This may involve data amendment techniques such as deletion of some variables, reducing the detail available in some variables (particularly geography), deleting some highly identifiable individuals, and random perturbation. The confidentiality is further protected by requiring a legal undertaking from all researchers accessing the microdata. In cases of breaches sanctions will be applied (including the withdrawal of the microdata service) to the researcher and possibly their institution. Legal recourse may also be sought.

- Advantages - Great flexibility and convenience to the researcher.

- Disadvantages - Not all the detail being sought is available. Generally CURFs are not available for data about businesses. There have been a small number of breaches of the undertaking (but not identification of individual records).

- Current State of Play - Will remain a significant dissemination stream for supporting research and secondary data analysis. Demand is high. We are trying to improve the timeliness of our releases.

- When to use? - Is regarded as one of standard outputs from household surveys. Used selectively for other surveys where data is still useful for research purposes after confidentiality protection has been applied.

Remote Access Data Laboratory (RADL)

- What does it involve? - Running jobs submitted by authorised users via the internet against CURFs held at the ABS, and returning analysis results after largely automated confidentiality checks. Similar to CURFs except it should be possible to provide access to more detailed data because matching risk can be controlled as data does not leave the ABS. Limited to range of analysis software provided through RADL (e.g. SAS, SPSS). Outputs will be manually inspected before onward release.

- Confidentiality Protection - Advice of Microdata Review Panel. Manual inspection of outputs enhanced by automatic triggers to identify output that may require rigorous inspection. Audit trails and records kept. Legal undertakings will need to be made. Sanctions against offenders.

- Advantages - Access to more detailed data. Access to analysis software that might not be available to the researcher. Free processing facilities.

- Disadvantages - Inconvenience compared with CURFs. Some delays in the release of outputs. More expensive for the ABS to administer.

- Current State of Play - Was launched in April 2003. Will be modified in light of user reaction. The number of data sets available through this facility will increase, over time.

- When to use? - Will use rather than CURF service when data matching risk of CURF is too great, and reliance on undertaking/sanctions is risky. For example, it may be used for linked data files, particularly if one of the linked files is available externally.

ABS Site Data Laboratory

- What does it involve? - Similar to RADL except that no downloading of unit record data is available (this is possible in RADL for up to 30 records to support outlier detection, etc.). Note that it is different to the situation in many other countries where a declaration of secrecy enables on-site access to unconfidentialised unit record files. We cannot do this unless the researcher is genuinely assisting the statistician to perform his functions and his employment status means that the researcher can be deemed an ABS officer. This would mean payment for services.

- Confidentiality Protection - Similar to RADL except that there is more control on output; no unit record data can leave the ABS.

- Advantages - Access to data that may not be possible through CURFs or RADL (e.g. longitudinal data files). More direct access to ABS experts.

- Disadvantages - Inconvenience of working on ABS premises. Expensive for ABS staff to manage, particularly across nine offices.

- Current State of Play - Is available now. Main use has been for longitudinal data files, particularly where the sample unit or some of the data has been derived from the administrative system of another agency.

- When to use? - Only when CURF or RADL service is deemed inappropriate for a data set or the researcher prefers this form of working and ABS is prepared to support.

Collaboration

- What does it involve? - Working collaboratively with a researcher to produce an output (often a published output) of relevance to the ABS. May or may not be a statistical output released by the ABS. The arrangements generally do not prevent researchers publishing or presenting the results of this work elsewhere, including in scientific journals.

- Confidentiality Protection - The research collaborator does not directly access unit record data. This is done by the ABS staff member working with them.

- Advantages - Mutual benefits from collaborative effort. Genuine knowledge transfer. Researcher could mostly work away from the ABS office. May result in funding being made available to the researcher to assist with research. Costs to researcher will generally be lower. Potential access to Australian Research Council grants.

- Disadvantages - No direct access to data. Limited to collaborative projects of interest to ABS.

- Current State of Play - Policy on collaborative arrangements has been put in place. Analysis Branch has been established and has been in operation for four years with about 30 staff members. This has provided a real focus for collaborative effort with the research community. Previously, arrangements were ad hoc.

- When to use? - In cases where collaboration will result in outputs of mutual benefit. For some higher priority projects, the ABS may seek collaborators. As well as confidentiality, principles that should govern collaborative work are consistency with government purchasing principles, deriving statistical value, even-handedness and transparency, and protecting the ABS reputation.

In-House Analysis

- What does it involve? - The ABS can engage persons as "officers" if they are undertaking functions to support the ABS in its activities. In these situations they can access unit record data although subject to the same secrecy provisions of other ABS officers. This may be appropriate when the ABS wishes to produce an output where the researcher can cover an identified gap in expertise. Generally, arrangements can be made to allow researchers to publish aspects of their work elsewhere with permission.

- Confidentiality Protection - Secrecy provisions apply as they are ABS officers. Liable to severe penalties for breaches.

- Advantages - Provides researcher access to unit record data. Mutual benefit from collaborative work.

- Disadvantages - Much of the work will need to be done on ABS premises. Limited to subjects of direct relevance to the ABS. Some restrictions on research outputs. May not always be possible to employ as an ABS officer.

- Current State of Play - This provision has been rarely used. Recent changes to public service arrangements make it easier to implement.

- Where to use? - When the ABS takes the initiative to engage a researcher to assist it with its statistical activities. (There still may be mutual benefits of course.)

Linked data sets

Linked data sets are a special case of a microdata set that users may want to access. Here I am talking about using data matching techniques to bring together unit records to form a set of composite records. The composite record may be based on a hard match using identifiers or a statistical match using a combination of variables (e.g. geography, age, sex, household characteristics). Both are of concern from the point of view of confidentiality. Hard matches are clearly of greater concern but research we have undertaken indicates a surprising high proportion of exact matches when undertaking statistical matches, particularly for files that include the household structure.

Linked data sets may comprise:

(a) matching ABS data sets;
(b) matching an ABS and non-ABS data set; or
(c) matching non-ABS data sets.

In (a) and (b), the ABS must be the custodian and access has to be through the dissemination streams discussed in this paper. It is not necessary for the ABS to be custodian for (c) but there are advantages. We have legislation which could underpin the arrangements for accessing these data sets and protect their confidentiality. Furthermore, our reputation is such that there is strong public confidence that we will be a trusted custodian. We also have the tools and systems to support access.

A linked data set can have considerable analytical power as illustrated by the following examples:

- studying the interactions of a person with different institutions - by say linking together the records of health services provided by medical practitioners, hospitals, nursing homes and the like;
- studying the relationships between inputs, outputs and outcomes by drawing together information on policing, courts and prisons; and
- studying through time patterns by assembling a longitudinal database.

Some additional principles are needed for creating/working with these data sets. The core principles will be as follows:

- consistency with the ABS mission to use statistical information to better support informed decision making, research and discussion;
- a demonstrable statistical benefit;
- integrity and openness about applications;
- publication of a statistical output from each linked data set;
- maintaining public trust by ensuring ABS legislation, privacy legislation and other relevant legislation is followed.

We are considering the establishment of an Ethics Committee to help us with decisions in this area.

What are our plans?

Until recently, the situation for each of the dissemination streams, to support external researchers, was as set out below.

- Standard Statistical Outputs - a standard service was available for all fields of statistics.

- Datacubes - under development but available for some statistical series (e.g. demography, labour force).

- Special Data Services - available but only used occasionally. Usually for production of detailed tables.

- CURFs - A regular output from household surveys, occasionally for business surveys, but needing to curtail detail released because of increasing matching risks. Also needing to strengthen the legal undertakings that are necessary for release.

- RADL - Service not available.

- ABS Site Data Laboratory - Used occasionally but not promoted.

- Collaboration - Used only occasionally in the past but over the last year or so, about forty collaborative arrangements have been established. This includes 12 collaborative arrangements as a result of the Australian Census Analytical Program.

- In-house Analysis - Used rarely.

In the future, all eight streams will be used to support external researchers. Because of their expense, we will try to limit special data services (stream 3) to key clients. Because of the increased matching risk, there will be some contraction of the detail available on CURFs. Nevertheless they will remain a key means of researcher access to microdata.

The key areas of development will be Standard Statistical Outputs (Stream 1), RADL (Stream 5) and Collaboration (Stream 7). We expect that more Datacubes (Stream 2) will be released but, realistically, it is only a suitable form of output for a limited range of statistical series.

Our objective under "Standard Statistical Outputs" will be to increase the amount of data that will be available in this form through our special web based services (e.g. AUSSTATS). All statistical areas will be asked to review their dissemination strategies with the view to reducing reliance on paper publications and increasing output available electronically.

RADL is a new service which will have just commenced operation by the time of the CES meeting. We see this as an area of further development in light of experience with the first version. It will be especially targeted at:

- providing microdata access to more detailed data sets; and
- providing access to linked data sets, especially where one of the data sets are held externally.

We are pursuing "collaboration" more actively now that we have a fully effective and highly respected Analysis Branch. We will attempt to initiate collaboration in these areas of greater interest to us, particularly when a new statistical output might result. Of special interest is adding value to existing data sets through analytical techniques. In practice, some researchers will approach us in the first instance. We will assess whether there are likely to be mutual benefits from collaborative arrangements. Dissemination Stream 8 may be appropriate for some collaborative projects but it is an approach we would use selectively.

