

Report on the CENEX-SDC inventory

Sarah Giessing
Andrea Harausz
Anco Hundepool
Helen Lindkvist
Jane Longhurst
Eric Schulte Nordholt
Andreja Smukavec

December 2006

1 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 Statistic 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 stressing emphasising the importance of the questionnaire. In total 25 countries responded to this questionnaire.

As well as most EU-member states responses were received from Norway, Switzerland, Bulgaria and Turkey.

This report summarises the results of the survey.

2 Results of the legal section of the inventory

In this section the results of the legal section of the inventory are discussed. Most of these countries responding to the questionnaire 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 often or very often pay attention to the mathematical and computing aspects of confidentiality as well as to the organisational aspects. Most countries reported that they have data protection legislation.

France introduced a *data protection law* (to the protection of statistical data) in 1978, the other countries participating in the inventory introduced or changed such a law in the last fifteen years. Most countries reported that they have principles and laws on public access to government information to the protection of statistical data. Most countries have specific regulations on statistical confidentiality. In most statistical offices internal regulations on statistical confidentiality have only recently been introduced or have changed over the last couple of years.

Different countries have different *definitions* of confidential data. Some countries only consider data as confidential if it is data that is (directly) identifiable. 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 specific statistical *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.

Only two statistical agencies responded that they conduct *assessments* of public attitudes, perceptions and reactions to confidentiality very often. The majority of the statistical institutes in this inventory use registers very often e.g. the population register and the business register. 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, remarkably legal authorities, 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 sexual activity 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 income were seen as sensitive. Data about professions and educational data were considered as sensitive in a minority of the countries.

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. About half of the countries screen the results of analysis with respect to disclosure control but only a minority screen the users. For statistics on persons and administrative data the option of access in a controlled environment does not exist for the majority of the statistical institutes, but for statistics on enterprises there is a small majority where this option exists. The option of trusted third parties that keep the keys necessary for identification and linking files hardly exists in Europe.

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 releases Microdata Under Contract (MUCs) also known as research use microdata files (MFR) , although not on administrative data. Synthetic datafiles are hardly produced in Europe and the majority of countries also have no on-site facility or online access option.

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 their particular software problems.

3 Microdata under contract

3.1 Introduction

The following analysis is concerned with microdata under contract. Microdata under contract are only available to selected researchers. The questionnaire contains details about disclosure protection and risk assessment methods for microdata under contract and the use of checklists and data protection software.

3.2 Evaluation

Question 23: Disclosure limitation methods

Initially the questionnaire focused on disclosure protection methods for microdata released under contract. Overall the results show that recoding data into broader categories is frequently used by many countries (see table 3 1). However, this does depend on the type of data. While over 60% of the countries questioned use recoding to protect data concerning natural persons and for business statistics, only about a quarter use this method for administrative data. Another method frequently applied is the use of geographical or population thresholds, which is more often used for data concerning natural persons than for data concerning enterprises. For data concerning natural person 15 and 16 countries indicated that they use this disclosure protection method for population census and other data and labour force data, respectively; for data concerning enterprises it is 9 (business statistics) and 4 (administrative data). The deletion of data items (local suppression) as well as top and bottom coding are applied less often but are still used in several countries.

When distinguishing by type of statistics, the following picture is observed: For data concerning enterprises, one third of the countries use top and bottom coding in business statistics and 8% for administrative data; for data concerning natural persons, it is between 36 and 52%. The use of local suppression (the deletion of data items) does not show significant differences between different data types apart from administrative data where that method is used less frequently (5 countries). Sampling and deletion of especially sensitive records rank in the middle area of use frequency. Where the method of sampling is used, this is mostly implemented for population census data. 40% of the countries questioned indicated that they use sampling to protect population census data. The deletion of especially sensitive records is applied most often (about one third) for business surveys as opposed to other data sources. The method of micro aggregation is applied rather seldom when compared with the above-mentioned methods. In business statistics, every fourth country uses that form of disclosure protection method, for other data concerning natural persons it is one fifth.

The other methods for microdata under contract (noise addition, data swapping, (multiple) imputation, generating synthetic data (e.g. by resampling) and post randomisation) are rarely used by the countries questioned. It should be mentioned that imputation and post randomisation

are used only in social statistics – if at all – and noise addition as well as generating synthetic data are used more often for data concerning enterprises.

Other additional methods reported by some countries are removing variables that provide direct identification (name, address, personal identification or registry code of the data subject), modification of data in various ways (e.g. forming ratios of the variables) and the qualitative estimation of representativeness.

