Lessons of the inventory

Introduction

In 2006 Eurostat took the initiative of setting up Centres of Excellence (CENEX). The idea behind this scheme is to combine the strengths of the leading National Statistical Institutes (NSIs) in Europe on a certain topic. Often in several NSIs small isolated groups are working on specific topics. Other NSIs even lack the resources to pay enough attention to certain methodological issues. This situation led to the Eurostat initiative on Centres of Excellence. A CENEX could bring together the knowledge on a certain topic at a higher level by supporting the research in the leading countries and to spread this work to the other NSIs. Statistical Disclosure Control (SDC) was selected as a pilot topic. One of the tasks of the CENEX-SDC was to conduct an inventory or survey, trying to capture the situation with respect to SDC in Europe. For this a questionnaire was designed. The questionnaire was divided into 6 sections

I. General questions and questions asked about the legal aspects/regulations
II. Public use microdata files (PUF)
III. Microdata under contract for researchers (MUC)
IV. Tabular data (Magnitude tables)
V. Tabular data (Frequency tables)
VI. Remote access/Onsite facilities

This questionnaire was sent to all EU-member states as well as a few other European countries. The response was very encouraging. Eurostat was very helpful emphasising the importance of the questionnaire. In total 25 countries responded to this questionnaire. As well as most from EU-member states responses were received from Norway, Switzerland, Bulgaria and Turkey. This report summarises the key lessons learnt from this survey. It can be used to highlight the future research needs in SDC.

The questionnaire on Statistical Disclosure Control has been filled in by 25 countries. Most of these countries consider the legal protection of their data to be very important. It does not make a difference in importance if these data concern natural persons or enterprises. Most countries pay attention to the legislative and administrative aspects of confidentiality. Most countries pay also often or very often attention to the mathematical and computing aspects of confidentiality as well as to the organisational aspects. Most countries have a data protection law.

General questions and questions asked about the legal aspects/regulations

Most countries have principles and laws on public access to government information to the protection of statistical data. Most countries have specific regulations on statistical confidentiality. Internal regulations of the statistical office on statistical confidentiality exist only since recently or were changed over the last couple of years in most statistical offices.

Different countries have different definitions of confidential data. Some countries mention the aspect that data should be (directly) identifiable to consider them as confidential. In some countries the statistical law only refers to personal data. In some countries the statistical law does not mention the concept of confidential data at all, and sometimes different definitions are used depending on the context. In the statistical laws of some countries a reference can be found to the implementation of EU legislation.

Most countries have no statistics specific rules for the release of confidential data. In almost all countries almost all enterprise data are considered confidential. Most countries have no special rules that apply to the transmission of data to Eurostat. Most staff members of statistical institutes in European countries have to sign confidentiality warrants. Penalties can be imposed for (intentional) breaches of statistical confidentiality.

A majority of the statistical institutes in the inventory uses registers as e.g. the population register and the business register very often. Specific confidentiality rules concerning the use of these register data for statistical purposes are applied in some countries.

In most offices a certain number of staff members have been made responsible for ensuring statistical data confidentiality. In almost all countries universities (and research centres) have the option to use individual data concerning natural persons for research purposes. In the majority of the countries this option also exists for individual data concerning enterprises. Business organisations, fiscal authorities and marketing organisations can in general not get access to individual data. However, remarkable is that legal authorities as e.g. the police can get access to individual data in a considerable minority of the countries. Finally, for other governmental organisations the picture is somewhat mixed. In about half of the countries these organisations can get access to individual data.

In a majority of the countries a review panel (e.g. an ethical or statistical committee) exists to judge whether statistical data are sufficiently safe for use by persons outside the statistical office. Most of these review panels are internal committees. In a majority of the countries respondents can authorise the agency to provide their own individual data to a specified third party (informed consent).

In most countries the variables on racial or ethnic origin, political opinions and religious or philosophical beliefs were considered as sensitive. Also data concerning health and sex life were considered sensitive in most countries. In addition in a majority of the countries trade union membership, data relating to offences, criminal convictions and security measures and data related to incomes were seen as sensitive. Data about professions and educational data were considered as sensitive in a minority of the countries only.

Special licensing agreements exist in a minority of the countries. Access under contract for named researchers exists in a majority of the countries. About half of the countries have the option of access only for specially sworn employees. Also about half of the countries screen the results with respect to disclosure control but only a minority screens the users.

Public Use Files (PUFs) and Microdata Under Contract (MUCs)

Most countries release microdata concerning natural persons and enterprises. However, Public Use Files (PUFs) exist only in a minority of the countries. The majority of the countries release Microdata Under Contract (MUCs), although not on administrative data. Synthetic data files are hardly produced in Europe.

