Strategies to Achieve SDC Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys

Daniela Ichim and Luisa Franconi

Istituto Nazionale di Statistica, DCMT,
Via C. Balbo, 16, 00184 Roma, Italy
{ichim,franconi}@istat.it

Abstract. Preliminary considerations and an initial proposal are made for the harmonisation of different statistical disclosure limitation procedures at European level. Here we present the case of microdata file but the same approach could be successfully applied to other types of releases as well. The 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 European Statistical System. A proposal for a possible implementation of both the methodological and procedural/organisational framework is sketched. 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 might be available: care must be taken in designing such files in order to avoid incoherence. Finally, the problem of the impact on the coherence of an anonymisation procedure of the release of a system of surveys is briefly explored.

Keywords: comparability, privacy in official statistics, SDC governance, public use file, microdata file for research.

1 Introduction

Under the umbrella of Regulation EC 831/2002 Eurostat releases European microdata for research purposes for several surveys ranging from social surveys such as the Labour force survey to business microdata (Community Innovation Survey ─ CIS ─ and Structure of Earning Survey ─ SES). Such microdata stem from a harmonised process usually ruled by European regulations which are mandatory for the Member States (MSs) sharing common definitions and common structure. However, in many cases, MSs do not agree with the anonymisation methodology proposed by Eurostat leading to the release of datasets that do not cover the whole of Europe. If an increase

in the number of MSs participating to the European dissemination is aimed, a change in strategy is needed. In this paper preliminary considerations and an initial proposal are made in order to allow for more flexibility in the implementation of statistical disclosure limitation strategies and harmonise the anonymisation of microdata files at European level. Harmonisation is needed also to deal with the release of multiple files from the same survey. In fact, besides the release of European microdata for scientific purposes (MFR) the new European legislation on statistics allows for the dissemination of public use file. This introduces a further dimension to the multiple countries problem due to the release of multiple types of microdata for external users (PUF and MFR) from the same survey. Finally, in the last few years, a new way to systematically investigate the complexity of modern societies has lead to the development of systems of surveys that, although focussing in different areas, still present common structures and characteristics. SDC methods applied to such systems ought to be coherent in order to avoid limitations for the users.

In section 2 we describe the current limitation to the release of microdata in Europe and show how the same type of problems may occour in other international settings. In section 3 we address the multiple countries dimension of the European release of microdata by proposing a general framework that allows for flexibility within known boundaries. In Section 4 we sketch the problem of multiple releases from the same survey. Finally, the need to address the coherence of SDC methods when applied to multiple related surveys is briefly explored in section 5. The conclusions are presented in section 6.

2   European Anonymisation Process: Structural Constraints and Different Situations

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 provides utility to users. The *output* of the masking process is the microdata file to be released.

In this paper the input microdata files contain the original survey data collected by twenty seven MSs of the European Union. Usually, but this is certainly the case for the surveys mentioned in EC Commission Regulation 831/2002 that deals with the release of European microdata for scientific purposes, data collection and processing are harmonised at European level. What makes the European anonymisation procedure different from an anonymisation procedure in a single MS is the complexity derived from several different approaches and situations. The anonymisation of European microdata files ought to take into account both organisational heterogeneity of MSs and their needs, rights and duties to respect their own national standards.

The organisational heterogeneity of MSs 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 MSs obliges the data owner, i.e. institution that collects the data, to guarantee the confidentiality of respondents. The responsible institution is the data owner although the possible harm is propagated throughout the whole European Statistical System (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 types 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 to the data owner as the institution who carries out the survey.

c) *Access to original confidential microdata*: Some MSs 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 MSs 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 MS to do so.

e) *Microdata dissemination*: Some MSs have the legal possibility of disseminating anonymised microdata files; some others may not have such possibility. Also it is possible that a MS 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).

Each MS should decide on its participation to a given dissemination channel. For example, the MSs 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 data owner, significant differences are visible amongst MSs. An anonymisation process needs to take into account all such different standards.