Table 3 1: The use of disclosure methods for microdata under contract

Disclosure methods	...data concerning natural persons							...data concerning enterprises				
	Population census		Labour force survey		Other data		Business statistics		Administrative data			
	N	%	N	%	N	%	N	%	N	%		
Recoding data into broader categories	Yes	15	60	17	68	16	64	Yes	16	64	6	24
	No ¹	10	40	8	32	9	36	No	9	36	19	76
Geographical or population threshold	Yes	15	60	16	64	15	60	Yes	9	36	4	16
	No	10	40	9	36	10	40	No	16	64	21	84
Deletion of data items (local suppression)	Yes	8	32	11	44	10	40	Yes	10	40	5	20
	No	17	68	14	56	15	60	No	15	60	20	80
Top and bottom coding	Yes	9	36	11	44	13	52	Yes	8	32	2	8
	No	16	64	14	56	12	48	No	17	68	23	92
Sampling	Yes	10	40	5	20	6	24	Yes	5	20	3	12
	No	15	60	20	80	19	76	No	20	80	22	88
Deletion of especially sensitive records	Yes	5	20	4	16	7	28	Yes	8	32	4	16
	No	20	80	21	84	18	72	No	17	68	21	84
Micro aggregation	Yes	3	12	3	12	5	20	Yes	6	24	1	4
	No	22	88	22	88	20	80	No	19	76	24	96
Noise addition	Yes	1	4	0	0	1	4	Yes	2	8	1	4
	No	24	96	25	100	24	96	No	23	92	24	96
Data swapping	Yes	0	0	1	4	2	8	Yes	1	4	0	0
	No	25	100	24	96	23	92	No	24	96	25	100
(Multiple) imputation	Yes	1	4	1	4	1	4	Yes	0	0	0	0
	No	24	96	24	96	24	96	No	25	100	25	100
Generating synthetic data (e.g. by resampling)	Yes	0	0	0	0	1	4	Yes	1	4	1	4
	No	25	100	25	100	24	96	No	24	96	24	96
Post randomisation	Yes	2	8	0	0	1	4	Yes	0	0	0	0
	No	23	92	25	100	24	96	No	25	100	25	100

1) aggregation of the response options 'No' and 'No answer' for all disclosure methods and statistics

Question 24: Risk assessment.

This question covered risk assessment methods for microdata under contract and the satisfaction with them. Two methods are distinguished here: The μ -ARGUS threshold rule and the Franconi-Benedetti approach. Altogether, 16% (4 countries) have so far used the μ -ARGUS threshold rule

and 4% (1 country) the Franconi-Benedetti approach. Those users indicated medium to high satisfaction.

Question25: Use of checklists

This section of the questionnaire analyses available checklists which are used in the various institutions. 6 countries indicated that they use checklists. The content of such lists differs considerably:

- Switzerland uses a checklist with general rules on the detail of the released variables for sensitive data only.
- Germany’s checklist contains details of the various anonymisation methods. (Sub-sample 70%, threshold 500,000 for variable ‘nationality’ and 5,000 for all other key variables, top/bottom coding for age and income, region: federal states (=Länder) level)
- According to legislation Estonia can release identifiable data for scientific purposes. Usually they remove direct identification variables and recode variables that may be used for indirect identification. Lists of variables and level of details are agreed for personal surveys: LFS, EU-SILC and HBS.
- Greece has a checklist which contains two different procedures. In case sampling micro data samples are released, the sampling weight factor for each record should be higher than 3. Global recoding of variables or local suppression of values is carried out for the creation of records with factors higher than 3. To avoid micro data sets relating to the whole population being released, the first check is the estimation of population frequencies of a combination of values of identifying variables. The population frequencies of a combination of values of identifying variables should be higher than 3.
- For researchers, there are no general rules of detail arising from confidentiality considerations in Norway. Data for researchers must be de-identified. The researcher must have an approved project and the employer must be on a list of organisations approved for receiving micro data from Statistics Norway. Rules for approval have been set by law and by Statistics Norway according to law. The unit in Statistics Norway owning a dataset has the basic responsibility for complying with these rules, but can seek advice with Statistics Norway’s Confidentiality Committee.
- The United Kingdom implements a checklist asking for details of the key microdata variables, visible and traceable variables, geographical details and sample design. In addition the DIS-SUDA software was used to highlight high risk records in the micro data samples from Census (SARs). Work is currently being undertaken to develop probabilistic modelling methods for micro data risk assessment based on work by Skinner¹ et al. (1998)

¹

Question 26: Software.

This question covered the general use of available software for microdata protection. Although 20% (5 countries) use μ -ARGUS, the majority of countries do not (yet) use that software. Users indicated their satisfaction with this software as rather high, while 5 countries indicate medium to high satisfaction, and one country is not satisfied at all.

Own procedures, however, are applied twice as often. Among those 10 countries, 3 provided information about their satisfaction with their own software, which is on a medium level.

9 countries provided details on their own procedures that they use for microdata under contract. Because of the different procedures for data processing of the various statistical surveys, not all countries could indicate their own procedures.

- The Czech Republic sometimes use their own accessory procedures in PL SQL e.g. for preparing files with respect to threshold rules with a minimum frequency of 3.
- Germany uses their own procedures regarding micro aggregation based on the implementation available in μ -ARGUS and noise addition.
- The program which is used in Greece estimates the population frequencies of a combination of identifying variables, values and examines in which combinations the frequencies are lower than 3. If the frequency is lower than 3, recoding and/or local suppression is carried out.
- In Estonia the use of their own procedures depends on expert decisions.
- Finland uses own procedures e.g. by generating synthetic datasets and by various in-house software.
- in Poland the data sets do not include information that provides direct identification of the units (surname, address, PESEL)
- Slovakia creates data protection software for every individual survey.
- Spain uses SAS macros among others.

