Strategies to achieve SDC harmonisation at European level: multiple countries, multiple files, multiple surveys

Daniela Ichim and Luisa Franconi
Istat, DCMT, Via C. Balbo 16, 00184, Roma, Italy

Version 1.0

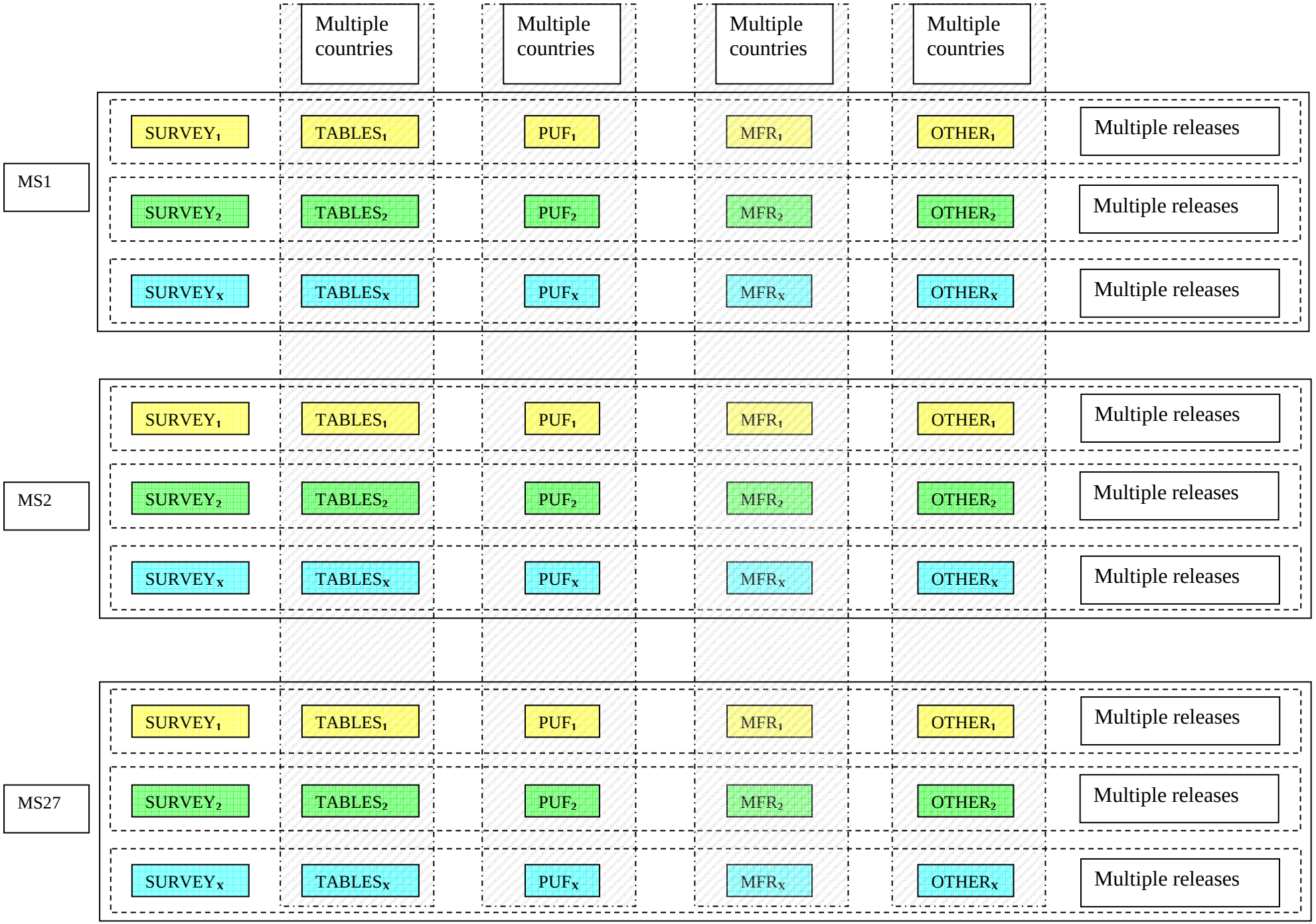
Summary

In this report some preliminary considerations and an initial proposal are made in order to harmonise the anonymisation of microdata files at European level. If an increase of the number of the Member States participating to the European dissemination is aimed, a change in strategy is needed. To reach this goal, tight and systematic collaboration between the partners of the European Statistical System is crucial as stated also in "Communication on the production method of EU statistics: "a vision for the next decade"" (Eurostat 2009).