2.1   Analysis of the Current Anonymisation Flow and Its Critical Points

Figure 1 summarily presents the current flow of the anonymisation of a European MFR. Usually Eurostat proposes a masking procedure to be adopted; subject to MS agreement, carries out the microdata anonymisation, produces the CD-ROM and takes 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.

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 responsibilities. 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 characteristics 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.
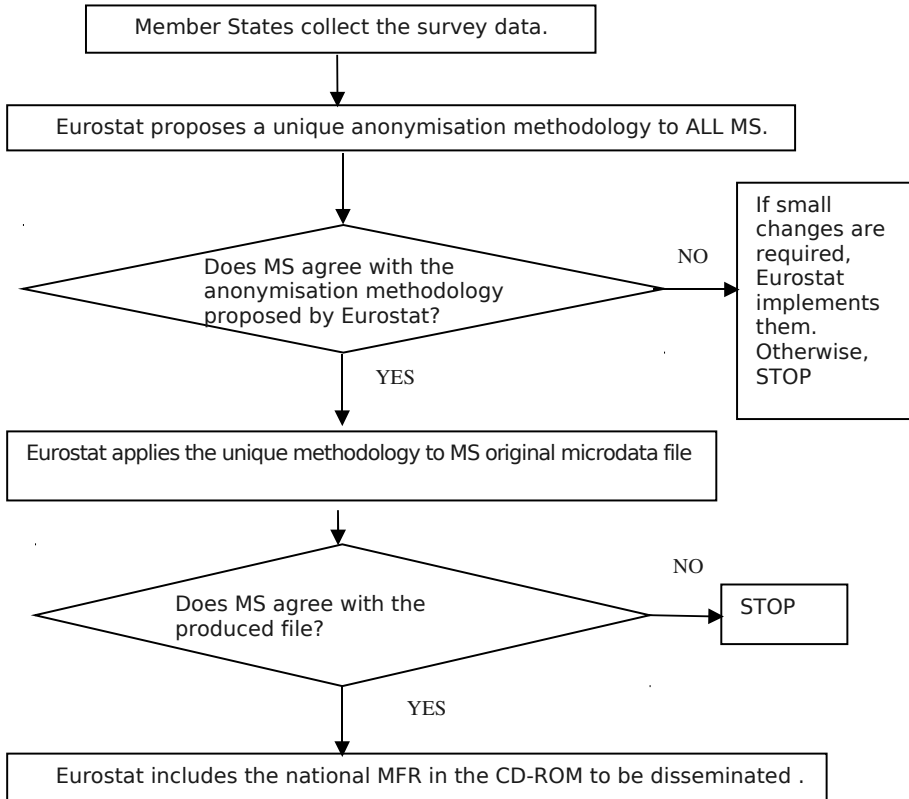

Fig. 1. Current flow of anonymisation of European microdata files for research purposes

Given a certain level of data utility, a European microdata file needs to satisfy a crucial requirement in order to meet the expectations of users: it has to be representative of all MSs in Europe. It might be possible that a single protection method may not be suitable for all MSs (and all real data sets and all waves of the survey). Indeed, the use of a single method does not take into consideration MSs organisational heterogeneity (discussed in section 2). If a MS wants to disseminate its own anonymised microdata file at European level, there is strong pressure to agree with the anonymisation methodology proposed by Eurostat.

In order to increase the number of MSs adhering to the release of a microdata file, some forms of flexibility need 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.

2.2   Other Possible International Settings

In this paper we analyse only some problems related to the European system of information dissemination. However, it should be noted that the same approaches, analyses and solutions could be applied to other information dissemination systems. International organisations like UNECE or OECD make efforts to disseminate information at transnational level. To cite only a few, well-known examples, a) OECD is currently running a project to disseminate microdata files stemming from a variety of labour force surveys, see Brackfield and Ruiz (2010), b) the World Bank is coordinating the International Household Survey Network, c) Integrated Public Use Microdata Series (census microdata for social and economic research) and d) Demographic and Health Surveys maintained by the U.S. Agency for International Development (USAID). Sometimes, due to legal, logistic and cultural differences, it is not realistic to constrain national organizations to strictly follow some approved guidelines; an alternative could then be the proposed approach.