3.3 Summary

To sum up, many of the disclosure protection methods are already used for microdata under contract. However, use depends on the relevant type of statistics and, consequently, cannot be examined in general. What can also be seen is that risk assessment methods are used less frequently by the countries questioned. Only a few countries reported the use of checklists for risk assessment. More frequent is the use of software for data protection. Many countries use their own procedures which differ considerably between countries.

4 Public use files

The previous section provided details on practices concerning microdata under contract. Here we describe the results for public use files.

Three countries apply the μ -ARGUS threshold rule and two countries apply the Benedetti-Franconi approach as the risk assessment method for the protection of public use files. One country is very satisfied with the μ -ARGUS threshold rule and one country is very satisfied with the Benedetti-Franconi approach. Checklists are used in Germany, Greece and the United Kingdom. Only four countries use μ -ARGUS for data protection, two countries are very satisfied with the software. Seven countries use their own procedures for the protection of public use files (Austria – under development, the Czech Republic, Germany, Greece, Slovak Republic, Spain and the United Kingdom).

Thirteen countries didn't answer what kind of disclosure limitation methods they use for the protection of public use files; therefore, in the continuation of this section twelve countries represent the total number of answers.

1. Population Census

More than a half of countries apply geographical or population thresholds, sampling, top and bottom coding and recoding into broader categories for the protection of public use files, a half of countries apply local suppression and less than a half of countries apply the deletion of especially sensitive records. Only one country applies (multiple) imputation, micro aggregation and post randomization for protection, while data swapping, generating of synthetic data and noise addition are not used for the protection of public use files in any country.

2. Labour Force Survey (LFS) and other data concerning natural persons.

Geographical or population thresholds, top and bottom coding, recoding into broader categories are used for the protection of public use files by more than a half of countries. Less than a half of countries apply local suppression, deletion of especially sensitive records and sampling, and only one country applies (multiple) imputation and micro aggregation. No country applies data swapping, generating of synthetic data, post randomization and noise addition for the protection of public use files.

3. Business data

Exactly a half of the countries (6 countries) apply recoding into broader categories, while less than a half of the countries apply geographical or population thresholds, sampling, top and bottom coding, local suppression, deletion of especially sensitive records and micro aggregation. Noise addition is used only by one country, while no country applies data swapping, (multiple) imputation, generating of synthetic data and post randomization for the protection of public use files.

4. Administrative data (concerning enterprises)

Geographical or population thresholds, sampling, recoding into broader categories, local suppression, deletion of especially sensitive records and micro aggregation are used only by one country; post randomization, noise addition, generating of synthetic data, (multiple) imputation and data swapping are not used for the protection of public use files by any country.

No country applies data swapping for protection. Micro aggregation, (multiple) imputation, post randomization and noise addition are used only by one country. Micro aggregation and (multiple) imputation are used only for the protection of micro data concerning the natural persons, noise addition only for the protection of non-administrative business data and post randomization only for the protection of the population census.

Countries didn't specify any additional methods for the protection of public use files.

5 Magnitude tables

5.1 Introduction

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 on one hand, but also with the resources spent on the task of tabular data protection. Our questionnaire involved four questions (questions 28 to 31) on magnitude tables. The first two questions address methodological issues, whereas the latter two are concerned with the technical implementation of secondary cell suppression methods.

In section 5.2 we will, question by question, present, analyse and comment on the results of the survey, summarizing the main results in section 5.3 and deriving conclusions for the future work agenda in section 5.4.

5.2 Evaluation:

Question 28: Methods for disclosure limitation.

This question asked information on the methods used to protect tabulations of data from various sources: from the population census, the labour force-, or other surveys concerning natural persons, and on the other hand for business data from business surveys and administrative sources. Some of the respondents ticked none of the boxes with respect to a particular type of statistics. This could be item non-response in the classical sense, maybe because the respondent is not familiar with this kind of data, and the SDC-practice of the institute regarding these data. Another possible reason can be that our questions do not apply to this kind of data for that country, i.e. it does not conduct a population census, or does not use administrative data for tabulations for publication. Or, on the other hand, perhaps those data (e.g. administrative data) are public in that country, and do not require disclosure control, or it may be that the institute publishes only highly aggregated data from a particular source, where it can be assumed that the re-identification risk is sufficiently small. Note, that other respondents may have mentioned the same kind of practice explicitly, considering it as ‘recoding’, and/or ‘limiting the number of table dimensions’.

It should also be noted that in some cases respondents may have ticked boxes regarding data concerning natural persons without considering the fact that question 28 relates to magnitude tables only, giving answers that would be adequate for frequency tables which are more common for this kind of data.

Table 5.1 below summarizes the individual reactions. The table does not contain a line referring to “Adding noise / blurring of cell totals”, because no country reported use of any such method. Neither does the table contain a line referring to the use of minimum cell count rules, because, considering the fact that all concentration rules cover a minimum cell count rule, we observed that all countries reporting some kind of assessment of cell sensitivity use at least a minimum cell count rule.