The document is divided into three parts. In the first part a proposal for the harmonisation of different statistical disclosure limitation procedures from several member states is presented. Here we present the case of microdata file but the same approach could be successfully applied to other types of releases as well. Our proposal is based on two pillars: in the methodological part, contrary to the proposal of Pérez-Duarte (2009), the harmonisation concept is defined by means of a set of minimal requirements on both the input and the output of the anonymisation process. In the organisational part the burden is shared among actors in the ESS: Eurostat and the MSs. A proposal for a possible implementation of both the methodological and procedural/organisational framework is sketched. In the second part issues related to the release of multiple files from the same survey i.e. from the same original dataset are sketched. The release of multiple files is a new feature at European level stemming from the introduction of the public use file (PUF) concept in the new regulation on European statistics. This implies that for the same survey both a public use file and a microdata file for scientific purposes (MFR) may be available: care must be taken in designing such files in order to avoid incoherence. Finally, in the third part the problem of the impact on the coherence of an anonymisation procedure of the release of a system of surveys is briefly explored.

The dimensions presented in the three parts of the document are summarised graphically in figure 1 where the vertical transnational anonymisation (part 1) is linked to the horizontal multiple releases of different types of files (part 2) on multiple surveys (part 3).

Figure 1: Summary of the dimensions of the harmonisation problem: vertical boxes represent transnational releases (part 1 comparable dissemination), horizontal dashed (uniform colour) boxes represent release of multiple files (part 2: different types of file) and horizontal large boxes represent release of related surveys (part 3)



Part 1: Multiple Countries

Harmonised dissemination of European microdata files

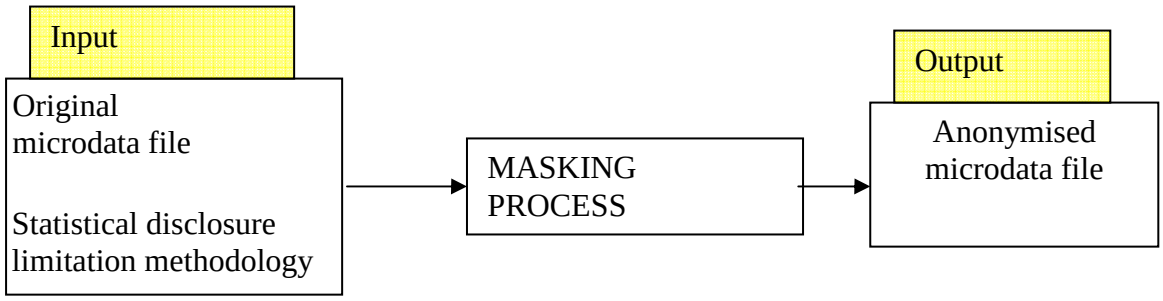
1. Introduction

This first part of the document proposes a framework, given a predefined survey, for harmonising the anonymisation procedures among different member states. This procedure could be applied to any type of microdata file, e.g. public use files or microdata files for scientific purposes (MFR) at European level under the umbrella of Regulation EC 831/2002 and also to other types of statistics. Their corresponding original microdata stem from an harmonised process usually ruled by European regulations which are mandatory for the Member States sharing common definition and common structure. In section 2 we describe the key characteristics of a European release as opposed to a release in a MS highlighting the structural constraints that should be considered and recognising different situations in Europe. In section 3 the current dissemination procedure is analysed. If an increase of the number of the Member States participating to the European dissemination is aimed, a change in strategy is needed. Section 4 contains a proposal of how to reach a harmonised dissemination by improving both the input and the output of the anonymisation process. Section 5 presents some conclusions.

