



**Economic and Social
Council**

Distr.
GENERAL

ECE/CES/2009/3 Add.1
2 June 2009

Original: ENGLISH

ECONOMIC COMMISSION FOR EUROPE

STATISTICAL COMMISSION

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-seventh plenary session
Geneva, 8-10 June 2009
Item 6 of the provisional agenda

**PRINCIPLES ON CONFIDENTIALITY AND PRIVACY ASPECTS OF STATISTICAL
DATA INTEGRATION**

**RESULTS OF THE CONSULTATION ON THE PRINCIPLES AND GUIDELINES ON
CONFIDENTIALITY ASPECTS OF DATA INTEGRATION UNDERTAKEN FOR
STATISTICAL OR RELATED RESEARCH PURPOSES**

Room paper by the UNECE Secretariat

Summary

The Conference, at its 2006 plenary session requested that a task force be appointed to examine the confidentiality and privacy concerns related to integrated data sets and to develop common principles (ECE/CES/70). The Task Force on Confidentiality and Privacy Aspects of Statistical Data Integration was set up in February 2007.

The Principles and Guidelines on Confidentiality Aspects of Statistical Data Integration Undertaken for Statistical or Related Research Purposes are presented to the 2009 plenary session of the Conference for endorsement, in accordance with the Procedure for Adopting Products and Recommendations by the Conference of European Statisticians (ECE/CES/2006/37/Rev.1). This room paper summarizes feedback from Conference members when the principles and guidelines were circulated for comment during April 2009.

I. SUMMARY OF FEEDBACK

1. The *Principles and Guidelines on Confidentiality Aspects of Statistical Data Integration Undertaken for Statistical or Related Research Purposes* were circulated to the Conference for written comments on 6 April 2009. Responses were received from 32 countries and 2 international organizations, and are summarized below.
2. Twenty-one countries and two international organizations agreed or endorsed the paper without any specific comments: Armenia, Australia, Belarus, Brazil, Bulgaria, Cyprus, Denmark, Estonia, Finland, France, Germany, Greece, Japan, Kazakhstan, Kyrgyzstan, Latvia, Mexico, Norway, Slovakia, Switzerland, United States, Eurostat and the World Bank.
3. Ten countries (Hungary, Ireland, Lithuania, Moldova, Netherlands, New Zealand, Russian Federation, Slovenia, Tajikistan and Turkey) supported the adoption of the principles and guidelines in general, but had specific comments on the text aimed at improving clarity and completeness. Poland had some reservations on the document and considered that it could be further improved. The comments received are organised by topic and presented below.

II. GENERAL COMMENTS

4. New Zealand suggested adding an explanation of why this document is needed – i.e. why should there be particular consideration of integrated datasets, if integration is for statistical purposes, and the same careful protection of privacy, security and confidentiality is applied as for other unit record datasets.
5. They added that some principles are more restrictive than the standards applied for "single source" datasets, so it would be useful to explain why this is. For example, they found Principle 5, guideline (e) quite restrictive, or alternatively, encouraging of broad definitions of statistical or research purposes, going beyond what would normally be imposed for single source datasets.
6. New Zealand also suggested that actual levels of public concern in a country be monitored and understood, as this would be relevant in the business cases evaluating whether the integration work should go ahead. They see a huge value in data linkage to the community (e.g. the linkage of health datasets); and some such linkages may fall outside some definitions of official statistics, so it is worth stressing both increased utility and increased risk. They add that paragraph 10 rightly raises the point as to whether statistical purpose is enough, or whether public interest should be added. Official statistical purposes are by definition in the public interest, but other purposes may not be.
7. Poland considered that the document would need further improvement before endorsement by the Conference of European Statisticians, as it includes some formulations that raise concerns of public opinion, especially in the light of the organization of housing and population censuses in the near future.

8. Slovenia suggested combining some principles to improve clarity, and make them more focussed. They also expressed a softer approach based on suggestions and descriptions of good practices, rather than principles and guidelines, adding that some of the guidelines may be rather country-specific, and other country-specific guidelines were missing. They also asked about the updating mechanism for these principles and guidelines.

9. Turkey proposed to add an additional principle “Administrative and survey records for the privacy policy should have detailed description. Identity and address information in the administrative records should be kept separate from the statistical records. When statistical data are used the personal key information should be hidden.”

Response

10. Text on the need for these principles and guidelines will be added to paragraph 2: “***In some cases the use of integrated data sets can introduce additional legal and policy concerns compared to the use of single-source data sets. These additional concerns typically relate to, but are not necessarily limited to privacy and data protection requirements***”.