On the other hand, depending on the data source type, between 17 (population census) and 25 (business surveys) countries state that they assess cell sensitivity on the individual cell level. Of those, between 32% (labour force) and 72% (enterprise data) use a concentration rule for that purpose. Concentration rules should be used, if it is likely that respondents to individual cells can be identified, especially if the size of their contribution is outstanding, and if it is assumed that statistical confidentiality is violated, when a published aggregate provides a close estimate of an individual contribution. In the class of concentration rules, the CENEX-SDC handbook/guidelines recommend use of the prior/posterior rule (also known as p%-rule). Considering that this method was widely unknown in Europe a few years ago, it is certainly a success that now 6 countries state that they use this method. However, dominance rules are still much more frequently in place – between 21% (labour force) and 64% (business surveys) of the cases.

Apart for tables presenting ratios, statistical tables are additive, because they usually involve totals and subtotals for groups of table cells. In order to avoid the problem of disclosure by differencing, if it has been assessed that there are sensitive cells which must not be published, some kind of table level protection measure is required.

A very effective and simple method is to reduce the amount of detail in the tables. Between 72% and 83% of the countries reporting use of any disclosure methods for the type of data specified in the questionnaire say that they protect the data by reducing the amount of detail in the tables, through recoding or by limiting the number of table dimensions. However, after the amount of detail has been reduced sufficiently, so that the number of sensitive cells is not a too big proportion of the number of non-zero cells, in order to avoid that too much of the information collected remains unpublished, a more focussed protection method should be used. One possible option is to ask those respondents to sensitive cells, whose data would be at risk in case of a publication, for their permission. Application of this method is explicitly mentioned by one country. Another well known method is secondary cell suppression. It is reported to be used by 10 countries on tabulations of data concerning natural persons, and by 19 countries on tabulations concerning enterprises². An alternative to secondary cell suppression are perturbative approaches, like “Adding noise / blurring of cell totals” which no country reported to use, or rounding which is mentioned to be used by 5 countries. In that context it should be noted that ordinary implementations of rounding methods often fail to protect tables sufficiently. This is especially a problem for magnitude tables computed from strongly skewed data typical for business statistics, where it is usually difficult to find a suitable rounding base. See for instance (Salazar, 2005) for a discussion of rounding methods and related concepts.

Under the label ‘other methods’ 3 institutes mention that survey estimates are not published, when their estimated precision is too low.

² 3 countries did not tick the ‘secondary suppression’ box for any of the statistics listed in the questionnaire, but nevertheless state that they assign secondary suppressions manually in question 30.

Table 5.1: Use of disclosure limitation methods by type of data source

Methods reported	data concerning natural persons						data concerning enterprises			
	Population Census		Labour Force Survey		Other data		Business Surveys		Administrative data	
<i>Any kind of disclosure control reported</i>										
No. of countries	18		21		20		25		18	
<i>Assessment of cell sensitivity reported</i>										
No. of countries	17		19		19		25		18	
	abs.	%	abs.	%	abs.	%	abs.	%	abs.	%
Any concentration rule	6	35	6	32	10	53	18	72	13	72
<i>Dominance rule</i>	5	29	4	21	9	47	16	64	11	61
<i>Prior/posterior</i>	3	18	5	26	4	21	6	24	3	17
<i>Disclosure control methods on the table level reported</i>										
No. of countries	18		21		20		25		18	
	abs.	%	abs.	%	abs.	%	abs.	%	abs.	%
<i>Secondary cell Suppression</i>	8	44	10	48	9	45	19	76	13	72
<i>Rounding</i>	2	11	4	19	3	15	3	12	2	11
<i>Recoding variables to reduce detail</i>	13	72	16	76	16	80	18	72	15	83

Question 29: Foreign trade statistics.

14 countries state that they use a special rule (e.g. passive confidentiality) to protect data from foreign trade statistics.

Question 30: Secondary suppression methods.

This question addresses the problem of how the selection of secondary cell suppression is solved technically. Table 5.2 summarizes the results. Evidently, manual procedures are still in place in most (e.g. 19) countries. However, 7 NSI's use the τ -ARGUS technology, choosing between the Modular, linear programming based and the Hypercube method, both offered by the program. We also observed that only 2 countries state that they do not use secondary suppression at all. 7 NSI's have developed their own technical solutions. In one of these cases the CIF interface to the hypercube method of τ -ARGUS which was earlier developed on behalf of Eurostat is reported to be used. One institute says they have developed an implementation of a hypercube method on their own.

Table 5.2: Technical solutions to assign secondary suppressions

<i>Method to assign secondary suppressions</i>	<i>No of countries</i>
<i>Manual procedures</i>	19
<i>Own software solution</i>	7
<i>τ-ARGUS</i>	7
<i>τ-ARGUS Modular</i>	6
<i>τ-ARGUS Hypercube</i>	5

Question 32: Use of τ -ARGUS.

The CENEX-SDC handbook/guidelines recommend use of the τ -ARGUS software to assign secondary suppressions. The purpose of question 32 was to identify the main reasons why some NSI’s do not use this program. In fact, we found that the majority (18 institutes) does not (or not yet) work with that software. Of those 18 institutes, 7 have not even tested the software (2 of those say they plan to test it). Of the 11 institutes who did test the software, none claimed that they do not use it because they were not happy with the results (‘too much suppression’), or because the software was not sufficiently stable. 3 say they actually plan to introduce the software. 7 NSI’s say that τ -ARGUS does not fit into their production environment, and 2 say it is difficult to use. In 3 cases NSI’s admit that they ‘do not know’ about the software, or do not have enough resources (in terms of educated staff) to attend to problems related to the introduction of the software.