2. European anonymisation process

The core of any dissemination procedure is the anonymisation process. The input of this process has two main parts: the original microdata file and the statistical disclosure limitation methodology that limits the disclosure risk and still provide utility to users. The output of the masking process is a microdata file to be released. This procedure is summarily described in figure 2.

Figure 2 Input and output of the statistical disclosure limitation process**.**



The input microdata files contain the original survey data collected by Member States. For the surveys mentioned in EC Commission Regulation 831/2002, data collection and processing are generally harmonised at European level. What makes the European anonymisation procedure different from an anonymisation procedure in a single Member State is the complexity derived from 27 different cases and situations. The anonymisation of European microdata files ought to take into

account both organisational heterogeneity of Member States and their need, right and duty to respect their own standards. In the next two sections we describe these two concepts.

2.1 Structural constraints: the organisational heterogeneity of Member States

The organisational heterogeneity of Member States is visible in several dimensions. Without being exhaustive, some of these dimensions are listed below. It should be observed that the dissemination of European microdata files should deal with all these features.

a) Law

Legislation in Member States obliges the collecting data institution to guarantee the confidentiality of respondents. The only responsible institution is the data collector although the possible harm is propagated throughout the whole ESS.

b) Organisation of the Statistical System

According to each national statistical system organisation, the data might be collected by a National Statistical Institute or by some other type of entity, for example a minister or a research institute. This is an important issue as national statistical laws may oblige **only** some type of organisations to preserve the confidentiality of respondents and not others. Moreover, the data collection via administrative registers is another type of organisation of a Statistical System. From now on, for simplicity, we will refer only to National Statistical Institutes as the institution carrying out the survey.

c) Access to original confidential microdata

Some Member States allow access to the original microdata, some others may not allow such access. Or, at least, (international) access to the original microdata might be extremely difficult.

d) Microdata transmission

Some Member States have the legal possibility of transmitting the original microdata to other institutions, under bilateral agreements. In some countries the transmission of microdata (even to Eurostat) is possible only if this is accounted for in a specific regulation that obliges the Member State to do so.

e) Microdata dissemination

Some Member States have the legal possibility of disseminating anonymised microdata files, some other MS may not have such possibility.

Also it is possible that a Member State may easily allow the dissemination of some kind of microdata (e.g. social), while strictly prohibiting the dissemination of other data types (e.g. enterprises, or indeed the other way around).