11. The structure of the principles has been refined several times, and has now been widely agreed, so it would be preferable to avoid further changes at this stage if possible. The Conference of European Statisticians will be able to revise and update these principles and guidelines in the future as it thinks fit.

III. COMMENTS ON DEFINITIONS

12. Concerning “Statistical activity”, Slovenia proposed to change the word “distribution” to “dissemination”, whilst Lithuania proposed “Statistical activity – the collection, storage, transformation of statistical data and distribution of statistical information”, adding that it is of the utmost importance to distinguish between statistical data and statistical information.

13. New Zealand considered the definition of “Statistical purposes” to be a bit circular, and suggested that statistical purpose is about finding results for a population (or part of), rather than individuals.

14. The Netherlands noted that paragraph 3 states that combining data sets from different countries is not considered data integration, however the definition of data integration in point 5c seems to contradict this, for new output is generated by combining data sets from different countries. They suggest either to accentuate the definition, or, if the definition is to be maintained, to adjust the text in paragraph 3 to: “*but as there are unlikely to be any units in common between the national data files, no confidentiality issues arise here*”, so as not to treat national and international statistical organisations differently with respect to data integration. Moldova questioned whether data integration includes the process of creating and updating of statistical registers, adding that if this is the case, it could cause difficulties. New Zealand also suggested that both “*data integration*” and “*data matching*” are about attaching a unit record from one data collection, to a unit record from another data collection.

15. Ireland said distinguishing statistical activities from research activities has always been problematic and indeed the document itself only defines statistical purposes. Their national legislation (the Statistics Act, 1993) and UN Fundamental Principle 6, refer only to "statistical purposes". Consideration should therefore be given to defining "research purposes" as distinct from "statistical purposes". Russia said that the phrase "*for statistical and research purposes*" should be replaced "*for official statistics and related analytical purposes*".

Response

16. This paper aims to follow existing international standards where possible. The definitions of "statistical activity", "statistical purposes", "data integration" and "data matching" were taken from the recently adopted 2009 version of the SDMX Content-oriented Guidelines. We would be reluctant to introduce different definitions.

17. The new wording proposed by the Netherlands for paragraph 3 will be used: "***but as there are unlikely to be any units in common between the national data files, no confidentiality issues arise here***". It will be made clearer in paragraph 2 that: "***these principles and guidelines, whilst having some relevance to the creation and maintenance of statistical registers, do not cover these tasks***". A definition of "Research purposes" will be added to paragraph 5: "***In the context of these principles and guidelines, "related research purposes" are defined as ad-hoc activities to investigate or explain economic or social phenomena, which result in statistical outputs. These activities may be undertaken by a statistical organization (in which case the results may not necessarily be published), or by external researchers (following the Conference of European Statisticians "Principles and guidelines on managing statistical confidentiality and microdata access").***"

IV. OTHER COMMENTS ON THE INTRODUCTION

18. The Netherlands noted in relation to paragraph 2, that data integration may lead to new statistics as well as to data sets for research purposes. These are different goals for different use and users; which might call for a distinction between the two with regard to regulating confidentiality aspects. They also noted that paragraph 7 states that these principles expand on Fundamental Principle 6, and suggested quoting this Fundamental Principle.

19. The Netherlands also noted in relation to paragraph 10, that the first notion is not entirely clear, and asked for an explanatory sentence. They noted that the second notion stresses the need to justify data integration for research purposes with respect to public interest, and emphasised the need to make proper agreements with data source owners regarding the use and confidentiality of the data, however, statistical organizations might not be able to influence the agreements between data source owners and data suppliers.

Response

20. Fundamental Principle 6 is already quoted in paragraph 1. Any difference in managing confidentiality according to purpose could be seen as deviating from this principle. Paragraph 10 will be re-drafted as follows: “***data integration must not occur when it will materially threaten the integrity of the source data collections, for example by posing a risk of reduced response rates***”.

V. COMMENTS ON PRINCIPLE 1

21. Poland said that the phrase "data integration should be undertaken by NSOs for statistical and research purposes and in circumstances in which release of data that could identify a natural person(s) (and, if legislated, a legal person(s) such as a business) is regulated by law”, is not compatible with the point 6 of “Fundamental Principles of Official Statistics”.

22. Both Hungary and New Zealand suggested that data integration might also be undertaken by other members of the national statistical system not only by NSOs. Hungary suggested to distinguish between the rules for matching using direct identifiers, and those for statistical matching. In the first case, regulation by law is important, but in the second one regulation by law can not be applied since it can not be controlled, because researchers are also allowed to make statistical matching on anonymized microdata.