5.3 Summary

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 more than 70% of the respondents. As pointed out in the CENEX-SDC handbook/guidelines, prior posterior rules perform better than dominance rules. Six countries have meanwhile started to replace the still much more common dominance rules.

It is widely accepted, obviously, that the amount of detail given by tabulations must be limited with respect to disclosure control. A majority of more than 70% of the respondents explicitly mention this as a technique for disclosure control. As a protection technique, compared to the reduction of detail in the publication tables, secondary cell suppression generally allows for more information to be published. On the other hand it requires much more effort. For data concerning enterprises, more than 70% of the respondents say that the institute protects tables by cell suppression. While recent research in the field of tabular data protection concentrates on the development cell perturbation methods, pointing out that certain methods indeed provide adequate protection to the data, and have the potential to outperform cell suppression regarding the loss of information due to

disclosure limitation by far, in practice those methods are not yet used in any of the NSI's.

With respect to the technical implementation of secondary cell suppression, the majority of statistical institutes use manual procedures. Although manual processing is time consuming, it can be carried out by staff not having to be especially skilled or educated. However, such manual procedures bear a high risk that some cells will not be sufficiently protected.

Some institutes say they have developed technical solutions for secondary cell suppression on their own. It is of course a matter of quality of these technical solutions, to what extent they are able to provide adequate solutions to the cell suppression problem both, in terms of protection, and also in terms of information loss.

The software τ -ARGUS offers methodologically adequate solutions for secondary cell suppression. However, the integration of the computation of those solutions into the production process of a particular survey requires expertise regarding tabular data protection, and also modifications in the IT applications plus eventually newly defined special IT applications, if processing of linked tables is required which is not part of the current version of τ -ARGUS. Consequently, 7 of the respondents say that τ -ARGUS does not fit into their production environment. In spite of such obstacles, 7 institutes managed to introduce τ -ARGUS into data production at least for some statistics.

5.4 Conclusions for future work

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 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 NSI's with the aim of defining, testing and implementing suitable procedures to use τ -ARGUS to protect all the tables produced by those NSI's 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 NSI's 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.

References

Salazar, J.J. 2005, "Protecting tables with Cell Perturbation", Work session on Statistical Data Confidentiality, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 291-298

6 Frequency tables

This section of the inventory asked NSIs which rules they apply to identify sensitive cells in frequency tables and which methods they implement in order to protect these sensitive cells. These questions were asked for different types of frequency tables; those generated from Census, Survey and Administrative data sources.

6.1 Rules applied to identify sensitive cells in frequency tables

For frequency tables generated from Census data the most popular rule for identifying sensitive cells was the frequency rule, 64% of NSIs adopt this rule. 24% of NSIs reported that they used the dominance rule for Census data and only 8% use the prior/posterior rule. Note, some NSIs may adopt more than one rule for protecting their Census tables, whereas others may use alternative methods to those proposed in the inventory.

For frequency tables created from survey data again the most popular rule for identifying sensitive cells was the frequency rule with 68% of NSIs using this rule. 52% of NSIs reported using the dominance rule for survey data and 16% adopt the prior/posterior rule.

For frequency tables of administrative data that pattern is the same, the most popular rule for identifying sensitive cells is the frequency rule, then the dominance rule then the prior/posterior rule. 56% of NSIs use the frequency rule, 32% use the dominance rule and 4% use the prior/posterior rule.

6.2 Measures for the protection of frequency tables

For Census frequency tables 76% of NSIs use recoding/collapsing to protect sensitive cells. 64% use cell suppression, 28% round the data and 8% adopt a noise addition or distortion methodology.

For frequency tables produced from survey data cell suppression and recoding/collapsing are the most popular SDC methods, 80% and 76% of NSIs implement these approaches, respectively. 32% of NSIs use rounding and 4% adopt a noise addition or distortion methodology.

For frequency tables generated from administrative sources the pattern is the same 56% of NSIs use recoding/collapsing, 60% use suppression, 20% round the data and 4% adopt a noise addition or distortion methodology.

6.3 Conclusions for future work

In general the most popular rule for identifying sensitive cells for frequency tables (generated from Census, surveys or administrative sources) is the frequency rule. The most popular methods for protecting these sensitive cells are suppression and recoding/collapsing.

CENEX SDC

a Centre of Excellence for Statistical Disclosure Control

The CENX SDC handbook/guidelines recommend 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 dominance or prior/posterior rules, that have been developed for magnitude tables. The CENEX handbook also recommends that recoding/collapsing 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. These results show that more work needs to be undertaken to encourage the use of rounding and understand why NSI's prefer suppression.

7 Onsite Data laboratory/Remote Execution/Remote Access

This section of the inventory asked NSI's about their facilities for microdata access in onsite data laboratories or using remote access/execution.