2.2 Different situations in Member States

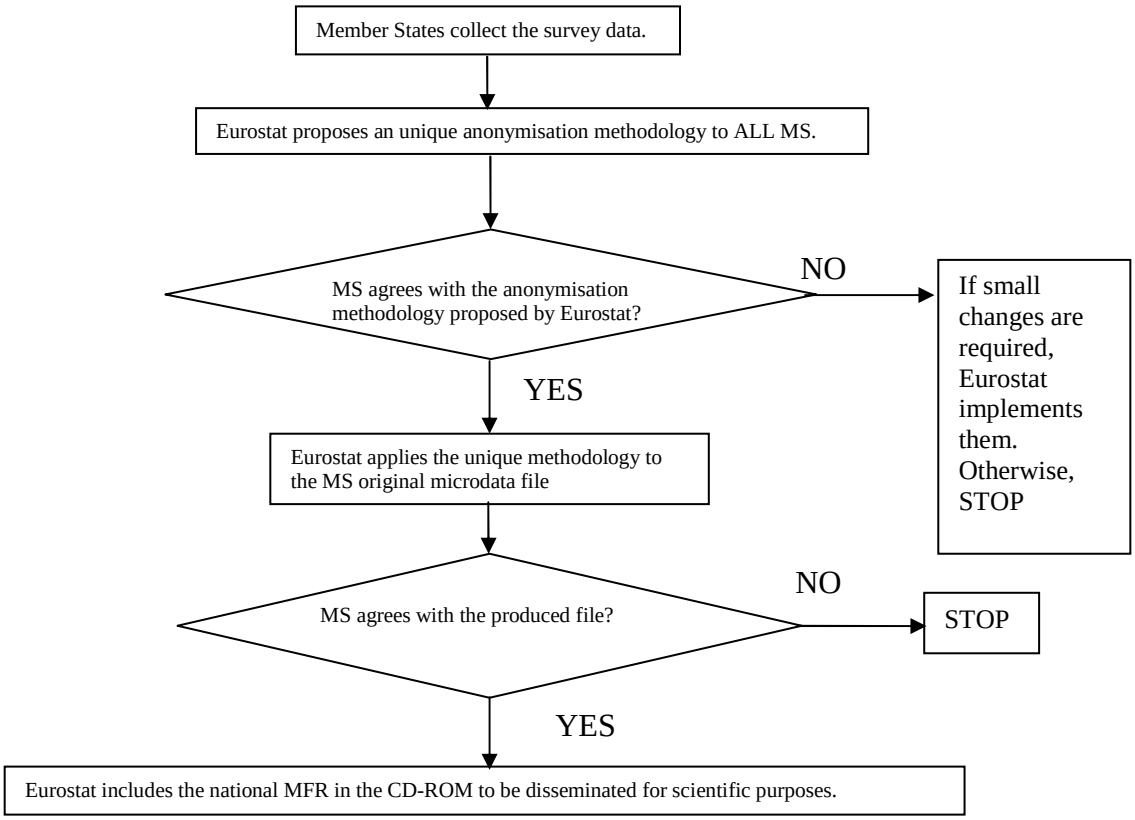
Each Member State should decide on its participation to a given dissemination channel. For example, the Member States may agree or disagree on an anonymisation procedure and they may or may not allow the release of a certain data set to a particular project under EC Regulation 831/2002. The release of an MFR is not compulsory. Nonetheless, the NSIs have the mission to provide society the information needed. That's why, if the national legislation allows it, MS are generally willing to disseminate anonymised microdata files, in the provision that the anonymisation process is up to their national standards. However, there are large countries and very small countries with completely different economic structures, with different perceptions of the disclosure risk and different approaches to confidentiality. Also, as disclosure limitation is a recent field of work for many NSIs, significant differences are visible amongst MSs. An anonymisation process needs to take into account all such different standards.

3. Analysis of the current European anonymisation flow

3.1 Current anonymisation flow

Figure 3 summarily presents the current flow of the anonymisation of an European MFR[1]. Usually Eurostat proposes a masking procedure to be adopted, carries out the microdata anonymisation, produces the CD-ROM and take care of its subsequent delivery. Of course, the same strategy could be easily extended to other forms of microdata dissemination, if agreement is got on this workflow.

Figure 3 Current flow of anonymisation of European microdata files for research purposes.



3.2 Analysis of critical points

The central role of Eurostat in the current anonymisation flow may be easily observed. Starting from the methodological proposal and ending by the CD-ROM dissemination, Eurostat is the most important actor and therefore the one that holds most of the work and responsabilities. From the harmonisation side, this is surely a very attractive feature of the European dissemination procedure. If on the one hand the implementation of a single method is an important feature for users and is a crucial simplifying step for Eurostat (who might find difficult to apply different methodologies for different MSs); on the other hand, being a rigid solution it might limit the possibility of anonymisation for a large number of MSs. We will see how such situation could be modified by creating formal forms of cooperation that allow to share the burden inside the ESS in order to develop more sustainable anonymisation procedures in a predefined methodological framework.

Given a certain level of data utility, an European microdata file needs to satisfy a crucial requirement in order to meet the expectations of users: it has to be representative of all MS in

[1] Currently, this is the only type of microdata released at European level.

Europe. It might be possible that a single protection method may not be suitable for all Member States (and all real data sets and all waves of the survey). Indeed, the use of a single method does not take into consideration Member States organisational heterogeneity (discussed in section 2.1). If a Member State wants to disseminate its own anonymised microdata file at European level, the only possibility is to agree with the anonymisation methodology proposed by Eurostat. Probably, the only exception to this rule holds for the Member States allowing the dissemination of original microdata to anyone.
In order to increase the number of MSs adhering to the release of a microdata file, some form of flexibility needs to be introduced in the anonymisation process to accommodate for organisational heterogeneity of MSs and their own standards. Alongside flexibility the other pillar on which to build harmonisation is the involvement of other MSs inside the ESS to share the burden of anonymisation.