Organizational arrangements

The leadership for these arrangements must come from the ABS Executive especially whilst they are going through a period of substantial change. Communication is important, both internally and externally. We are supported by the ABS Branch responsible for policy and coordination.

The actual management and administration of the arrangements lies with our Information Services Division. Within this Division, they have a unit responsible for the administration and

distribution of CURFs, RADL and the ABS Site Data Laboratories. They are also responsible for promoting these services and managing the relationship with clients.

Access methods are still under development in many respects. To strengthen our research capabilities, including research done elsewhere, and to provide more focus, we have created a Data Access and Confidentiality Methods Unit within Methodology Division. This is headed by a senior methodologist.

A special project team (over sighted by a Project Board) was established to support the development of RADL. The ongoing responsibility for maintenance of these systems has been transferred to Information Services Division.

Analysis Branch is responsible for the setting up and managing most of the collaborative arrangement that rely on access to microdata. Some may be managed through the statistical areas but this will be an exception. We are using Analysis Branch to ensure greater consistency of approach. Furthermore, they have the technical know-how to work most effectively with research collaborators. A Project Board of our most senior subject matter statisticians oversights this work.

Finally, the statistical areas need to be closely involved. They are responsible for providing the underlying data sets for all the dissemination streams. Furthermore, researchers will need to call on their subject matter expertise from time to time.

Key issues

It is becoming more and more difficult to provide truly "safe data" so it is inevitable that we will need to rely more on "safe settings", including legal arrangements, to support secondary data analysis. This is more labour intensive - requiring additional resource commitments when NSOs are often under resource pressure. Still, we believe it is an appropriate reallocation of resources if our data is being used effectively.

Researcher acceptance of these arrangements may be an issue. From their point of view, they may provide unnecessary constraints or inconveniences. They ask why can we not trust them to do the right thing? The communication strategy is vital. We not only need to inform the researchers of these new arrangements but why the constraints are necessary. They are much more likely to work within the system if they understand the rationale.

We are really moving from a paradigm of risk avoidance to risk management. There are greater risks of a loss of public confidence in the degree to which we protect the confidentiality of their data. The risks may be small, and justified by the value being added to our statistical data, but they still exist. The value system of researchers is different to that of official statisticians. The research imperative dominates and researchers can be frustrated by what they see as unnecessary impediments and bureaucracy. It is inevitable that some will "step across the line". It is unlikely that a researcher will try to identify an individual - that is not the motivation. Rather, from our

experience, they are more likely to bend the rules to advance their research agenda (e.g. we have found cases of our microdata being on-sold to support further research albeit with added value). It is important that we act in these cases. Legal sanctions may be appropriate in some cases. These can be difficult and drawn out. Withdrawal of service, including from the host institution, is easy to apply and very effective particularly if the message gets around the research community that the ABS is prepared to undertake this step.

Finally, there is a lot of international collaboration among the research community. They will point out what they can do in country A compared with country B. We know from personal experience. There would be considerable benefits if there was a greater degree of uniformity in our approaches. We have agreed on a Fundamental Principles of Official Statistics - why not fundamental principles for use of microdata by external researchers? I elaborate on this in the next section.

Conclusions

Supporting external research use of our statistical data is an important way of getting more value out of our statistical activities. We regard this support as an important ABS objective. Furthermore, our legislation provides us with the authority to support this type of activity.

In the past, we have interpreted this legislation in a conservative way - focussing on approaches that result in safe data. The increasing sophistication of technology, and the availability of external databases make it more difficult to release truly "safe" microdata (or safe datacubes for that matter). The increasing prevalence of private sector databases may be the biggest concern as there is generally less regulation about their use or misuse.

Consequently, there is a need to move towards dissemination approaches that rely on a "safe environment". We have been assisted in this respect by confirmation that legally binding undertakings signed by researchers can be taken into consideration when assessing whether we are complying with our enabling legislation. That is, we do not need to rely on safe data alone.

We will continue to take a somewhat conservative approach to interpreting our legislative authority for releasing microdata, although not as conservative as previously was the case. This is because of our concern that one significant incident could create severe damage to our reputation and our ability to maintain public confidence in the degree to which we protect the confidentiality of their data. This will affect response rates in our collections and the quality of statistics we produce.

The most promising new approach to providing microdata access is RADL. Its use as a means of increasing access to linked data sets is of particular interest.

Like many statistical endeavours there is great scope to learn off each other - both good and bad experiences. Microdata access may be an area of activity on which we may want to agree on

some core principles. The research communities work across countries and make comparisons. In fact, there are already arrangements (e.g. Luxembourg Income Study) where microdata from several countries are brought together for convenient research access.

Legal and administrative arrangements will vary from country to country of course. But there still may be some core principles on which we agree. To start the debate, I suggest the following:

- it is appropriate for microdata collected for official statistical purposes to be used to support research and secondary data analysis under prescribed conditions that prevent misuse;
- the use of microdata to support external use for other than research and statistical purposes is not supported;
- there should be a legal or other arrangement to support use of microdata in order to increase public confidence in its appropriate use;
- the uses of microdata should be transparent, and publicly available, again to increase public confidence that microdata is being used appropriately;
- external researchers should not be engaged by the NSO as an employee unless they are contributing to work which will lead to a new output;
- the arrangements for microdata access should be cleared with the privacy authorities of the country.

IV.2 CHALLENGES FOR TRADITIONAL APPROACHES TO CONFIDENTIALITY PROTECTION – THE DANISH EXPERIENCE

Supporting paper by Lars Thygesen, Statistics Denmark

The Danish concerns

Confidentiality protection is one of the top priority policy issues of Statistics Denmark because of the conflicting interests between, on the one hand, making best use of the data and, on the other hand, protecting personal data of citizens and enterprises who have, directly or indirectly, entrusted their data to us. Therefore, the Board of Statistics Denmark made a number of important decisions on the issue in the past year, and will continue discussions.

It should be noted that because of the way statistics are produced in Denmark, making intensive use of data from administrative registers, Statistics Denmark possesses an immensely rich database with billions of individual data. Most of the data are endowed with precise identifiers and can thus be linked. These data are never passed on to outsiders, not even for research purposes ("the One Way Street"). However, the needs of researchers and analysts are so strong and qualified that they may be allowed, under specific circumstances and strict control, to access unidentified microdata.

Data users versus data subjects

The conflict between people who want to make use of data and respondents is a real one. But in recent years in Denmark, the data users have been much more prominent in the public debate than the data subjects.

It is obvious that the total data warehouse of Statistics Denmark is a virtual gold mine for researchers and others – including the Government and Parliament – who want to generate a strong and detailed basis for their decisions. The richness of the data on individual citizens and enterprises allows for innumerable valuable analyses. Statistics Denmark recognises this and has for many years sought ways to accommodate the needs in ways that do not threaten confidentiality.

There has been continuous and strong pressure, especially from researchers, to obtain access to "everything". Why should there be restrictions at all, since the researchers only want to identify structures in the data, not to snoop into the private lives of persons or enterprises? Danish researchers claim that practices of NSOs in other countries are more "research friendly". It is difficult to argue that Statistics Denmark must be more confidentiality conscious than other NSOs – and for this reason international discussions like this one in the Conference are important. The result of the pressure has been a steady move towards increased access.

At the same time, however, we have to take the perception of respondents into account. They have not been very active in the debate, maybe because they do not know exactly what is going on. Yet we see the risk of a major publicity disaster. This, together with the European data

protection legislation and its Danish implementation, makes it necessary to keep up strong measures. And the Board, having people from the world of business as well as researchers as members, is very much aware of the importance of protection, as well as the user perspective.

The Danish solution

The 2001 negotiations between Statistics Denmark, the Ministry of Research and the Research environment resulted in the signing of a contract on the establishment of a special unit (the Research Service Unit) in Statistics Denmark with the particular duty to improve researchers' access to microdata through a better infrastructure and to lower the costs of using the data. The budget for the Research Service Unit is 6 million D.kr. per year (approx. 800,000 Euro). Some of the money is used to upgrade the special Unix computers, cf. below.

Principles for access

Statistics Denmark has created an advanced solution allowing researchers and policy makers to have access to microdata – always unidentified – based on the Need to Know Principle. This means that they can only obtain the data needed for their purpose, and Statistics Denmark has to make the judgement on that. Microdata must never leave the NSO, but users obtain access to "their" data sets in Statistics Denmark via a virtual private network, the use of which is under strict control and observation. A log will be used for investigation in case of confidentiality violation.

The technical solution

Since 1986, Statistics Denmark has given access to researchers to analyse micro datasets from work stations in Statistics Denmark ("the in-house researcher arrangement").

In 2001, Statistics Denmark launched a new system granting access for specially authorized research and analysis environments to approved datasets from their own workplaces. A research or analysis environment can apply for an authorization from Statistics Denmark. As of 15 March 2003, 43 environments had been granted authorization. The wording of the authorization appears in Appendix 1.

The technical solution is based on a virtual private network on the Internet, see the chart in Appendix 2.

The relevant microdata are produced by the staff in Statistics Denmark and the de-identified microdata are transferred to the disk storage connected to special Unix servers dedicated to the researchers. These Unix servers are separated from the production network.

Communications via the Internet are encrypted by means of a so-called RSA SecurID card, a component that secures Internet communications against unauthorized access. In practice, the

researcher rents a password key (a token) from Statistics Denmark. The token ensures that only the authorized person obtains access to the computer system.