7.1 Data laboratory

Out of the 25 NSIs, there are just more than half of them (13) that provide access to micro data to researchers in a secure setting at their office. Most of those offices that provide a data laboratory (8) have between 2 and 10 users per year. Only 2 NSIs have more than 25 users per year. The NSIs that provide this 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 outputs.

7.2 Remote Access/Execution

6 NSIs provide a remote execution service or remote access (via the Internet) of microdata to researchers. Out of these, 3 NSIs have on average less than 40 requests per year, whereas one has more than 150 requests. 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 used by most NSIs.

7.3 Conclusions

There is a difference in what services the NSIs provide for letting the researcher get access to microdata. About half of them provide a secure setting at their office and 6 of them provide a remote access service. Also, for almost all NSIs that provide these sorts of services the average number of requests per year is rather small (less than 40). The NSIs indicate that the work of checking the statistical outputs is made manually. Probably, the number of requests will increase in the future. Therefore, it might be of interest for many NSIs to look further into how to avoid manual checking.

§ Appendix A: The questionnaire

In 2005 Eurostat has launched the idea of establishing European Centres of Excellence in the field of Statistics as a way to reinforce the cooperation between the National Statistical Institutes. In this way the various institutes in Europe could benefit from each others experiences and together raise the level of their statistical production process.

As a first project/centre the topic of “Statistical Disclosure Control” has been chosen. In this field there was already a tradition of European cooperation thanks to the CASC (FP5) project. This cooperation will be continued in this CENEX.

This CENEX project will be a one year project and be active in 2006. An overview of the activities in this CENEX can be found at the CENEX-Website (<http://neon.vb.cbs.nl/cenex/default.htm>).

One of the activities of the CENEX is to conduct a survey on SDC activities in the EU-member states. This questionnaire can be seen as a follow-up/extension of a survey implemented 5 years ago by the UN-ECE in Geneva. The UN-ECE only aimed at the so-called new member states.

You will find the questionnaire attached. We kindly ask you to fill it in and return it to the CENEX-SDC project leader (Anco Hundepool), if possible before: 25th May 2006. The questionnaire is divided into six sections:

- I. General and legal issues.
- II. Micro data (Public Use Files)
- III. Micro data (Micro data Under Contract)
- IV. Tabular data (magnitude tables)
- V. Tabular data (frequency tables)
- VI. Remote access/ Onsite facilities

After the processing we will publish a report of the results on the CENEX website. We very much appreciate your cooperation.

Anco Hundepool,
CENEX project leader
Statistics Netherlands
P.O. Box 4000
2270 JM Voorburg
Netherlands
Email: ahnl@rnd.vb.cbs.nl
Phone: +31-70-3375038
Fax: +31-70-3375990

I. General questions

1. How important is the protection of statistical data to your agency?
(Please put an 'x' in the relevant cell)

Data concerning...	<i>not important</i>	<i>slightly important</i>	<i>important</i>	<i>very important</i>
natural persons				
enterprises				

2. To what extent does your agency pay attention to each of the following aspects of confidentiality?
(Please put an 'x' in the relevant cell)

	<i>very often</i>	<i>often</i>	<i>sometimes</i>	<i>never</i>
Legislative and administrative aspects				
Mathematical and computing aspects				
Organisational aspects				

3. Legal issues

3.1. Does your country have any of the following types of legal regulation applicable to the protection of statistical data?

Relevant legal regulations	yes	Date
Data Protection Law		
Principles and laws on public access to government information		
Specific regulations on statistical confidentiality		
Internal regulations of the statistical office		
Other (please specify):		

If yes, please indicate the name of the legal regulation and the year when these regulations came into force.

.....
.....
.....
.....
.....

Please attach if possible the relevant articles of the relevant legal regulations to this questionnaire.

3.2. Is there a definition in national law of confidential data? What is it?

.....
.....
.....
.....
.....

3.3. Are there any special articles related to the implementation of EU legislation in the statistical confidentiality domain? If yes, please specify

.....
.....
.....
.....

3.4. Are there any specific rules for the release of confidential data in specific sectors (SBS (Structural Business Survey) national accounts,...)? If yes, please specify.

.....
.....
.....

3.5. What kind of data concerning enterprises is seen by your agency as confidential?

.....
.....
.....

3.6. Are there special rules that apply to the transmission of data to Eurostat? If yes, please specify.

.....
.....
.....

3.7. Do the statistical staff of your institution sign legal confidentiality commitments on appointment? Are there penalties imposed for intentional breaches of statistical confidentiality?

.....
.....
.....

4. Does your agency conduct any assessment of public attitudes, perceptions and reactions to confidentiality?
(Please put an 'x' in the relevant cell)

<i>very often</i>	<i>often</i>	<i>sometimes</i>	<i>never</i>

5. To what extent do you use administrative registers (e.g., population register, business register)?
(Please put an 'x' in the relevant cell)

<i>very often</i>	<i>often</i>	<i>sometimes</i>	<i>never</i>

6. Are there specific confidentiality rules concerning the use of data from administrative registers (e.g., population register, business register)?

.....
.....
.....
.....
.....

7. Are there special staff in your office responsible for ensuring statistical data confidentiality?
(Please describe number and role)

.....
.....
.....
.....