4. Proposal for a harmonised European anonymisation

4.1 Need for harmonised anonymisation

A harmonised anonymisation of microdata files would be surely profitable for all actors in the release process: users, NSIs and Eurostat.

A harmonised anonymisation would increase the number of MSs releasing their microdata and therefore increase the data utility. Moreover, the corresponding European data set would still share the same harmonisation properties of the original data files as the building phase would be harmonised as well.
At the same time, the National Statistical Institutes should be aware that a harmonised anonymisation would greatly benefit them, too. First, the recognition of structural differences and internal standards would allow more MSs to adhere to the anonymisation. Second, the exchange of experience and competence surely generates improved results.
Finally, Eurostat with the help of MSs with sound experience in the area of SDC would enforce its co-ordinating role at European level by promoting the definition and adoption of a set of common guidelines and by sustaining the implementation of software routines able to be applied to different microdata files.

4.2 A way to achieve harmonised anonymisation

It was pointed out in the previous section that it should be advantageous to design an European anonymisation procedure that takes into account the organisational heterogeneity of the Member States. At a first glance, one might believe that a harmonisation of the disseminated microdata files is very difficult. Since the organisational heterogeneity of Member States is a rigid constraint, we believe that a harmonised European anonymisation of microdata files could be achieved twofold: 1) modelling the input of the anonymisation methodology and 2) modelling the output of the anonymisation methodology. In other words, the harmonisation concept is defined by means of a set of minimal requirements on both input and output of the anonymisation process. The proposed dissemination flow is presented in figure 4. In principle, on the input phase, a significant improvement might be reached by using flexible statistical methods. On the output phase, the definition of a battery of benchmarking statistics and corresponding quality criteria/thresholds could be used to put in practice a comparability concept.
The changes to the European anonymisation flow are survey independent. Nonetheless, benchmarking statistics and quality criteria should be survey specific and should be applied to each survey wave. If appropriate, the same benchmarking statistics and quality criteria/thresholds could be applied to consecutive waves.

4.2.1 Working on the INPUT of the process: statistical methodology

a) A single method

Currently the European anonymisation procedure foresees the application of a single statistical disclosure limitation methodology. This strategy surely has the lowest costs in terms of implementation, testing and application. It might be believed that this strategy also produces highly harmonised results. Nonetheless, the application of the same statistical disclosure limitation method to different data sets might produce different qualitative and quantitative results.

Given the organisational heterogeneity of the Member States, it is hard to believe that there exists a method that best suits the requirements and standards of 27 countries. The best practical option would be the choice from a list of candidate methods. Anyway, it should be observed that the choice of the statistical disclosure limitation methodology is not an easy task. Today many statistical disclosure control methods exist, each one with its own merits and drawbacks. To our knowledge, there is no final winner. The situation is much more complicated when both risk and data utility are considered as the scientific community didn't find a rigorous way to compare all the protection methods.

The choice (selection/definition) would not completely solve the acceptance problem of the Member States. Because the participation to this dissemination channel is not mandatory, even if a method is agreed, one Member State could still refuse its application. This could mainly concern the Member States that today cannot legally disseminate anonymised microdata files. If, in future, their national law would change, those Member States could still not agree with an *a-priori* selected anonymisation method.

b) More methods

A simple strategy that possibly could take into account the Member States organisational heterogeneity is the creation of a **list** of pre-defined candidate methodologies. This approach would surely require some more resources spent in implementation and testing.

An advantage could be the possibility to increase the number of Member States agreeing to disseminate anonymised microdata files at European level. For example, in the framework of enterprise microdata European dissemination, the MSs that could have accepted the individual ranking applied irrespective of the categorical structural key variables (i.e. irrespective of the stratification) have already agreed on. **If an increase of the number of the Member States participating to the European dissemination is aimed, a change in strategy is needed.**

The usage of a list of candidate statistical disclosure limitation methodologies could activate a sort of virtual competition among methods. Different strategies could be implemented and tested on real survey data. In medium-long term, empirical evidence would guide the selection of the most suitable strategy for the analysed survey.