A farm of Citrix Servers ensures that the researchers from their own workplace can "see" the Unix environment in Statistics Denmark. All data processing is actually done in Statistics Denmark and data cannot be transferred from Statistics Denmark to the researcher's computer. The researcher can work with the data quite freely and can make new datasets from the original data sets.

All results from the researchers computer work can be stored in a special file and such printouts are sent to the researchers by e-mail. This is a continuous process (every five minutes) and has proven to be quite effective. The advantage to Statistics Denmark is that all e-mails are logged at Statistics Denmark and checked by the Research Service Unit. If the unit finds printouts with too detailed data, contact is made with the researcher in order to agree on details of the level of output. No severe violation of the rules, established in the authorisation formula, has taken place.

In December 2002, the Board of Statistics Denmark decided to consider the on-site scheme and the remote access scheme as equivalent concerning data security and, as a consequence of this decision, all data sets which can be accessed from on-site can also be accessed from remote. This has led to an explosion in the use of the remote solution, and it is envisaged that the in-house arrangement will soon become obsolete and will be closed down.

With this decision it has been very important to revise the rules for granting authorisation to microdata. The new rules can be seen in Appendix 3

Personal data versus business data

In Denmark we have found it expedient to distinguish between microdata on citizens and on enterprises, the latter containing more problems than the former, since it is difficult to avoid the fact that people obtaining access will immediately recognize well-known companies even without identifiers. For this reason, we have decided to restrict access to business microdata: data have to be at least one year old, and access cannot be granted to research departments of private companies.

Automated confidentiality protection

It has been argued in international discussions that protection may be performed using automatic tools, e.g. giving access to scrambled microdata or using tools like a "statistical fire wall", giving people access to process microdata without revealing the identity. We have not found evidence that this is a fruitful path, partly because it may distort the inference which the users are looking for; if this risk is avoided, we fear the risk that there may be ways to circumvent such tools, e.g. by multiple data requests.

Consequently, we have not found good solutions to providing general access to extremely detailed data via the web, e.g. geographical information systems where users would be able to compose their own detailed geographic breakdown “on the fly”. Our database on the Internet, www.statbank.dk, is extremely detailed and rich, but the levels of geographical breakdown are predefined and fixed.

APPENDIX 1: AUTHORIZATION FORM
Statistics Denmark

AUTHORIZATION

Statistics Denmark hereby grants

[Institution] represented by [Chief Researcher]

Authorisation for

Remote electronic access to selected datasets at Statistics Denmark

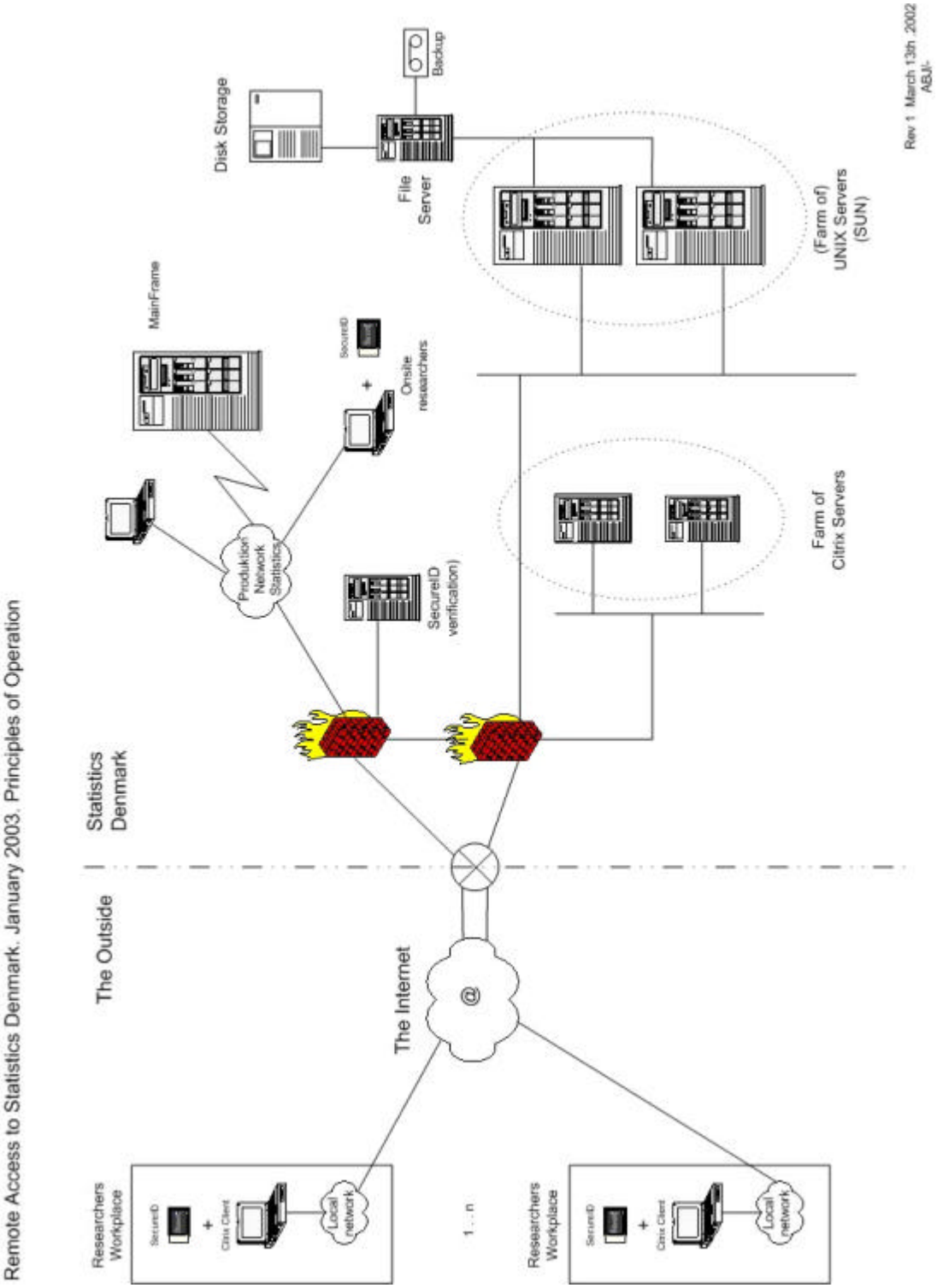
Remote Access via the Internet is subject to the following terms:

1. A project description must be submitted, which states the project objectives and renders it possible to select the data required for successful project execution.
2. Based on the project and data description, Statistics Denmark decides whether external electronic access to data can be granted for the specified project. If the authorisation is not granted, the researcher is referred to use the ordinary scheme for the on-site arrangement for external researchers at Statistics Denmark.
3. The researcher to whom external electronic access is granted shall sign a special agreement with Statistics Denmark, cf. appendix.
4. All datasets are confidential, cf. §27(3) of the Danish Public Administration Act and §152 of the Danish Criminal Code.
5. The researcher obtains access to make batch runs on Statistics Denmark's special researcher machines (UNIX system) from one or more PCs specially assigned for that purpose in the research/analysis environment. Access is denied for batch runs from remote PCs, PCs at home or PCs which cannot be properly supervised.
6. Only the client software assigned by Statistics Denmark may be applied in connection with the RSASecurID card provided. A PC connected to Statistics Denmark may not be made available to unauthorised persons, and when the user leaves the PC, the PC must be either shut down or disconnected, i.e., protected from any unauthorised use.
7. The password of the individual researcher is personal and strictly confidential.
8. The researcher may not, directly or indirectly, download the dataset or any datasets derived there from. All transfers of output for printing or further statistical processing (in spreadsheets or similar) must be executed in accordance with the guidelines and methods laid down by Statistics Denmark. Statistics Denmark will create a log file of such authorised transfers. Furthermore, individual records may not be printed, and all output must be aggregated to an extent that eliminates any risk of direct or indirect identification of persons or enterprises. The researcher may not attempt to make such identification.

9. Statistics Denmark shall be entitled at unannounced visits to check that the rules of this agreement are observed.
10. The person signing this agreement on behalf of the research/analysis environment shall ensure that publications by the environment do not contain any information that may identify individual persons or individual enterprises.
11. The person signing this agreement on behalf of the research/analysis environment undertakes personally to supervise or to appoint a person to supervise that the provisions of this agreement are observed.
12. In case of breach of the provisions of this agreement, the researcher in breach will be excluded from using any researcher schemes of Statistics Denmark permanently or for a period of not less then three years. Furthermore, in the case of breach hereof, this authorisation will be withdrawn for a period.

This agreement, which is signed in two copies, enters into force on [date] and may be terminated by either party at three months' notice.

APPENDIX 2: THE TECHNICAL SOLUTION



APPENDIX 3: RULES ON ACCESS TO DE-IDENTIFIED MICRODATA UNDER STATISTICS DENMARK'S RESEARCH SCHEMES

At its meeting on 2 April 2003 the Board of Governors laid down the following rules on access to de-identified microdata[1] under Statistics Denmark's research schemes:

Who can obtain access?

Access is only granted to authorised research and analysis environments. Only research and analysis environments of a more permanent nature with a chief researcher and several researchers/analysts can be authorised, as sanctions in case of violation of the rules would otherwise have limited effect. Authorisation is granted by the Director General. The authorisation form is shown in Appendix 1.