3   Proposal for a Harmonised European Anonymisation

A harmonised anonymisation of microdata files would be surely profitable for all actors in the release process: users, data owners, mainly National Statistical Institutes, and Eurostat. A harmonised anonymisation would increase the number of MSs releasing their microdata and therefore increase 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, data owners 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 experiences and competences 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.

At a first glance, one might believe that a harmonisation of the disseminated microdata files is very difficult. Since the organisational heterogeneity of MSs 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 dissemination flow, as described in section 1, presents an *input* phase and an *output* phase. 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 the 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.

3.1   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 MSs, it is hard to believe that there exists a method that best suits the requirements and standards of twenty seven 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 MSs. Because the participation to this dissemination channel is not mandatory, even if a method is agreed, one MS could still refuse its application. This could mainly concern the MSs that today cannot legally disseminate anonymised microdata files. If, in future, their national law would change, those MSs could still not agree with *a priori* selected anonymisation methods.

b) More methods

A simple strategy that possibly could take into account the MSs 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 MSs 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 MSs 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 MSs 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 MS 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 MSs.

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.

3.2   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 by the following steps:

  a. Given a single survey (CIS, SES, etc.);

  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 $C_1, C_2, \ldots, C_M$, $\quad M \geq m$ 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 MSs would be guaranteed.

Examples of non-statistical indicators are: fulfilment of a dissemination deadline, compatibility with a predefined electronic format, preservation of the original microdata file structure; examples of statistical indicators could be: preservation of an informative content of the most important variables, preservation of an informative content of the survey specific variables (generally the confidential variables), means of the most important variables, by stratification domain, variances of the most important variables, by stratification domain, distributions of the most important variables, by stratification domain, already published statistics (tables). Finally examples of quality criteria/thresholds could be: preservation of a minimum level of detail on categorical variables (for example NACE 2-digits or NUTS at regional level), bounds on variations (e.g. the anonymised total should not differ from the original total by more than given percentage), and coherence with the already published statistics. Some remarks on the process are outlined:

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 statistics 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). A proposal of governance structure is presented in Appendix 1.

4   Release of Multiple Types of Files

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 are however growing very fast as international institutes or EU or world based projects urge 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). In Appendix 2, an overview of the problem of multiple releases is given.

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 embryonic 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 MSs 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 control on reasons for access nor on

uses. This implies a completely new approach to the definition of a PUF with respect to the ones MSs 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 its inclusion in the MFR; non nested classifications for the same variable should not be allowed, and so on. The use of strategies outlined in section 3 (comparable dissemination) would 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 could allow both the definition of interesting PUF for the users and coherent multiple releases.

5   Release of Multiple Related Surveys

In many MSs 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 is becoming common; this is done 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.

6   Conclusions

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 framework 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 MSs 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 Eurostat (2009).

References

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

Brackfield, D., Ruiz, N.: Harmonised Labour Force and Migration Statistics Based on Microdata. In: Joint UNECE/Eurostat work session on statistical data confidentiality (2009), `http://www.unece.org/stats/documents/2009.12.confidentiality.htm`

Eurostat, Communication on the production method of EU statistics: "a vision for the next decade" (2009)

Franconi, L., Ichim, D.: Community Innovation Survey: Comparable Dissemination. In: Joint UNECE/Eurostat work session on statistical data confidentiality 2007 (2009), ISBN 978-92-79-12055-8, Theme: General and regional statistics, Collection: Methodologies and working papers, `http://www.unece.org/stats/documents/2007/12/confidentiality/wp.2.e.pdf`, doi:10.2901/Eurostat.C2007.004