8. Is access to individual micro data possible for specific users outside the agency?
(Please put an 'x' in the relevant cell)

Access to individual micro data concerning ... granted to ...	<i>natural persons</i>	<i>enterprises</i>
Universities		
Research centres		
Business organisations		
Legal authorities, e.g. police		
Fiscal authorities		
Other governmental authorities		
Marketing institutions		
Other (please specify)		

9. Review panel

- a. Does your agency use a review panel (e.g. an ethical or statistical committee) to judge whether statistical data are sufficiently safe for use by persons outside the agency?

yes	no

- b. Is this panel internal or external?

internal	external

10. Can a respondent authorise your agency to provide original data about himself to a specified third party (informed consent)?
(Please put an 'x' in the relevant cell)

A respondent can authorise	<i>Data concerning natural persons</i>		<i>Data concerning enterprises</i>	
	<i>yes</i>	<i>no</i>	<i>yes</i>	<i>no</i>
the agency to divulge statistical data				

11. What kind of data concerning natural persons is seen by your agency as sensitive?
(Please put an 'x' in the relevant cell)

Sensitive data concerning natural persons	yes
Racial or ethnic origin	
Political opinions	
Religious or philosophical beliefs	
Trade union membership	
Data concerning health and sex life	
Data relating to offences, criminal convictions and security measures	
Data related to incomes	
Data about professions	
Educational data	
Other data (please specify)	

12. Do you use specific organisational or administrative measures to prevent disclosure of confidential data?

Organisational and/or administrative measures	<i>.. concerning natural persons</i>			<i>.. concerning enterprises</i>	
	<i>Population Census</i>	<i>Labour Force Survey</i>	<i>Other data</i>	<i>Business Statistics</i>	<i>Administrative data</i>
Special licensing agreements					
Access under contract for named researchers					
Access only for specially sworn employees					

C E N E X S D C

a **CEN**tre of **EX**cellence for **Statistical Disclosure Control**

Screening the results with respect to disclosure control					
Screening of the users					
Access only in controlled environment					
Trusted Third Parties that keep the keys necessary for identification					
Other (please specify):					

13. Do you release microdata?
(Please put an 'x' in the relevant cell)

Release of microdata ..	<i>..concerning natural persons</i>		<i>..concerning enterprises</i>	
	yes	no	yes	no

14. If microdata are released, what type of files are released and by which methods?
(Please put an 'x' in the relevant cell)

Type of files Methods for release	<i>..data concerning natural persons</i>			<i>..data concerning enterprises</i>	
	<i>Population Census</i>	<i>Labour Force Survey</i>	<i>Other data</i>	<i>Business Statistics</i>	<i>Administrative data</i>
Public Use Files (PUFs)					
Microdata for research (MUCs)					
Synthetic Datafiles					
Access in a controlled setting (OnSite facility)					
Online access					
In any other form (please specify)					

15. What are the major problems in your office concerning statistical confidentiality development (legal, methodological, organisational, software, training, etc.)?

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

16. Does your country/organisation receive technical assistance from other countries to help with the implementation of disclosure control? If yes, please indicate from which countries or organisations you receive technical assistance and the nature of the technical assistance.

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

17. In which confidentiality related areas could your office require technical assistance?

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

II. Microdata (Public use files)

18. What kind of disclosure limitation methods do you use when releasing **Public Use Files** (PUF's that are generally available)

(Please put an 'x' in the relevant cell)

Disclosure Limitation Method for PUF's	<i>..data concerning natural persons</i>			<i>..data concerning enterprises</i>	
	<i>Population Census</i>	<i>Labour Force Survey</i>	<i>Other data</i>	<i>Business Statistics</i>	<i>Administrative data</i>
Geographical or population thresholds					
Sampling					
Top and bottom coding					
Recoding data into broader categories					
Deletion of data items (local suppression)					
Data Swapping					
Deletion of especially sensitive records					
(Multiple) imputation					
Generating synthetic data (e.g. by resampling)					
Micro aggregation					
Noise addition					
Post Randomisation					
Other Techniques (please specify)					

C E N E X S D C

a **CEN**tre of **EX**cellence for **Statistical Disclosure Control**

19. Do you apply risk assessment methods for Public Use Files (eg. μ -ARGUS threshold rule or Franconi-Benedetti approach);
 (please indicate, how satisfied you are with your current solution. (Scale 1–5))

	Yes	No	If yes: are you satisfied? Scale 1-5 (1 not satisfied 5 very satisfied)
μ -ARGUS threshold rule			
Franconi-Benedetti approach)			

20. Do you use checklists available in your institution (eg containing rules on the detail of the released variables)? Please specify.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

21. Do you use generally available software for data protection (e.g. μ -ARGUS) or own procedures?
 (Please indicate, how satisfied you are with your current solution. (Scale 1 – 5))

	Yes	No	If yes: are you satisfied? Scale 1-5 (1 not satisfied 5 very satisfied)
μ -ARGUS			
Own procedure			

22. If you use your own procedure please indicate.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

III. Microdata (Micro data under contract)

23. What kind of disclosure limitation methods do you use when releasing **Micro data Under Contract** (MUC's that are only available to selected researchers)