Prior to granting the authorisation, Statistics Denmark makes a concrete assessment of the applicant's reliability as a data recipient. In respect of non-governmental organizations and enterprises it is relevant to examine the ownership, the staff (qualifications) and the assignments handled for public-sector clients in particular. The examination may include enquiries to such clients to in order obtain a statement.

When a research or analysis environment has been authorised, agreements may be concluded with specific researchers/analysts.
The following environments can be authorised:

1. *The user group defined under the framework agreement between Statistics Denmark and the Ministry of Science, Technology and Innovation* can be authorised and thus obtain access. This group comprises all employees in government funded research projects, employees in public research and analysis environments (i.e., universities, government research institutes, ministries, government agencies, etc.) and researchers employed with non-profit foundations in Denmark.
2. In the private sector, the following organizations with research and analysis environments of a more permanent nature are eligible for authorisation:
   a. *Non-governmental organizations*
   b. *Consulting firms* may be authorised, but cannot generally obtain access to microdata containing business data. The Director General may grant exemption to consulting firms that carry out investigations or research for a public authority, or to a non-governmental organization that would be eligible for authorisation if its client guarantees, in writing, the correct use of data in terms of security.
   c. *Other individual enterprises* may be authorised, but cannot obtain access to microdata containing business data.
3. Danish researchers *who are working abroad for a period*, but who are attached to an authorised Danish research environment, may obtain external electronic access from their place of research abroad. In these cases the responsibility lies with the Danish research environment.

[1] There is no access to microdata with identification.

4. *Foreign research institutes* may, in exceptional cases and following a concrete assessment, be authorised to use the on-site research scheme for external researchers for as long as this scheme exists. They cannot use the external electronic access.
5. *Foreign researchers* who are working in Denmark in a Danish research environment may achieve external electronic access for the duration of their stay under the authorisation of the particular research environment, which assumes the overall responsibility.

What data can be accessed?

Access can only be granted to de-identified data, i.e., data for which all identification details such as name, ID number and address have been removed.

Access is granted according to the need-to-know principle, which implies that researchers/analysts can obtain access to the data required for the specified purpose.[2] Accordingly, the applicants have to document a reasonable relationship between the requested data content and the project description. If the issue requires total population coverage, access may be granted to total data material, otherwise a sample will be made available. In addition, data may be limited in the form of grouping or segments for confidentiality reasons.

Generally, authorised persons have access to all types of personal and business data with the limitations following from the above rules on consulting firms and other individual enterprises, and the need-to-know principle. However, anonymised business data cannot be accessed until one year after the reference period. Detailed product data on individual enterprises are anonymised before they are made available.

In certain cases, Statistics Denmark may deny the requests of a researcher or analyst because of insufficient data quality, primarily in connection with compilation of information from different statistical fields. This applies to both personal data and business data.

5.1 Cases of doubt
Acting on the recommendation of the heads of divisions, the Director General decides any cases of doubt resulting from interpretation of the rules.

[2] This is in accordance with the principles of the Danish Act on Processing of Personal Data, particularly section 5(3): "Data which are to be processed shall be adequate, relevant an not excessive in relation to the purposes for which the data are collected and the purposes for which they are subsequently processed"; and section 10(1): "Data as mentioned in section 7(1) or section 8 may be processed where the processing is carried out for the sole purpose of carrying out statistical or scientific studies of significant social importance and where such processing is necessary in order to carry out these studies."

IV.3 RESEARCH DATA CENTRES OF THE OFFICIAL STATISTICS

Supporting paper by Tom Wende and Markus Zwick, German Federal Statistical Office

Developments in informational infrastructure

On 1 October 2001 a Research Data Centre (RDC) of the German Federal Statistics Office was established in Wiesbaden and one in Berlin. On 1 April 2002, RDCs in the Statistics Offices of the federal states with one location in each federal state were founded. The RDCs offer a lot of opportunities for microdata access and thus an extraordinary improvement of the informational infrastructure between official statistics and empirical science.

The Research Data Centres provide a well-balanced service proposition for users. They are independent but cooperate closely with each other. The main focus of the Federal States Research Data Centres is centralised data storing, a widespread web of visiting researchers desktops[1] and a supply of metadata for decentralised surveys. The focus of the German Statistical Offices Research Data Centres is on the development of Scientific and Public Use Files[2], the improvement of controlled remote data processing[3] and the supply of metadata for central surveys. Together, all Research Data Centres are keen on developing a high quality metadata system, consulting data users and the further improvement of the informational infrastructure.

History

Before going further in this paper about the work of Research Data Centres and the development of a better informational infrastructure, it may be useful to take a short glance at the history of microdata use in Germany. In the past, it was considered sufficient for data users to work with aggregated data-like tables and indexes given out by the statistical offices. But the accelerating change of society and the increasing number of questions changed the scientific interest, and aggregated data was no longer enough. The first requests for official statistics microdata by scientists were made in the early 1970s. A group of scientists at the Universities of Mannheim and Frankfurt founded a research project called SPES that tried to create a "social-political decision- and indication system for the Federal Republic of Germany" ("Sozialpolitisches Entscheidungs- und Indikatorensystem für die BRD") by using official microdata. From this project evolved the so-called special research sector 3 (Sonderforschungsbereich 3 – SFB 3) which, through to the present day, deals with matters of social policy and econometrics. This pioneering work, which demonstrated the desire to use microdata for societal research, paved the way for still ongoing changes in the law and the development of an informational infrastructure for the empirical use of microdata bases. At almost the same time, a project called VASMA dealt with a comparative analysis of the social structure by population data.

[1] See VISITING RESEARCHER DESKTOP AND "ONE DOLLAR MAN".
[2] See SCIENTIFIC USE FILES AND PUBLIC USE FILES.
[3] See CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING.

Legal basis

The first legal regulation for the passing on of official microdata was included in the 1980 federal law on statistics. §11 subsection 5 allowed the passing on of completely anonymised microdata. This, of course, imposed a lot of restrictions, as anonymisation always involves a certain loss of information. But it was still an epoch-making invention, as it offered the first legal opportunity for official statistics to distribute so-called Public Use Files[4], which are completely anonymised datasets of social statistics, to everyone who needed such information. This regulation also showed the way ahead. A more satisfying solution for empirical researchers was the next legal improvement: the 1987 federal law on statistics brought about the so-called "privilege of science", which means that from that point on, scientists were allowed to receive factual anonymised microdata. Factual Anonymisation means that the data is not absolutely anonymised, so there is a chance to de-anonymise data and to draw conclusions back to single persons or organizations, but the work involved in de-anonymisation is too great to make it worthwhile[5].

Excursus 1: The Development of Anonymisation Criteria

Between 1988 and 1991, a large-scale research-project aimed at anonymisation of selected microdata was completed. Representatives of the Federal Statistical Office and of the statistical offices of the federal states worked alongside representatives of the data protection registrars of the federal states and the Federation, the University of Mannheim and the ZUMA – the Centre for Survey Research and Methodology. This project was directed by Prof. Dr. Walter Müller at the University of Mannheim, mainly because it was concerned with the factual anonymisation of person- and household related microdata, which can be of particular interest to social sciences. In the course of this project some measures were developed for a specific factual anonymisation of the sample survey of income and expenditure and the Microcensus. Special issues were, for example, coarsening of data files and drawing of sub-samples. The results of this research project culminated in two similar reports: "Textbook for the building of factual anonymised data regarding the Microcensus" and "Textbook for the building of factual anonymised data regarding the sample survey of income and expenditure".

End of Excursus

National and international data requests

In the previous section, the access to Public (PUF) and Scientific Use Files (SUF) was described. You may ask yourself why this is such an important development: the example of international data access will illustrate this. Before PUF and SUF, access to microdata was almost impossible. After 1980 - with the invention of Public Use Files - data access was possible for everyone, but with a lot of restricted and non-accessible information. After 1987, scientists' requirements for less restricted data were solved with the invention of Scientific Use Files, which

[4] See SCIENTIFIC USE FILES AND PUBLIC USE FILES.
[5] See SCIENTIFIC USE FILES AND PUBLIC USE FILES.

are today only provided to German scientists. But what if a scientist from a foreign country wants access to German official single data? Today, it is almost impossible. One solution is the EU-Regulation 831/2002, which regulates the data access possibilities for members of the EU, such as controlled remote data processing[6] or the possibility for a guest researcher to work in the safe area of the statistical offices[7]. For Non-EU countries, the legal situation is still unsatisfactory.

Research data centres (rdc) of the official statistics

Due to this exigency, resulting from the conflict between data-protection and the eminently reasonable interest in microdata access by the science community, the federal ministry on education and research (BMBF) created a "commission for the improvement of the informational infrastructure between science and statistics" (KVI). It was the constitutional task of the commission to revise the informational infrastructure of the Federal Republic of Germany (BRD) with respect to its capacity and to work out new concepts on the exchange of data between science and statistics. The KVI elaborated some advice which is described in detail in their final report[8].

Some elementary advice from the KVI was the establishment of the so-called Research Data Centres (RDCs). The implementation of this advice was almost immediate. The main functions of the RDCs are:

- to continue further development and implementation of the advice given by the KVI;
- to serve as an interface between official statistics and science;
- to provide consultations and service for the use of official microdata;
- to create and provide possibilities for access to microdata with a lower level of anonymisation.