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

Pérez-Duarte, S.: Harmonisation of anonymisation practices through partially synthetic files. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao (December 2009), `http://www.unece.org/stats/documents/2009.12.confidentiality.htm`

Trottini, M., Franconi, L., Polettini, S.: Italian Household Expenditure Survey: A Proposal for Data Dissemination. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 318–333. Springer, Heidelberg (2006)
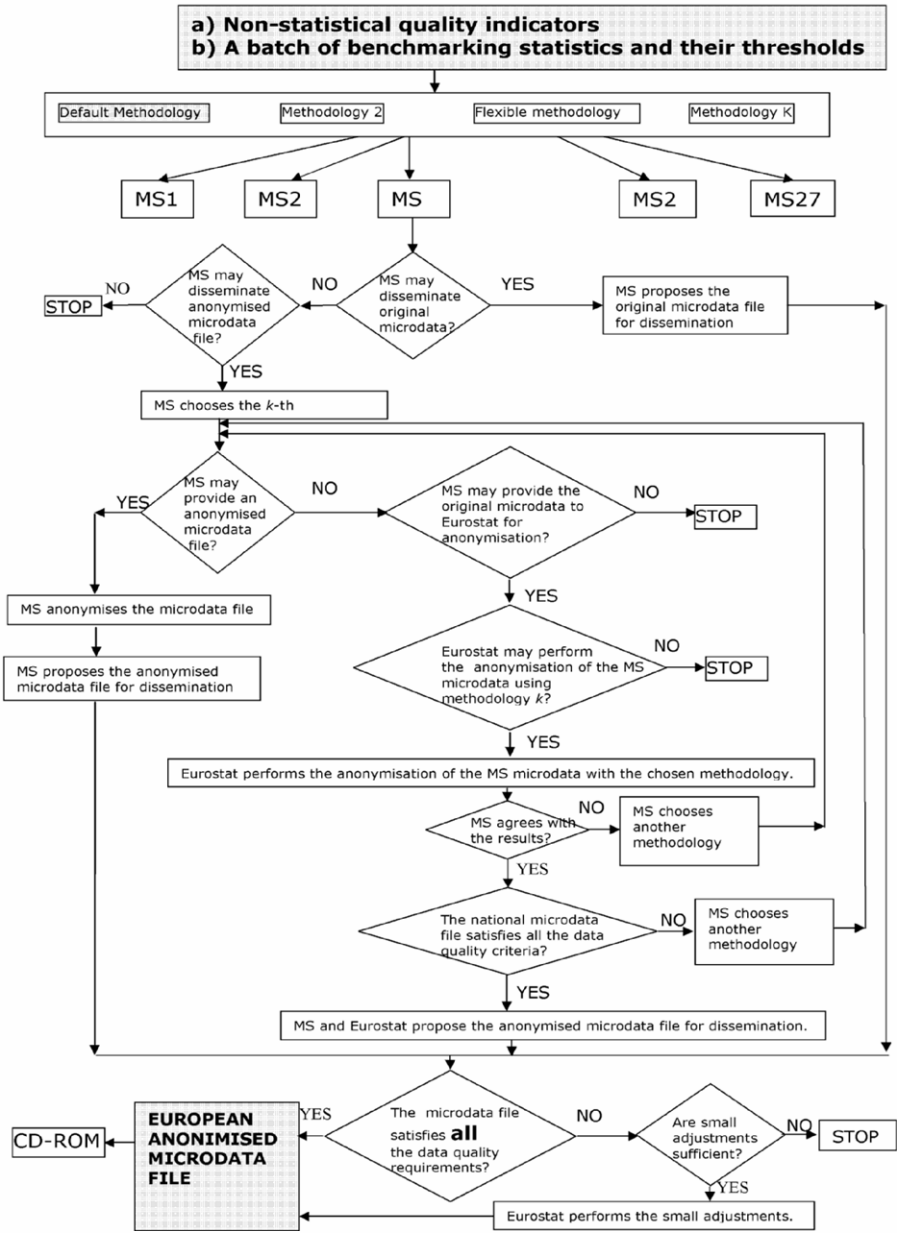
Appendix 1: Harmonised Anonymisation Flow