c) Flexible methods – parameterisation

This proposal is just an extension of the previous one (point b), aiming at increasing the number of Member States disseminating anonymised microdata files at European level. Different variants of the same statistical disclosure limitation methodology could be easily implemented and tested. For example, the implementation of the individual ranking could depend on the microaggregation parameter *p*; then, each Member State could select its own value for this parameter *p*, e.g. 3 or 5 or some other value. The implementation of a statistical disclosure limitation methodology with respect to different stratification domains is another form of flexibility. For example, the methodology could be applied to the entire microdata file or to the domains defined by the categorical key variables (generally the structural categorical variables). In other words, by simply

changing the values of some parameters, the statistical disclosure methodology could be more easily adapted to many Member States.

Another option could be the usage of sound statistical methods allowing, by definition, the output control. That is, some output quality indicators could already be taken into account by the statistical disclosure limitation methodology. For example, in the framework of continuous variables, if the preservation of weighted totals is required, using a methodology that by definition satisfies this constraint (e.g. adding noise or regression models) could be very helpful. Moreover, the usage of such statistical methods would allow a sound study of the statistical properties of the anonymised microdata files.

4.2.2 Working on the OUTPUT of the process: Comparable dissemination

Data utility / data quality are one of the most important characteristics of the output of the European anonymisation flow. Timeliness, consistency, efficacy and comparability are only some dimensions of data quality which are of interest to the users. Data utility is neither easy to define nor easy to quantify. We propose to assess it through the definition of benchmarking statistics for the type of data under analysis. Then, thresholds / quality criteria on these benchmarking statistics should be set. Moreover, possible remedies should be indicated for the cases when the quality criteria are not met. For the definition of both benchmarking statistics and their corresponding thresholds / quality criteria, cooperation between survey experts and methodologists is strategic. The most relevant statistics (benchmarking statistics) could be identified from a review of previous analyses performed on the survey data and from information given by users groups.

The comparable dissemination procedure may be summarised in the following steps:

a. Given a single survey (ECHP, FSS, FOBS, CVTS, XXX or YYY),

b. Indicate a list of non-statistical quality indicators $Q_1, Q_2, \ldots, Q_n$

c. Indicate a list of benchmarking statistics $S_1, S_2, \ldots, S_m$,

d. Indicate the thresholds / quality criteria $C_1, C_2, \ldots, C_M$, $M \geq m$ associated to the statistical indicators $S_1, S_2, \ldots, S_m$

e. Suppose that a candidate statistical disclosure limitation methodology is applied to the original microdata file.

f. If the anonymised microdata file satisfies each of the non-statistical criteria $Q_1, Q_2, \ldots, Q_n$ and each of the quality criteria $C_1, C_2, \ldots, C_M$ corresponding to the statistical indicators $S_1, S_2, \ldots, S_m$, then the file should be accepted for dissemination at European level.

Using the above procedure, at least from the point of view of the considered statistical (and non statistical) indicators, the comparability among the Member States would be guaranteed.

Examples:

i. Non-statistical indicators
   1. Fulfilment of a dissemination deadline
   2. Compatibility with a predefined electronic format
   3. Preservation of the original microdata file structure

ii. Statistical indicators
   1. Preservation of an informative content of the most important variables
   2. Preservation of an informative content of the survey specific variables (generally the confidential variables)
   3. Means of the most important variables, by stratification domain

4. Variances of the most important variables, by stratification domain
5. Distributions of the most important variables, by stratification domain
6. Already published statistics (tables)

iii. Quality criteria/thresholds

1. Preservation of a minimum level of detail on categorical variables (for example NACE 2-digits or NUTS at regional level)
2. Bounds on variations (e.g. the anonymised total should not differ from the original total by more than given percentage)
3. Coherence with the already published statistics

Observations

1. The procedure should be constructed and applied to each survey. This dependency on survey is due to the fact that the benchmarking statistics and their quality criteria/thresholds are strongly related to the survey type, to the kind of microdata and to the kind of analyses performed on such microdata.
2. The procedure should be constructed and applied to each survey wave (see item 3, too). The same motivations as above.
3. In order to ensure the comparability among distinct waves of the same survey, the same statistics and quality indicators should be chosen.
4. For each statistic S indicated in step c, different quality criteria/thresholds may be indicated. Consequently, $M \geq m$. For example, one might bound the total variation, but at the same time, the total computed on anonymised data should be nonnegative.
5. The key point in the comparable dissemination procedure is the definition of the benchmarking statistics and their thresholds / quality criteria. Anyway, the importance of non-statistical criteria should also be stressed.