23. Russia suggested that the phrase “data held in government departments” in guideline (c) should be replaced by “data held in government departments, public authorities and local governments, organizations”.

24. New Zealand suggested the following change “Data integration should ~~only~~ be undertaken by NSOs only for statistical and research purposes” to make this principle clearer. They added that the intent of this principle with respect to legislation is also unclear.

Response

25. We propose to clarify principle one by re-drafting it as follows: “***Data integration should be undertaken by NSOs (and other organizations within national statistical systems) only for statistical and related research purposes.***” Guideline (c) in paragraph 11 will be re-written: “***data held in national or sub-national government departments or public authorities***”.

VI. COMMENTS ON PRINCIPLE 2

26. The Netherlands noted that whereas the first line under principle 2 seems to call for restraint regarding data integration, guideline (b) strongly suggests that data integration is the alternative for new surveys. When a NSO has no authorization to integrate data (stated in guideline 12a) and a law to protect confidentiality (Principle 1), this part of the guideline can not be followed. They suggest re-phrasing this guideline: “Before undertaking a new survey for statistical purposes, consideration should be given to whether data integration of data sources already available at the NSO could be used as an alternative means”.

27. Poland expressed (unspecified) reservations about guideline (a), and suggested the document needs further work, offering to provide detailed remarks later.

28. Tajikistan noted that a new Statistical Law is currently under development, which will provide more detail about confidentiality. Guideline (c) requires using a form of “business case” for the approval of data integration projects. They ask whether the new Statistical Law should include such a requirement.

Response

29. The proposed re-wording from the Netherlands of guideline (b) in paragraph 12 is accepted: “***Before undertaking a new survey for statistical purposes, consideration should be given as to whether integration of data sources already available at the NSO could be used as an alternative means***”. The type of business case mentioned in guideline (c) is seen more as a procedural than a legal requirement, and as such, it should not be necessary to include it in statistical legislation.

VII. COMMENTS ON PRINCIPLE 3

30. New Zealand felt that this principle could be more explicit about confidentiality, and identified specific risks around confidentiality for data integration, such as respondents objecting to data collected for one purpose being integrated and used for another, and integrated datasets typically containing a larger number of variables than any of their sources. The heightened risks suggest heightened care is needed concerning who is able to access integrated data, but the document says nothing specific on what should be done about published results. A risk assessment would be useful, including a statement about risks to confidentiality, and how they are managed. This is needed in the Annex.

31. Poland expressed doubts that guideline (e), concerning the advisability of obtaining consent for data integration, is realistic in practice.

32. Russia noted that guideline (b), and other parts of the text, refer to direct identifiers. It is necessary to consider also indirect identification (e.g. aggregated data on a municipality, where only one or two companies have a particular type of activity). This requires further explanation of the terms “privacy” and “confidentiality”, subject to the provisions of Principle 8.

Response

33. An additional guideline is proposed in paragraph 13: “*(f) The notions of privacy and confidentiality also require careful management of the risks of indirect identification (typically for units with unusual characteristics), and the increased sensitivity of integrated data sets, which may contain a wider range of variables than any of their sources*”.

34. A comment about risk assessment will be added to the Annex as point D, the subsequent points will be re-numbered: “*D. Risk Assessment – The business case should include an assessment of the risks to confidentiality, risks to the integrity of the data sources, any other relevant risks, and a statement on how these risks will be managed*”. The first sentence of the current paragraph 7 of the Annex will be deleted, as it is now covered in the proposed addition above.

35. The recommendation to obtain consent “where reasonable and practicable” in guideline (e) is seen as best practice rather than an absolute requirement, so we propose to retain it.

VIII. COMMENTS ON PRINCIPLE 4

36. Ireland noted that national legislation can facilitate data integration even for sources where an explicit indication has been given to respondents that the data will only be used for the purposes for which it has been collected (e.g. The Statistics Act, 1993). Clearly an NSO should always consider the ethical and data implications of undertaking data integration involving such sources but perhaps the principle might benefit from a reference to this type of circumstance.

Response

37. These principles and guidelines are not mandatory, so would not prevent integration in the scenario mentioned, however, we would strongly recommend that this principle is not breached if at all possible, as this could lead to a loss of public trust.

IX. COMMENTS ON PRINCIPLE 5

38. Ireland noted that guideline (d) implies that where the method of integration changes then a new approval for the data integration should be prepared. The how (technical methodology for data integration) shouldn't really impact on the business case/decision for sanctioning the data integration activity unless a change in the data integration methodology has the potential to significantly change the risk of disclosure for a natural/legal person.