The invention of the RDC is a great improvement for the informational infrastructure because, for the first time, every service related to official microdata access is located in one place. There are different ways to access official microdata such as controlled remote data processing and visiting researcher desktops[9]. The RDCs also offer consultation and service for the use of official microdata. Let's now talk about the Research Data Centres' work in practice.

As has already been mentioned, the RDC offer different ways to access microdata:

- Scientific and Public Use Files;
- Visiting Researcher Desktop;
- Controlled Remote Data Processing;
- Special Data Processing.

[6] See CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING.
[7] See VISITING RESEARCHER DESKTOP AND "ONE DOLLAR MAN".
[8] KVI (HRSG.) 2001: WEGE ZU EINER BESSEREN INFORMATIONELLEN INFRASTRUKTUR. BADEN-BADEN: NOMOS VERLAGSGESELLSCHAFT.
[9] See CONTROLLED REMOTE DATA PROCESSING AND SPECIAL DATA PROCESSING.

Scientific use files and public use files

The first possibility for a scientist to access microdata is to purchase a Scientific or Public Use File (PUF). Different surveys are already available in that format. For example, you can obtain the Microcensus, the Sample of Income and Expenditure or the Statistics of Road and Traffic Accidents and many more. The Time Use Survey, the Wage and Income Statistics or the Social Welfare Statistics are also available as Public Use Files.

One important aim of the Research Data Centres is to broaden the range of PUF in the near future. Scientific and Public Use Files are anonymised with different grades of anonymisation. The Public Use Files offer no way to draw conclusions about single characteristic carriers anymore. The Scientific Use Files do theoretically offer that possibility, but the expense is much higher than the use of de-anonymising the factually anonymised data[10]. The rights to use Scientific Use Files are reserved – as the name implies – to scientists (at the moment almost without exception to German scientists). That is another confidentiality function of these files, because in case of a breach of confidentiality the scientist can be prosecuted by law. The advantage of giving out anonymised files is that the scientist is able to work with his own Software on his own PC; the disadvantage is the loss of information resulting from anonymisation and consequently the difficulty to close from the sample to the complete population.

Excursus 2: Anonymisation Procedures

For a better understanding of the problem, it is necessary to know how anonymisation is realised in practise. Basically there are three ways of anonymizing data: enlarging the pitch, clearing critical data and drawing of samples. For example: imagine an offender gets paid by a company to find out strategic information about competitors of this company (this example was chosen for its simplicity; actually it is not yet possible to anonymise company data to an extent sufficient for a Scientific Use File, but right now specialists from the official statistics are working together with highly decorated scientists to find a solution to that problem). First of all, de-anonymisation can occur if an offender connects different information – maybe different variables in the data set or external information about the population.

With Scientific Use Files, de-anonymisation is impossible most of the time without additional knowledge about the population, because the internal critical variables are already anonymised. So the offence scenario is based on the fact that a data offender is able to connect knowledge which he receives from the data set with knowledge from other sources like the Internet or other Scientific Use Files of earlier surveys about the same topic. Let's look at ways of reacting: enlarging the pitch means, for example, if you have a very small high-turnover class of companies in one region - maybe less than three - you can easily find out who these three are through the press or Internet and then research concrete data in the dataset.

[10] §16(6) BstatG.

De-anonymisation is easy to prevent if you include high and very high turnover in one category or if you enlarge the geographic area, i.e. if you include two or three small regions into one region, like the federal states, or the different points of the compass into north, south, east and west. Another possibility is to cut off critical data, like the turnover numbers, but that involves a certain loss of information and research quality. The third way is to draw a sample of the original file, with the result that the offender does not really know if all companies are in the sample and therefore cannot estimate exactly whether all very high turnover companies are in his sample.

End of Excursus

If SUF and PUF were the only ways to research microdata, a lot of empirical questions would remain unanswered. But the Research Data Centres offer some more possibilities of data access which, in combination with the supply of Public and Scientific Use Files, close the circle of informational infrastructure and in combination with each other are able to provide a more satisfying balance between empirical research interests and data confidentiality. Above all, there are the visiting researchers desktop and the option of controlled remote data processing, special data processing, which will be described in the following chapters.

Controlled remote data processing and special data processing

If the researcher needs more information than a Public or Scientific Use File can offer, there is a way to work with less or even non-anonymised data via the Research Data Centres. One way is to work, as a first step, with the anonymised dataset, for example a Scientific Use File - or if a SUF is not available, with a so-called structural dataset, which corresponds to the original dataset in all structural attributes but not in content attributes. The second step is to send the thus produced syntax for Software like SAS, SPSS or STATA back to the RDC, where it is processed under internal control over the original data. This is called Remote Controlled Data Processing. A special form of Remote Controlled Data Processing is Special Data Processing, where the scientist informs a representative of the Statistical Office of his research interest, and the representative does the empirical work. As you can imagine, Special Data Processing is more cost-intensive than simple Remote Controlled Data Processing and, on top of this, it is unsatisfactory for a scientist to let others do his work.

One advantage of Remote Controlled and Special Data Processing for data confidentiality is that the computing process is not beyond control and the representatives of the Research Data Centre know exactly what information is given to the researcher. Another advantage is that the output is not microdata but aggregated data in the form of tables, which can be anonymised more easily. The advantage for the researchers is that they have the possibility to make an exact predication about the whole population with a lower standard error and in general a low error variance. Further advantages are that the consulting function of the research Data Centres can be engaged and there is a possibility to work with company data, an option which did not exist before. The disadvantages are that these processes mean a lot more work and cost for both the scientist and the representative and, as a result, require a lot more time.

Visiting researcher desktop and "one dollar man"

The RDCs provide another new method of data access in the protected area of the German Statistics Offices. The empirical researcher can access microdata over sealed-off computers at the visiting researchers desktop in the statistical offices. Generally speaking, there are two different methods of data access for a visiting researcher. One way – the Visiting Researcher Desktop – is the future, the other way – the "One-Dollar-Man" – is the past. In the past, it was possible for a scientist to sign an employment contract (with the symbolic payment of one Dollar) with a statistical office and to work with microdata in the area of the statistical offices as an employee and therefore be bound by confidentiality like every employee of the statistical offices. But that was a very excessive solution and has now been replaced by the regulated method of the visiting researcher desktop, where the researcher remains as an employee of his actual institution and obtains on-site-access to factually anonymised data as a guest of the Research Data Centre. The difference in anonymisation with the distributed Scientific Use File, which is also factually anonymised, is that the anonymisation criteria in the latter case are lower, because of other means of confidentiality control, such as the fact that the guest researcher can only take aggregated data – in the form of tables – out of the statistical office. He also has no method of data transfer other than his aggregated output. The researcher is given a special password protected folder, where he is given the ability to save his research data for a limited period in the statistical office.

Projects in the future

The Research Data Centres are working on the expansion of low cost microdata access in the form of Scientific and Public Use Files. Also the production of PUF and SUF for on-site use will be forced. Controlled Remote Data Processing will be simplified and improved in the future. Looking further ahead, there will be an improvement of consultancy capacity for visiting researchers and researchers who use controlled remote or special data processing. The RDCs are working on the central availability of all official microdata and also on the elaboration of a widespread metadata-system for all official data.

IV.4 ACCESS TO MICRODATA – THE DATA STEWARDSHIP MODEL OF THE U.S. CENSUS BUREAU

Supporting paper by Gerald Gates, Patricia Doyle, Sam Hawala, Arnold Reznek and Rochelle Wilkie Martinez, U.S. Census Bureau[1]

Public use microdata: importance, limitations, and threats

To conduct public policy analysis and research, federal, state, and local policy makers - and researchers from many disciplines - rely heavily on the Census Bureau to provide high quality information on the population and the economy of the U.S. The Census Bureau makes these data available to external users in the form of tables or public use microdata files that have been "disclosure proofed" to protect the identity and privacy of respondents.

A recent book and conference on confidentiality and data access brought home the Census Bureau's growing challenge to maintain its historical commitment to respondent confidentiality and still meet the American public's growing data needs (Doyle et al., 2001). The latest research suggests several reasons we will have a problem maintaining confidentiality (as defined by current legislation) in the future if we continue with our current disclosure and data dissemination methods:

- there is a growing wealth of individual- and business-level information available in the public domain;
- data from other agencies that do not follow strict disclosure guidelines are publicly available;
- technology to mine public information is increasing in sophistication;
- the general public has increasing concerns over privacy.

To meet critically important public policy and research needs, the entire federal statistical system faces increasing demands for more, better, and more recent data. The Census Bureau is responding to this situation with improved disclosure techniques, but the methods that reduce disclosure risk also reduce the level of detail and the quality of the data disseminated publicly. To remain the pre-eminent provider of data for public policy and research, the Census Bureau must be proactive in addressing the challenges posed by the simultaneous increase in stress on our system of maintaining confidentiality and increase in demand for our data.

A change from our current approaches to disclosure and dissemination would involve recognizing that disclosure risk is composed of both opportunity and incentive. Our disclosure practices to date have focused on minimizing opportunity, since we have had little control over incentive. However, to make substantial strides toward making public data more usable, without disclosure risk, we need to extend our focus to address the incentives users have for attempting to identify individuals in public use microdata. Activities needed to minimize both opportunity and

[1] This paper reports on results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited Census Bureau review than official Census Bureau publications. The report is released to inform parties of research and to encourage discussion on work in progress.

incentive involve technological advances, legal strategies, policy enhancements, interagency coordination, new disclosure techniques, and privacy research.