An example of such an approach can be found in Franconi and Ichim (2009).

Part 2: Release of Multiple Types of Files

Issues on the release of a public use file and a microdata file for scientific purposes from the same survey

The release of different files from the same microdata is a new issue at European level. It derives from the entry into force of the new regulation on European statistics, Reg. (CE) 223/2009, introducing the definition of public use file (PUF) besides the already implemented file for scientific purposes (MFR). Although new at transnational level, the instances of the production of multiple files from the same dataset is however growing very fast as international institutes or EU or world based projects urges the need to develop "customised files" that could be compared at international level: recent examples are the generation and gender project (http://www.unece.org/pau/ggp/Welcome.html) or the IPUMS project (https://international.ipums.org/international/). The problem encountered in such situation is a simple one: the file required by international institutions is generally not a problematic one *in itself*, but it might differ for some classifications from other files already released at national or EU level. For example nowadays an international organisation could require for a certain survey a level of geography not extremely detailed but, at the same time, it would need of indications on the socio-demographic characteristics of the municipality. Such requirements could then be in contrast with previously released files with more detailed geography where information on the size of the town or its rural/urban nature were not present. This type of problem is the microdata counterpart of the linked tables problem and as for the latter an optimal solution can be found only when the different

data to be released are anonymised at the same time. Therefore to be optimal at European level, the anonymisation of different types of microdata files should be planned at the same time.

At national level the multiple types of files (multiple releases) problem has already been encountered as the production of different files for different users is becoming a widespread practice (see for example Trottini *et al.* (2006) for a dissemination strategy proposal for the household expenditure survey in Italy).

However, despite of the need of data anonymisation procedures targeted to the different data users, the problem of releasing different files is still at an embrional stage and indeed very rarely approached in practice (besides the previous citation an example of such implementations can be found in Abowd and Lane, 2003). This is due to the cost associated with a real differentiated data dissemination strategy and the complexity of its implementation. What is most commonly applied in most Member States adopting a dual dissemination (PUF and MFR) is the mere adoption of more aggregated classifications for the categorical variables and various forms of top and bottom coding as well as the introduction of bands for the continuous variables. This causes the needed drastic decrease of the risk of disclosure but presents, as a side effect, a severe drop in the information content of the microdata file. Also, till the present time, the dual release process at national level has been, in most cases, a *controlled release* also when "general use files" were involved. This means that in most countries the current procedure to release a microdata file implies the need for a formal request (therefore implying the clear identification of the user), specifications of the foreseen uses to be provided and some sort of confidentiality statement agreed. However, a new concept of PUF need to be developed where the dissemination mean will be the web and where a simple download could be the procedure to gain it. Possibly, in the future, for European PUF there will be no list of users, no guarantee of reasons for access, no control on uses. This implies a completely new approach to the definition of a PUF with respect to the ones we are currently used to. The risk of disclosure will be surely higher as the risk is related also to the dissemination mean. However, new methods and a new attitude towards statistical disclosure control could supply strategies where the public nature, i.e. the free availability of the PUF, should not be the synonym of the production of files showing very limited interest and analytical validity for the final users. Targeted utility-based perturbation methods or, more recently, synthetic data generation methods can be used to release perturbed data that still present interesting level of information content.