39. The Netherlands stated that the words ‘significant variation’ do not immediately make clear what is a small or a large change, and asked for an example in guideline (a).

40. New Zealand asked that if the primary statistical purpose of data integration carries sufficient weight, why should further approval processes be imposed for other purposes, as long as those purposes fit within the definition of “statistical or related research purposes”.

Response

41. Guidelines (a) and (d) in paragraph 15 will be re-written in line with the comments: “(a) *The sources (data sets) used in the integration process change significantly (for example a category of units is added or deleted, or there is a change in the type of variables covered), or a new source is proposed to be added to the integration process*”; “(d) *The method of integration changes (e.g. from statistical to exact matching), and this change could significantly alter the risk of disclosure for a natural/legal person*”.

42. Separate approval is recommended for the use of integrated data for additional purposes, as these purposes might entail a different degree of risk, or may deliver different levels of benefits to official statistics.

X. COMMENTS ON PRINCIPLE 7

43. Hungary noted that sometimes results of the integration work are used only for validation, and quality check purposes. They suggest to re-word guideline (b) “Metadata of statistics published from composite databases should contain information about the original data sources used for data integration.”

44. Moldova stated that it is important, along with the necessity of assuring the confidentiality of data, to establish the procedures which would allow NSOs to fulfill their duties efficiently, if possible without the bureaucratization of the process. In this respect they propose to remove the requirement in guideline (c) to inform respondents (New Zealand also saw this requirement as difficult to meet in practice). Moldova also noted that rules for integration of data on legal persons could be less restrictive than those for natural persons. Legislation in Moldova foresees special rules for the protection of personal data of natural persons.

45. The Netherlands supported the view in guideline (b) that results of data integration work should be made publicly available. However, in some cases, research work may fail to achieve its goal, and thus will not be published. To remove the risk that this guideline may be used to prevent research on integrated data sources, they suggested adding “the results of successful research”.

Response

46. The suggested re-wording from Hungary will be added to guideline (b) in paragraph 17: “*Metadata of statistics published from composite databases should contain information about the original data sources used for data integration*”.

47. The recommendation to inform respondents “wherever reasonable and practicable” in guideline (c) is seen as best practice rather than an absolute requirement, so we propose to retain it.

48. The point raised by the Netherlands has been incorporated in the new definition of “Research purposes”. It is recommended that if external researchers are given privileged access to an integrated dataset at least some results of their research should be published. This is essential to maintain transparency in how composite micro datasets are being used.

XI. COMMENTS ON PRINCIPLE 8

49. The Netherlands noted that confidentiality of microdata has to be guaranteed, whether data are obtained by integration or otherwise. They suggested making a general reference to confidentiality rules and regulations.

50. New Zealand asked whether there is a possibility for Confidentialised Unit Record File (CURF) versions of microdata for licensed users, as this principle seems a little restrictive at present.

51. Poland expressed (unspecified) reservations about principle 8, and suggested that the document needs further work, offering to provide detailed remarks later.

52. Russia suggested that the expression “consistent with the purposes set out in the standard approval process” should be replaced by “consistent with the purposes of use of data for official statistics”.

53. Turkey said that the phrase “*External bona fide researchers*” should be clearly defined, and validation criteria proposed.

Response

54. The text “*As for other statistical microdata*” will be added at the start of the second sentence of principle eight. The question from New Zealand seems also to be covered by this sentence in that if CURF versions have a clear legal basis, they can be provided. The proposed re-wording from Russia will be added at the end of the same sentence: “*consistent with the purposes of use of data for official statistics*”.

55. It is proposed to remove the words “bona fide” from guideline (c). The access to microdata for research purposes is covered in the Conference of European Statisticians principles and guidelines on “Managing Statistical Confidentiality and Microdata Access”.

XII. COMMENTS ON THE ANNEX

56. The Netherlands suggested adding “reduction of costs or response burden” to the list of improvements in paragraph 3.

Response

57. This point will be added to paragraph 3: “*reduction of costs or response burden*”.

XIII. CONCLUSION AND RECOMMENDATIONS

58. These principles and guidelines have been through several rounds of review by members of the Task Force on the Confidentiality and Privacy Aspects of Statistical Data Integration, national experts and the Bureau of the Conference of European Statisticians. Most of the comments received in this latest round of comments can be addressed with relatively minor changes and additional explanations. The most substantial change proposed is to principle 1, which has been simplified and brought closer in line with the Fundamental Principles of Official Statistics.

51. The chair of the Task Force on Confidentiality and Privacy Aspects of Statistical Data Integration, and the UNECE Secretariat recommend acceptance of the Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, subject to the modifications proposed above.