Many offices have organisational, methodological and software problems concerning statistical confidentiality development. Most countries do not receive technical assistance from other countries to help with the implementation of disclosure control. However, many countries would like to receive help to solve in particular their software problems.

Different data protection methods exist. Which methods are used depend on the statistics to protect. Most countries do not include direct identifiers in their PUFs and MUCs. Global recodes and local suppressions are key methods. Some countries use in addition top and bottom coding. Other methods (e.g. micro aggregation, noise addition, data swapping, imputation and synthetic data) are hardly applied in practice.

Risk assessment methods are only applied in a few countries, but many other countries are interested in applying these methods as well. With some teaching and support it is likely that more countries will make use of risk assessment methods.

A few countries make use of the software package μ-ARGUS. Many more countries are interested in making use of this package to run the data protection procedure more efficiently. Nowadays many countries use their own microdata protection programs. This is very costly and sometimes difficult to maintain. It is clear that countries that do not yet use μ-ARGUS are interested in help to introduce this software package.

Tabular data

Tabular data protection is a challenging task for statistical institutes. They are supposed to produce tables that can be released safely. This goal must be balanced with the information loss caused by table protection, but also with the resources spent on the task of tabular data protection.

With only very few exceptions, all countries report use of minimum cell count rules to assess sensitivity on the individual cell level for magnitude tables presenting sensitive information on an aggregate level. Additionally, for data concerning enterprises, concentration rules are reported to be used by most of the respondents. As pointed out in the CENEX-SDC handbook, prior posterior rules perform better than dominance rules. Some countries have meanwhile started to replace the still much more common dominance rules.

Different rules are thus used to determine the unsafe cells. Depending on the country and the statistics a method is applied with which secondary suppressions are added to the tables to make sure that the primary suppressions cannot be recalculated. For foreign trade statistics most countries use passive confidentiality (i.e. only protect the records of the enterprises that have asked for protection) to make sure that the very detailed tables on this topic can be published.

For secondary suppressions some countries make use of the τ-ARGUS software package. Some countries use their own programs or make use of manual protection. It is clear that this last option is very time consuming and often leads to severe over or under protection. A few countries tested τ-ARGUS but did not start using it. The main reason for those countries was that it does not fit in their production environment.

In order to harmonize and improve the SDC practices in the European statistical systems, on the basis of the findings of this report, in particular the observed disagreement between the actual, and the recommended practices, both regarding the use of methodologies, and software we suggest the following activities:

- With respect to primary confidentiality, it could be a possible way to promote the methodologically superior prior posterior rule by adapting the definitions used in the context of the structural business surveys and PRODCOM for safety of European aggregates.
- With respect to promoting the introduction of τ-ARGUS for secondary cell suppression, we propose the following for a future work agenda.
    - Extension of the package with respect to the processing of linked tables which will make it much easier to use it in practice.
    - A co-operation project with some interested NSIs with the aim of defining, testing and implementing suitable procedures to use τ-ARGUS to protect all the tables produced by those NSIs on the basis of a particular business survey or census selected for the project.
    - Consider migration into open-source software. Such a migration might help to deal with some of the obstacles found by some NSIs concerning the eventual integration of the package into their production system.
    - Further research and development of perturbative methods for tabular data protection, implementation into practical tools (such as τ-ARGUS), and pilot studies on the basis of real data.

For frequency tables most countries apply a frequency rule to identify sensitive cells. Not only cell suppression is used, also global recodes and rounding are popular methods to protect frequency tables.

The CENEX-SDC handbook recommends that frequency rules should be used to identify sensitive cells for all frequency tables. More work is required to understand why some countries are using other rules (that have been developed for magnitude tables). The CENEX-SDC handbook also recommends that recoding and rounding should be implemented to protect frequency tables based on whole population data sources, e.g. Census and administrative data, and outlines the disadvantages of using cell suppression.

Remote access/Onsite facilities

About half of the European countries provide access to microdata for researchers in a secure setting at their office. The NSIs that provide this onsite facility check manually the results that the researchers want to bring out of the NSI. This is done either randomly or on a complete basis. A couple of NSIs conduct a training programme for the researcher in order to prevent them from producing disclosive statistical output.

A few NSIs provide a remote execution service or remote access (via the Internet) of microdata to researchers. Examples of how the NSIs have set up a secure connection are by use of VPN authorisation, Citrix server and biometrics. Manual checking of the results that are made available to the researchers is done by most NSIs.

There is a difference in what services the NSIs provide for letting the researcher get access to microdata. The NSIs have indicated that the work of checking the statistical output is made manually. Probably, the number of requests will increase in the future. Therefore, it is of interest for many NSIs to study further how to limit the checking burden.