Certainly PUF and MFR must be hierarchically designed in terms of information content (see Trottini et al. 2006). This means that all the information in the PUF should also be contained in the corresponding MFR. The hierarchical structure of the two data sets greatly simplifies assessment of the disclosure risk and information loss associated with the anonymisation procedure. Because of the hierarchy, in fact, there is no gain for a user having access to the MFR, to access the PUF. The hierarchy requires coherence in the choice of the variables to be included in both files and on the corresponding level of details. The inclusion of a variable in the PUF implies the inclusion in the MFR; non nested classifications for the same variables should not be allowed, and so on. The use of strategies outlined in part 1 of this report (comparable dissemination) will allow the selection of the list of variables to be included and the agreement on the basic and broad classification for the PUF. Then, details on single respondents could be provided inside the broad band by means of perturbation or synthetic generation. So if a ten year classes for the variable age is agreed for the PUF, the age in single year of a the respondent could be generated inside the corresponding broad ten year band using also different methodologies. We foresee that the use of comparable dissemination coupled with perturbation/data generation procedures couls allow both the definition of interesting PUF for the users and coherent multiple releases.

Part 3: Release of Multiple Related Surveys

In many MSs is becoming common the definition of a system of surveys structured in such a way that a basic questionnaire is present yearly but different modules are rotated year after year in order to monitor cyclically a phenomenon of interest (in Italy the multipurpose system of surveys, at European level the future general social survey). Without reaching this level of definition social surveys present always the same socio-demographic characteristics: gender, age, marital status, etc. It would be extremely appealing if a systematic recognition of such variables would be identified and harmonisation of the SDC practices applied in related surveys would be achieved.

Conclusions: tight collaboration between Eurostat and MSs

The harmonisation of surveys and processes throughout Europe is recognised as a key feature for the future of European statistics. In this report we identify the dimensions in which the anonymisation process at European level should develop, highlight some of the corresponding critical points and cast possible ways to approach a solution.

The underlying idea is to develop a framewok for harmonising the anonymisation process with the active cooperation of MSs by proposing possible sound alternative methodologies and by setting benchmarking statistics and thresholds on such statistics in order to guarantee the users with a minimum standard of quality throughout the continent. The framework and such indicators could be simply part of the structure of the quality report that each survey under European regulation needs to comply with. The flexibility allowed by the process will increase the number of MSs adhering to the dissemination and therefore the number of data sets available to users and will foster the development of knowledge in the field of the statistical disclosure limitation methods within the ESS.

The comparable dissemination framework implies an initial investment in identifying the benchmarking statistics and relative thresholds / quality criteria but, then, the whole procedure is expected to become part of the production process. Also this initial stage can be performed with the help of Member States who have gained already experience in this field by creating an institutional form of collaboration on this particular area of expertise. It would be extremely beneficial if Eurostat would formally join together the experienced and willing MSs to develop and test the anonymisation process or even to take part to the production of the anonymised files. This systematic collaboration between the partners of the European Statistical System would allow sharing the burden that is currently on the shoulder of Eurostat and transferring the knowledge and expertise across the ESS, as suggested in the “Communication on the production method of EU statistics: "a vision for the next decade"” (Eurostat 2009).

References

J. M. Abowd and J. Lane (2003): Synthetic data and confdentiality protection. In *Workshop on Microdata*, Stockholm, Sweden, August 2003.

L. Franconi and D. Ichim (2009): Community Innovation Survey: Comparable Dissemination, *Joint UNECE/Eurostat work session on statistical data confidentiality*, 2007, ISBN 978-92-79-12055-8, Theme: General and regional statistics, Collection: Methodologies and working papers, available at

http://www.unece.org/stats/documents/2007/12/confidentiality/wp.2.e.pdf, DOI number DOI: 10.2901/Eurostat.C2007.004**.**

D. Ichim (2008): Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research, *Proceedings of Q2008, European Conference on Quality in Official Statistics*, available at http://q2008.istat.it/sessions/24.html.

S. Pérez-Duarte (2009) Harmonisation of anonymisation practices through partially synthetic files. Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, December 2009. available at http://www.unece.org/stats/documents/2009/12/confidentiality/wp.7.e.pdf.

M. Trottini, L. Franconi, S. Polettini (2006): Italian Household Expenditure Survey: A Proposal for Data Dissemination, in: *Privacy in Statistical Databases* 2006, (eds.) J. Domingo-Ferrer, L. Franconi, LNCS 4302, Springer, Berlin pp. 318−333.

Figure 4: proposed anonymisation flow