Disclosure Protection Method for Micro data Under Contract	<i>..data concerning natural persons</i>			<i>..data concerning enterprises</i>	
	<i>Population Census</i>	<i>Labour Force Survey</i>	<i>Other data</i>	<i>Business Statistics</i>	<i>Administrative data</i>
Geographical or population thresholds					
Sampling					
Top and bottom coding					
Recoding data into broader categories					
Deletion of data items (local suppression)					
Data Swapping					
Deletion of especially sensitive records					
(Multiple) imputation					
Generating synthetic data (e.g. by resampling)					
Micro aggregation					
Noise addition					
Post Randomisation					
Other Techniques (please specify)					

C E N E X S D C

a **CEN**tre of **EX**cellence for **Statistical Disclosure Control**

24. Do you apply risk assessment methods for Micro data Under Contract (eg. μ -ARGUS threshold rule or Franconi-Benedetti approach);
 (please indicate, how satisfied you are with your current solution. (Scale 1–5))

	Yes	No	If yes: are you satisfied? Scale 1-5 (1 not satisfied 5 very satisfied)
μ -ARGUS threshold rule			
Franconi-Benedetti approach)			

25. Do you use checklists available in your institution (eg containing rules on the detail of the released variables)? Please specify.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

26. Do you use generally available software for data protection (e.g. μ -ARGUS) or own procedures?
 (Please indicate, how satisfied you are with your current solution. (Scale 1 – 5))

	Yes	No	If yes: are you satisfied? Scale 1-5 (1 not satisfied 5 very satisfied)
μ -ARGUS			
Own procedure			

27. If you use your own procedure please indicate.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

IV. Tabular data (magnitude tables)

28. What kind of disclosure limitation methods do you use when releasing magnitude tables

(Please put an 'x' in the relevant cell)

Disclosure Protection Method	.. data concerning natural persons			.. data concerning enterprises	
	Population Census	Labour Force Survey	Other data	Business Surveys	Administrative data
Minimum cell count rules ('rule of 3' etc)					
Dominance rules					
Prior-posterior rules (p% rule)					
Secondary cell suppression					
Rounding					
Adding noise / blurring of cell totals					
Recoding variables to reduce detail					
Other techniques (please specify)					

29. Foreign trade statistics.

Some countries apply a special rule for the protection of foreign trade statistics. Only those enterprises that actively ask for protection are checked for disclosure risks. Do you apply this procedure?

(Please put an 'x' in the relevant cell)

Yes	No

30. If you do cell suppression, how do you assign secondary suppressions?
(Please put an 'x' in the relevant cell)

	Yes	No
Manual procedures		
Own software solution		
T-ARGUS Modular		
T-ARGUS Hypercube		
Other	if yes, please specify	

Other software used:

.....

.....

.....

.....

.....

31. If you neither use nor plan to use τ -ARGUS:

Have you ever tested τ -ARGUS?
(Please put an 'x' in the relevant cell)

Yes	No

Which are your reasons for not using τ -ARGUS
(Please put an 'x' in the relevant cell)

	Yes	No
Too difficult to use		
Too many secondary suppressions		
Does not fit into our production environment		
Not stable enough at the time of testing	if yes, please report date and version of last testing: Version: Date:	
Other	If yes, please specify	

Other reasons for not using τ -ARGUS:

.....
.....
.....
.....
.....
.....
.....
.....

V. Tabular data (Frequency tables)

32. Which rules are applied to identify sensitive cells?
(Please put an 'x' in the relevant cell)

	<i>Census data</i>	<i>Survey data</i>	<i>Administrative data</i>
Dominance rule			
Prior/posterior rule (p% rule)			
Frequency rule			

33. Which measures are taken for the protection of the tables?
(Please put an 'x' in the relevant cell)

	<i>Census data</i>	<i>Survey data</i>	<i>Administrative data</i>
Recoding/collapsing			
Cell suppression			
Rounding			
Noise addition/distortion			

VI. Onsite Data laboratory/Remote Execution/Remote Access

34. Do you provide access to micro data to researchers in a secure setting (Data laboratory, OnSite facility) in your office? *(Please put an 'x' in the relevant cell)*

Yes	No

If yes could you quantify the number of requests per year?

35. How do you check the results that will leave the secure room?

.....
.....
.....
.....
.....
.....
.....
.....

36. Do you provide a remote execution service or remote access (via the internet) to micro data to researchers?
(Please put an 'x' in the relevant cell)

Yes	No

If yes could you quantify the number of requests per year?

.....

37. If you provide remote access, could you please indicate how the secured connection is set up?

.....
.....
.....
.....
.....

.....
.....

38. How do you check the results that will be made available to the researchers?

.....
.....
.....
.....
.....
.....
.....
.....

SDC contact

It is without saying that we will treat the answers of the above questionnaire with the utmost confidentiality

However for further contact and exchange of information the CENEX team would be grateful to have a list of SDC-contact persons for each country.

Country	
Institute	
Contact person for legal/organisation issues	
Name	
Address	
Tel	
Email	
Fax	
Contact person for statistical/technical issues	
Name	
Address	
Tel	
Email	
Fax	