Technical advances are those that allow methods like remote access to be more useful substitutes for public use microdata files. These also include new techniques to reduce opportunities for disclosure. Legal strategies are those that would provide shared data protection responsibility with the user or would severely penalize anyone conducting data linkages to identify individuals in Census Bureau data. Policy enhancements consist of formal guidelines related to confidentiality and privacy aspects of collecting data, controlling access to data, linking data, and providing data for research uses. Interagency coordination takes a Government system-wide view of disclosure risk, since the biggest threat to public use microdata tends to come from administrative data maintained by other government agencies. Finally, we need to pursue research in how to improve communication of our confidentiality procedures in a way that bolsters the confidence of our respondents (rather than calls their attention to an unrealistic potential for misuse of the data). While we consider how best to pursue each of these areas, we cannot lose momentum in the core approaches that depend on disclosure avoidance techniques.

Disclosure avoidance techniques: overcoming new threats

The usual review and approval of a release of new microdata sets requires judgments by reviewers based on, among other things:

- the size of the geographic entity - either directly identified by the Census Bureau or indirectly identified by contextual variables (such as sampling information, area mean income, population density, or percent minority population);
- the proportion of the study population included in the sample;
- the sensitivity of individual data items;
- the age of the data.

Notwithstanding the fact that released data contain no direct identifiers (such as name, address, telephone number, social security number), statistical disclosure limitation (SDL) experts recognize that the release of "truly safe" microdata (or raw individual data records) is extremely difficult. Data releases do not preclude, by all means, the disclosure of the individual respondent's identity. However, data are released in such a way that attempts at re-identification would require investments in manpower, time, and other costs that would be unreasonably high. In light of rapid changes in the technological and data environment, there may be an increased risk that a data user could match microdata records to another file containing identifiable information with reasonable accuracy - leading to the discovery of identities or of sensitive information. To better understand these types of elevated risks of disclosure, the Census Bureau conducts re-identification experiments to attempt matching files with overlapping information.

Re-identification experiments can shed additional light on the particularities of a microdata set. Hence, before the Census Bureau releases a microdata set, the Disclosure Review Board may

decide to consider some additional information on the nature of the data file. The information includes:

- the number and distribution of unique records;
- the amount of error in the data;
- the availability of external files with comparable data content[2];
- the resources that may be needed by an "attacker" to identify individual units.

Experience in re-identifying respondents from de-identified microdata sets show that the experiments should be run on a periodic basis to continually update SDL strategies. This is especially true for microdata sets published from recurrent large-scale sample surveys. Re-identification research is only one of the research areas the Census Bureau relies on to update SDL strategies. Research areas also target other aspects of dealing with disclosure risk such as measuring the risk, modifying of the data, and releasing synthetic (not observed) data.

Measuring disclosure risk for a microdata set usually entails the study of unique combinations of values in the data, and an assessment of whether an intruder can infer whether given sample unique records are also population unique (Bethlehem, Keller, and Pannekoek, 1990; Feinberg and Makov, 1998; Skinner and Elliot, 2002; Skinner and Holmes, 1998; Zayatz, 1991). Most work in this area assumes that there are no measurement errors in the data and that sub-sampling and other aspects of data releases are often not sufficient to protect against disclosure. Once records at risk of disclosure are identified, or a measure of disclosure risk for the entire file is calculated, traditional SDL strategies center on reducing the amount of information released. The Census Bureau considers statistical data as a public good and, therefore, does not want to rely on this as the best response to disclosure risk.

Methods of modifying the data include data swapping (Willenborg and de Waal, 2001) and adding noise (Kim, 1986). Records or blocks of records that are unique in their geographic area are sometimes swapped with partnered records or blocks of records that have identical characteristics but are in different geographic locations. The proportion of records that are swapped has a direct affect on the quality of the data. The Census Bureau modifies quantitative data - such as dollar amounts, travel time and dates - by adding small random quantities or noise, without affecting certain characteristics of the distributions of the original data. However, it is not possible to guarantee that the results of all analyses that can be done using the original data are reproducible using the perturbed data.

An alternative to releasing confidential observed data is the release of fabricated or synthetic data (Raghunathan, Reiter, Rubin, 2003, and Abowd, Woodcock in Doyle et. al. 2001). The obvious advantage of this method is that releasing entirely simulated data guarantees protection of respondents' confidentiality. One drawback is that the quality of inferences from the synthetic data

[2] All forms of public or propriety external files are considered: other microdata files, macro data files (or tabular data), and databases allowing queries of microdata records.

depends on the imputation models. The research in this area follows earlier, related but different, research efforts on masking microdata (Cox, 1994) to preserve confidentiality.

Restricted access: the census bureau's center for economic studies and its research data centers[3]

Several modes exist for providing restricted access to confidential data while limiting the risk of their disclosure. The Census Bureau has adopted (and pioneered) Research Data Centers (RDCs). RDCs permit restricted use of confidential files at secure sites under Census Bureau control, using limited access to dedicated computing equipment and enhanced physical and computer security.

Protecting the confidentiality of the data and ensuring their appropriate use are paramount in establishing and operating RDCs. To accomplish this requires several activities: providing physically secure offices and secure computer systems; selecting projects that use the data appropriately, benefit Census Bureau programs (as required by law), and present low disclosure risks; imparting to researchers at the RDC the Census Bureau "culture of confidentiality"; putting in place policies and procedures that protect confidentiality in the RDC office; and releasing only research output that is within the scope of approved projects and that does not reveal confidential information.

Each RDC has a security plan developed and approved according to established Census Bureau procedures. The RDC office is in a restricted access environment with locks and key cards that meet Census Bureau specifications. In response to increasing concerns about security (and to promote efficiency), the Census Bureau RDC system is now completing conversion from secure local RDC networks of PCs and Unix workstations to a centralized "thin client" environment. Under this arrangement, data are stored on secure servers at the Census Bureau headquarters. The RDCs are connected to the servers via dedicated T-1 lines. From the RDC offices, researchers use X-terminals ("thin clients") to access the data authorized for their projects. No confidential data are stored at the RDCs. Researchers are accountable for their computer use, through the use of passwords and system logs. Researchers have no access to any non-Census Bureau network (including the Internet) from within the RDC facility. They may not bring laptop computers or other portable mass storage devices into the RDC facility.

Access to an RDC facility is given only to Census Bureau employees or other persons with special sworn status (SSS) who are approved to use the facility - including researchers carrying out active, approved projects at the RDC. To be granted SSS, any researcher must have an approved project, must obtain a security clearance, and must sign the Census Bureau's standard sworn agreement to preserve the confidentiality of the data. Researchers are given access only to the confidential data needed for their approved projects. Persons with SSS are subject to the same legal penalties for revealing confidential information as are regular Census Bureau employees - up to a $250,000 fine or five years in prison. Another equally important legal requirement for SSS is that the researcher's project must benefit the Census Bureau's data programs. The Center for Economic

Studies and its RDC partners have set up a formal project selection process to ensure that all approved projects satisfy these requirements[4].

The Census Bureau stations a Center for Economic Studies' employee (the RDC administrator) at each RDC. Among the administrator's most important duties are to instil the Census Bureau's "culture of confidentiality" into the researchers and to train the researchers regarding the security and confidentiality restrictions. The administrator also examines any research output a researcher wishes to remove from the secure facilities – to ensure that the output is covered under the approved project and to prevent the release of confidential data. This examination of research output is called disclosure analysis. In carrying out disclosure analysis, the administrators use disclosure avoidance techniques.

Perceptions of confidentiality: the lurking threat to microdata

Beyond the quantifiable threats to microdata from intruder attacks and security breaches lies the little understood - but no less important - field of public perception (see Gates, 2001). Data collectors must not only be confident in their ability to protect data from determined intruders, but must also be confident that the public believes the collectors have taken all necessary precautions. In the past, the public (in its role as survey participant) was mostly unaware of who used the survey results and how they used them. Today, with our ability to make data easily accessible to the masses through the Internet, the survey participant has become the survey user. That fact, combined with advances in data mining and data fusion methodology, creates a real risk that the public will not support the data access approaches that have served so well in the past. Our challenge is to ensure that the data we release are clearly labelled for what they are and what they are not.

As a result of declining mail response in the 1990 census, the U.S. Census Bureau has been concerned that individuals' concerns for privacy may be playing an increasing role in their decision to provide information in our census and surveys. Census Bureau surveys of public attitudes have attempted to measure what the public knows and thinks about our legal requirements and our practices. We have found that the majority of the U.S. population does not believe we keep their personal information confidential - even though we have legal requirements to do so and strongly convey this message to all potential survey participants (Gates and Bolton, 1998). The extent to which attitudes will ultimately influence an individual's decision to participate in a survey is not well understood. Nevertheless, just as we cannot take a risk that our data products are vulnerable to attack, we cannot take the risk that misunderstandings about data access and protection procedures will cause respondents and potential respondents not to respond to our survey.

Some examples of possible misperceptions that could result from new access tools and methodologies include:

[4] For more details on the project selection process, see the CES Web site: <http://www.ces.census.gov>.