Fig. 2. Example of harmonised anonymisation flow

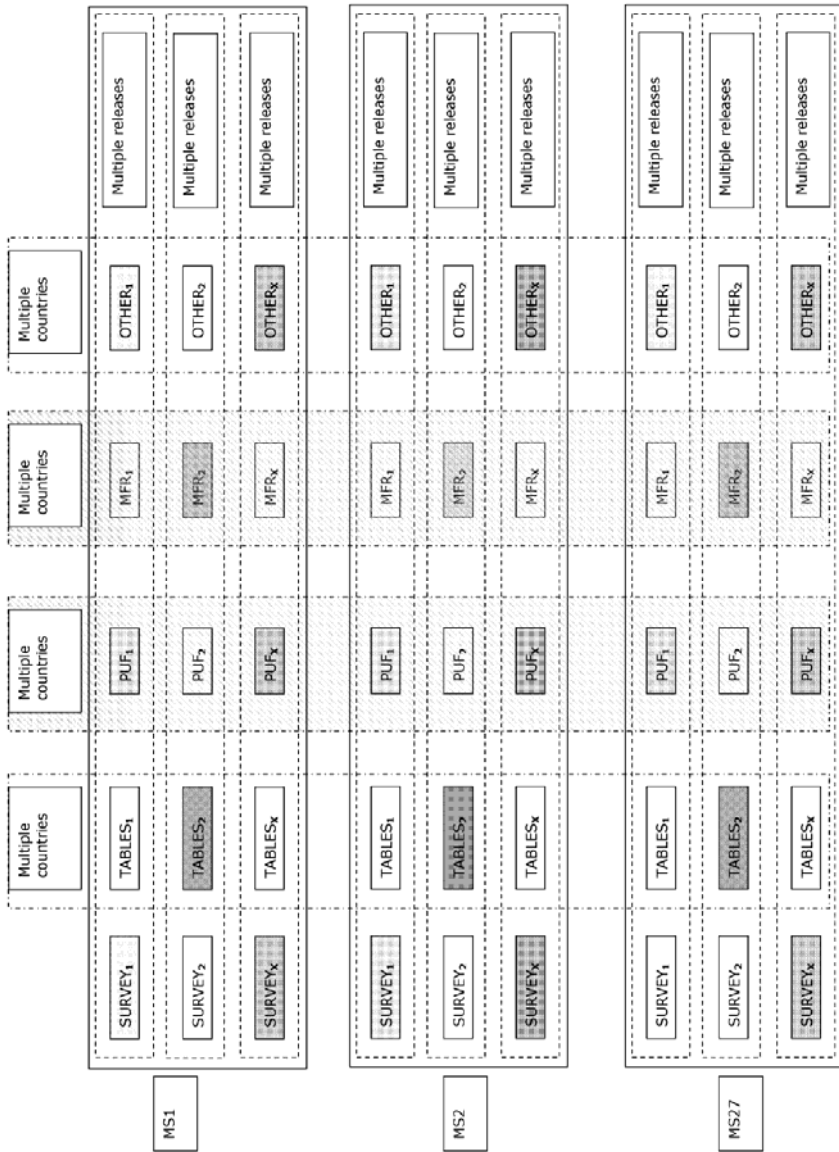## Appendix 2: Dimensions of the Harmonisation Problem

Fig. 3. Summary of the dimensions of the harmonisation problem: vertical boxes represent transnational releases, horizontal dashed (uniform colour) boxes represent release of multiple files and horizontal large boxes represent release of related surveys