- "finding oneself" on a public use microdata file (a relatively easy matter);
- questioning the occurrence of a cell of size one or two on a table where data may have been swapped or perturbed;
- being able to use published data to isolate and profile sensitive population groups;
- learning that data miners can combine data from diverse sources with new technology and methodological tools;
- questioning the agency's commitment to confidentiality when researchers are permitted access under special agreements.

These examples can potentially lead to negative reactions and signal the need to better understand how activities that seem so reasonable and appropriate may create misunderstandings. Once we understand these concerns, we need to develop education and awareness programs to address them. Fortunately, we have new avenues to interact with the public. In the past, our only contact came at the time of interview. We have always provided our respondents with basic information on our authority to collect the data, the purpose and uses for the information, and our pledge to keep the information confidential. Today, we have re-established contact with the respondent in his new role as data user. That fact creates both the problem and the solution.

By way of the Census Bureau's Internet dissemination tool, the American FactFinder, we reach millions of novice data users who now can access the entire decennial census data files and request tables of their choosing. The process is fast, easy, and free. Our challenge is to take this opportunity to reinforce the messages we provided at the time of collection and to address any misperceptions that may arise. With this technology, we can target the messages to specific users and their specific concerns by providing general information (at first) and progressing to more specific details (if desired).

The challenge is not so much in how to deliver the message, but rather in what messages to convey. The Census Bureau has approached this in two ways: public opinion surveys and cognitive research. Through public opinion surveys, we have learned what relevant beliefs about privacy and confidentiality are most widely held. Since attitudes are affected by personal experiences and societal events, it is not sufficient to measure attitudes at only one point in time. Surveys need to be conducted periodically and trends monitored. Results will identify key areas of concern that may translate into changes in behaviour (for example, reluctance to participate in surveys).

Armed with this information, we are able to develop and cognitively test messages that are clear, understandable and relevant. As research has shown, what may be intuitively appropriate is not always the best option. For instance, work done by Singer shows that overemphasizing the confidentiality promise at the time of data collection can have the unintended consequence of raising concerns that were previously not expressed (Singer, Hippler, Swartz, 1993). Cognitive interviewing and focus groups will offer insights into where these perceptions lie and how to best alleviate them.

Data stewardship approaches to confidentiality and data access

In the last few years, the Census Bureau has introduced a data stewardship approach to making decisions about how to collect and provide useful data: balancing data quality and access on one side of the scale and privacy and confidentiality on the other. The concept of "stewardship" is borrowed from environmentalists, the objective being to create a sustainable balance that supports one's needs over the long term.

In June 2001, the Census Bureau established the Data Stewardship Executive Policy (DSEP) Committee. The DSEP Committee is composed of top agency executives who are charged with identifying and developing policies related to data stewardship. This executive decision-making body is staffed by the Policy Office and supported by the analyses and recommendations of four staff committees, including the Disclosure Review Board (Potok and Gates, forthcoming).

One goal of the DSEP Committee is to ensure that strategic goals, corporate ethics, policies, controls, and operational practices are integrated and consistent. This means that strategic goals are shaped by corporate ethics and drive policies. Policies in turn drive the creation of organizational controls, and these controls incorporate practices that ensure compliance.

The Census Bureau has considered a number of sources for guidance in strengthening its data stewardship approach. We conducted a benchmarking exercise, a literature review, and an evaluation of the DSEP structure; and we drew on a U.S. General Accounting Office report published in 2001. From these sources, we gained an understanding of four pillars needed to strengthen our data stewardship program:

- culture and tradition;
- awareness and outreach;
- an integrating authority, such as a Chief Privacy Officer;
- technical and administrative tools.

The final item includes providing safe settings (such as RDCs), releasing safe data (by applying disclosure avoidance methodologies), as well as introducing automated tools that restrict access and limit uses within the organization. Finally, it includes ongoing research to ensure that these tools remain up-to-date.

At this writing, the Census Bureau is deliberately working toward full implementation of an enhanced data stewardship framework, based on the four pillars listed above. In so doing, the Census Bureau is also responding to new U. S. Office Management and Budget requirements for privacy impact assessments. These requirements offer an opportunity to integrate principles and policies into ongoing reviews throughout the lifecycle of data collections and supporting systems - allowing proactive planning to minimize risks (including those that are disclosure related). A key component for these assessments will be to build on a set of four privacy principles and sub principles that the Census Bureau identified as the ethical basis for the data stewardship structure.

The principles cover mission necessity, informed consent, protection from unwarranted intrusion and confidentiality.

It is important to note that developing and maintaining a viable data stewardship structure requires a significant commitment and investment of resources from an agency. Nevertheless, this more structured approach to data stewardship is integral to striking a balance between the tensions inherent in meeting data user needs and honouring the privacy and confidentiality commitments to its respondents. In the end, privacy and confidentiality - which are typically perceived as business constraints - can actually enable an agency's mission and business objectives by establishing the public's trust and cooperation as respondents.

References

Abowd, M.J. and D.S. Woodcock (2001). Disclosure Limitation in Longitudinal Linked Data, in Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds.), Amsterdam: Elsevier Science B. V., 215-277.

Bethlehem, J.G., W.J. Keller, and J. Pannekoek (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, Alexandria, VA: American Statistical Association, 85: 38-45.

Cox, L. (1994). Matrix Masking Methods for Disclosure Limitation in Microdata. Survey Methodology, Ottawa: Statistics Canada, 20, 165-169.

Doyle, P., J.I. Lane, J.J.M. Theeuwes, and L.M. Zayatz, eds. (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Amsterdam: Elsevier Science B. V.

Feinberg, S.E., and U.E. Makov (1998). Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data, Journal of Official Statistics, Stockholm: Statistics Sweden, 14: 385-397.

Gates, G. (2001). A Holistic Approach to Confidentiality Assurance in Statistical Data. Statistical Journal of the United Nations Economic Commission for Europe, United Nations, 18: 299-307.

Gates, G., and D. Bolton (1998). Privacy Research Involving Expanded Statistical Use of Administrative Records, Proceedings of the Government Statistics and Social Statistics Sections of the American Statistical Association, Alexandria, VA: American Statistical Association: 203-208.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, Proceedings of the Section on Survey Research Methods, Arlington, VA: American Statistical Association: 370-374.

Potok, N., and G. Gates (forthcoming 2003). Federal Committee on Statistical Methodology. Statistical Policy Working Paper 35, Washington, DC: U.S. Office of Management and Budget.

Raghunathan, E.T., P.J. Reiter and B.D. Rubin (2003). Multiple Imputation for Statistical Disclosure Limitation, research report, Washington, DC: U.S. Census Bureau.

Singer, E., H. Hippler and N. Swartz (1993). The Impact of Privacy and Confidentiality Concerns on Survey Participation, Public Opinion Quarterly, 4: 256-268.

Skinner, C.J., and M.J. Elliot (2002). A Measure of Disclosure Risk for Microdata, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64: 855-867.

Skinner, C.J., and D.J. Holmes (1998). Estimating the Re-identification Risk Per Record in Microdata, Journal of Official Statistics, Stockholm: Statistics Sweden, 14: 361-372.

U.S. General Accounting Office (2001). Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information, Report Number, Washington, DC: GAO-01-126SP.

Willenborg, L., and T. de Waal (2001). Elements of Statistical Disclosure Control, New York: Springer.

Zayatz, L. (1991). Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample. Statistical Research Division Report Number: Census/SRD/RR-91/08, Washington, DC: U.S. Census Bureau, <http://www.census.gov/srd/papers/pdf/rr91-08.pdf>.

BIOGRAPHICAL NOTES OF THE AUTHORS

BERIDZE, Teimuraz A., Education: MA, Economics, Tbilisi State University, 1978; Doctor of Economic Sciences, Moscow Institute of Economics, 1992. Appointments: Senior Researcher, Institute of Economics, Georgian Academy of Sciences, 1983-86, 1988; Visiting scholar, Moscow Institute of Economics, 1986-87; Head, Microeconomics Department, Georgian Academy of Sciences, 1993-95; Professor, Tbilisi State University; Visiting Professor, University of Maryland, USA, 1995-96; Chairman, State Department for Statistics of Georgia, 1996-. Publications include: The Republic of Georgia: Problems of Transition to a Market Economy, 1996; Industrial Policy and Trade Regime in Georgia, 1997; Numerous articles and papers in professional journals. Honours: Medal, Georgian Academy of Sciences, 1987; Fulbright Fellow, University of Maryland, USA, 1995-96. Memberships: The Georgian Academy of Economic Sciences, 1996; New York Academy of Sciences, 1997; International Statistical Institute, 1998.

BICIUNAS, Sigitas is Head of the data security and confidentiality service in Statistics Lithuania. He graduated from the Vilnius University, faculty of Mathematics and Informatics with a Bachelor and Master of Science degree. Since 1999, he works for Statistics Lithuania in the position of Head of data security and confidentiality service.

COOK, Len was born in Dunedin, New Zealand, in 1949. He attended the University of Otago and graduated with a BA (Hons) in Mathematics and Statistics, before joining the New Zealand Department of Statistics in 1971. In 1992 he was appointed Government Statistician in New Zealand. In May 2000 he crossed the globe to become the United Kingdom's first ever National Statistician. His contract in the UK was recently extended to 31 December 2005. He has a wide range of professional interests, such as social policy, superannuation, taxation, demography, statistical methodology and marketing. Earlier in his career he sat on the Royal Commission on Social Policy (1987/1988) and was a member of the secretariat of the Prime Minister's Task Force on Tax Reform (1981/1982). Len Cook is also concerned with the application of technology to information issues, and his more recent duties in New Zealand included chairing the Chief Executives' information technology steering committee and the Board of NZ Government Online.

DOYLE, Pat works at the US Census Bureau as Survey Improvement Coordinator for Demographic Programs, and leads the program of research and development for the 2004 Panel of the Survey of Income and Program Participation.

GATES, Gerald works at the US Census Bureau. He is Chief of the Policy Office. He leads the development of Census Bureau policy on issues related to confidentiality, privacy, data access and administrative records use.

GUDKOVA, Nataliya joined the state statistical office of the Kyrgyz Republic in 1972 as a specialist in labour and wages statistics division. In 1974 she graduated from the Kyrgyz State University, specialising in “Economy of Labour”. In 1993 she became Head of Labour and

Employment Statistics Division. She dealt with employment statistics improvement, and the development of special republican programmes supporting employment. She has also released publications on population employment. In 1998 she was appointed Head of Methodology and Statistics Organization Division. She coordinates statistical activities both at central and regional levels. She is a Deputy Chairman of methodological board, Chairman of a task group developing statistical documentations, and is the responsible person for the preparation of a publication "Methodological Provision on Statistics" for regional and local statistical offices.

HARRIS-KOJETIN, Brian is a Statistician at the United States Office of Management and Budget. He chairs the Federal Committee on Statistical Methodology and is also engaged in interagency work on confidentiality, statistics on the elderly, non-response in surveys, measures of educational attainment, and measures of insurance coverage. His other areas of work and interest include survey methodology, standards for statistical surveys and data quality.

HAWALA, Sam, PhD, works at the US Census Bureau as Mathematical Statistician in the Statistical Research Division, where he conducts research in microdata disclosure protection.

KING, John is now a consultant statistician, having recently retired from the Office for National Statistics (ONS) in the United Kingdom. At the time of writing this paper he was a Detached National Expert in the areas of Statistical Confidentiality and Statistical Disclosure Control in Eurostat on secondment from ONS. John King studied statistics and economics at the London School of Economics and is a member of the ISI and a Fellow of the Royal Statistical Society. He has worked in official statistics for over 25 years. In ONS he was responsible for the Family Expenditure Survey for 6 years. Earlier, for many years he was in the UK's Overseas Development Administration (now DFID) and worked with statistics offices in Africa, the Far East and the South Pacific, including a spell as Government Statistician in Fiji. This work ranged widely over economic and social statistics, including population censuses, balance of payments, national accounts and SAMs, consumer price indices, tourism statistics and household surveys.

KUDABAEV, Zarylbek is chairman of the National Statistical Committee of the Kyrgyz Republic since 1997. From 1995-1997, he was head of the Free Economic Zones Department of the Apparatus of the Government of the Kyrgyz Republic. Before, he was professor at the Kyrgyz Technical University. He has a master in Physics, a scientific degree in physico-mathematical sciences and a PhD in economic sciences. He has published various articles and monographs, both nationally and internationally, and has contributed papers to international conferences.

LANE, Julia is the Director of the Employment Dynamics Program at the US Urban Institute and, in that capacity, has worked with and for a number of statistical agencies across the world, as well as in the capacity of Professor of Economics at American University and as a consultant for the World Bank.

LYBERG, Ingrid is department statistician for Publishing and Information at Statistics Sweden since 2001. Before that she was Head of the department of Labour Market and Education Statistics

and prior to that Head of the department of Population and Welfare statistics. She has a PhLic in statistics from the University of Stockholm and worked as survey statistician within Statistics Sweden for a number of years before she was appointed as Head of department.

NIVA, Matti is chief international officer at Statistics Sweden since 1997. Before joining Statistics Sweden in 1989, he worked as lecturer and researcher at the Nordic Institute for Urban and Regional planning 1978-1989, and prior to that at the Department of Economics, University in Turku 1973-1978.

OSAULENKO, Olexander is Chairman of the State Statistics Committee of Ukraine. Mr. Osaulenko graduated from Kiev Institute of National Economy in 1974, specializing in organization of mechanical processing of economic information. Since 1980, Mr. Osaulenko has been working in statistics bodies as Head of division, Deputy Head of Computing Centre of Kiev City Statistical Department (1980-1987), First Deputy Head, Head of Kiev City Statistical Department (1987-1994), Deputy Minister, First Deputy Minister (1994-1996), Minister of Statistics of Ukraine (1996-1997) and since 1997 – Chairman of the State statistics Committee of Ukraine. Mr. Osaulenko has authored about 70 scientific publications, including 7 monographs; he is a Doctor of Science in Economics, Professor, and an Honoured Economist of Ukraine.

PETTERSSON, Birgitta works as chief legal adviser at Statistics Sweden since 2001. Before that, she worked for seven years as a legal adviser at the Ministry of Finance and she also has a background as junior judge in the Administrative Court of Appeal in Stockholm.

REZNEK, Arnold, PhD, works at the US Census Bureau. He is economist in the Center for Economic Studies, where he is the Administrator of the Headquarters' Research Data Center.

SOKOLIN, Vladimir L., Chairman of the Goskomstat of Russia since July 1998. In 1972-1992, while working at the Central Statistical Office of the USSR and later in Goskomstat of the USSR, he specialized in macroeconomic and finance statistics. In 1992-1993 he held leading posts at the Centre of Economic Analysis under the Government of the Russian Federation. In 1993 he was appointed Vice Chairman of the Goskomstat of Russia and was responsible for the development of the National Accounts System in Russia. Since 1991, he has participated in developing governmental programs on reforming the Russian economy. He is a member of the International Statistical Institute and Vice-Chairman of the Conference of European Statisticians.

SUNDGREN, Bo has worked for Statistics Sweden since 1968 as a researcher, project leader, and head of IS/IT. At present, he is senior advisor to the Director General of Statistics Sweden. He obtained a PhD degree in information processing from the University of Stockholm in 1973. In 1979 he was appointed professor by the Ministry of Culture and Education, and he has served as professor at the University of Uppsala, the University of Linköping, and at the Stockholm School of Economics. He has published a large number of articles and books on data/metadata management and statistical information systems, and he has done consultative work for several statistical agencies, universities, and international organizations.

THYGESEN, Lars is Director of User Services in Statistics Denmark. He has played a key role in establishing the Danish system of register-based statistics, and has been working with data protection and confidentiality for many years. He is responsible for servicing the research community.

TOCZYNSKI, Tadeusz, a graduate from the Central School for Planning and Statistics in Warsaw, is President of the Central Statistical Office of Poland since 1996. He began his professional career in the CSO in 1976 in the Industrial Statistics Division, and his previous assignments comprise various positions in the CSO such as Head of Section of the Price Statistics Division, Deputy Director of the Industrial Statistics Division, Director of Price Statistics Division, and Vice-president of the CSO. During his career he has also worked outside the CSO: from 1980-1982 in the UNECE, and from 1992-1996, he was president of the Board of the Joint Venture company.

TREWIN, Dennis, as Australian Statistician, has been head of the Australian Bureau of Statistics since July 2000. Prior to that he was Deputy Australian Statistician and, from 1992-1995, Deputy Government Statistician in New Zealand. He holds other senior appointments in Australia such as Electoral Commissioner and as a member of the committee which is preparing an independent report on the State of the Environment. He is an Adjunct Professor at Swinburne University. Internationally, he is currently President of the International Statistical Institute, having previously been Vice-President and President of the International Association of Survey Statisticians. He is Chairman of the Global Executive Board of the International Comparison Program and was recently elected as a member of the Bureau of the Conference of European Statisticians.

WALLMAN, Katherine K. currently serves as Chief Statistician at the United States Office of Management and Budget. In this capacity, she is responsible for overseeing and coordinating Federal statistical policies, standards, and programs; developing and advancing long-term improvements in Federal statistical activities; and representing the Federal government in international organizations such as the United Nations Statistical Commission. During her tenure in this position, Ms. Wallman has placed particular emphasis on increasing collaboration among the agencies of the Federal statistical system, strengthening the protection of confidential statistical information, fostering improvements in the scope and quality of Federal statistics, and making the products of the system more readily accessible to the public. Ms. Wallman, a Presidential Meritorious Executive, is an elected member of the International Statistical Institute, a Fellow of the American Statistical Association and the American Association for the Advancement of Science, and a Founder Member of the International Association for Official Statistics. In 1992, she served as President of the American Statistical Association. She is currently a Vice-Chairman of the United Nations Statistical Commission and recently was elected Chairman of the Conference of European Statisticians, United Nations Economic Commission for Europe.

WENDE, Tom is a sociologist. He works as empirical researcher in the Research Data Centre (RDC) of the German Federal Statistical Office. His main tasks include international microdata access-ways, anonymisation of complex datasets, and controlled remote data processing.

WILKIE MARTINEZ, Rochelle works at the US Census Bureau as Team Leader in the Policy Development Team in the Policy Office. She supports Census Bureau policy making, including developing the emerging data stewardship program

ZWICK, Markus is a National Economist. He is head of the Research Data Centre (RDC) of the German Federal Statistical Office. His main tasks include distribution analysis of tax- and social-transfer-systems, anonymisation of complex datasets and micro-analytical modelling for tax law variations.

Printed at Statistics Sweden, Stockholm
November 2003