UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

Principles and Guidelines

on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes

UNITED NATIONS

Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes

These Principles and Guidelines were endorsed by the Conference of European Statisticians at their June 2009 meeting.

INTRODUCTION

1. At its eighth meeting, the United Nations Economic Commission for Europe (UNECE) adopted its decision C(47) on the fundamental principles of official statistics in the region of the Economic Commission for Europe (15 April 1992 see http://www.unece.org/stats/documents/e/1992/32.e.pdf). Fundamental Principle number six states that "Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes."

2. Data integration is concerned with integrating unit record data from different administrative and/or survey sources to compile new official statistics which can then be released in their own right. In addition, these integrated data sets may be used to support a range of economic and social research not possible using traditional sources. 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. These principles and guidelines, whilst having some relevance to the creation and maintenance of statistical registers, do not cover these tasks.

3. These principles and guidelines apply to data integration work carried out in national statistical organizations (NSOs). In some cases international statistical organisations combine micro-data sets from different countries, but as there are unlikely to be any units in common between the national data files, no confidentiality issues arise here.

4. For the purpose of these principles and guidelines, the use of other sources in validation and imputation processes for a single source, is not considered as data integration, though similar issues may apply. Two instances of the same survey are considered to be a single source.

5. To assist in the use of these Principles and Guidelines the following definitions are used:

(a) Composite microdata - unit record data resulting from data integration;

(b) Confidentiality - an obligation to the provider of information to maintain the secrecy of that information;

(c) Data Integration - the process of combining data from two or more sources to produce new outputs;

(d) Data Matching - the linkage of micro-data from different sources based on common features present in those sources;

(e) Data Provider - An organization which produces data or metadata. For the purposes of these principles and guidelines, this term includes providers of data files from statistical or non-statistical sources, but not individual respondents to statistical surveys;

(f) Natural or legal persons - individuals and legal entities recognised by national legislation;

(g) Official Statistics - any statistical activity carried out within a national statistical system, or under the statistical programme of an intergovernmental organization;

(h) Privacy - someone's right to keep their personal matters and relationships secret, involving an obligation of the holder of information to the subject of the information to do so;

(i) Research Purposes - 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”).

(j) Statistical Activity - the collection, storage, transformation and distribution of statistical information;

(k) Statistical Purposes - the use of data in a way that complies with the Fundamental Principles of Official Statistics, fits into one or more phases of the statistical business process, and contributes to the production of official statistics.

6. Integration of data may include exact matching, probabilistic matching and/or statistical matching. The benefits of integrated data sets can include:

(a) Production of new or enhanced statistics;

(b) Production of more disaggregated information for measures where some information currently exists;

(c) Ability to carry out research using composite microdata that cover a wider range of variables for a larger number of units than available from any single data source;

(d) Potential to improve or validate existing data sources;

(e) Potential to reduce respondent burden.

7. The attached principles and associated guidelines expand on Fundamental Principle six by providing a common framework for assessing and mitigating legislative and other confidentiality aspects of the creation and use of integrated datasets for statistical and research purposes. In particular they recognise that the

fundamental principles of official statistics apply equally to integrated data sets as to any other source of official statistics.

8. In developing these principles, it is recognised that it is government policy in some countries to first seek to use (integrated) administrative data sources, such as registers, for the production of official statistics, only conducting surveys if crucial variables are missing or are of too low a quality in the available administrative sources. In these countries, integration of statistical data sets is a normal part of the operations of the national statistical office. These countries usually already have a strong framework of legislation and clear rules about protection of the confidentiality of personal and individual business data, irrespective of whether or not the data has been integrated from different sources.

9. However, for many other countries the notion of integrating data to produce composite microdata from different sources for statistical and related research purposes is relatively new. The attached principles, associated guidelines, and the example of a business case outline, are designed to provide a framework for such work that can provide some clarity and consistency of application.

10. In addition to the principles and guidelines below, two other notions have been identified that may be relevant in certain circumstances, particularly for countries that do not have a strong tradition of data integration activities for official statistics purposes. The first is that 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. The second is that data integration for research purposes should only be considered where the approval process can justify that this is in the public interest. It is assumed that official statistical purposes will always serve the public interest if they comply with the fundamental principles of official statistics. These notions are mentioned here for completeness, but are not as widely accepted, and therefore do not have the same status as the principles below.

PRINCIPLES AND GUIDELINES

Principle 1

Data integration should be undertaken by NSOs (and other organizations within national statistical systems) only for statistical and related research purposes.

Guidelines:

(a) The above principle should be enshrined in either the statistics legislation and/or in the legislation on data protection, and be strictly respected by governments;

(b) In circumstances where no explicit legislative protection exists, NSOs should abstain from data integration concerning natural and legal persons;

(c) Unless national legislation stipulates otherwise, the use for statistical or research purposes by NSOs of any existing data held in national or sub-national government departments or public authorities, from administrative or statistical sources, does not contravene the privacy of a specific natural or legal person.

Principle 2

NSOs should only undertake data integration activities consistent with their official statistics mandate and after completing a standard approval process (for example, a business case).

Guidelines:

(a) Where a NSO has a mandate(s) that goes beyond statistical and related research purposes, such as involving the use of data for administrative or regulatory purposes related to natural persons, it should abstain from any data integration activities for statistical or related research purposes pertaining to these units, unless this is specifically authorized by law;

(b) 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;

(c) A standard approval process should be followed for any new data integration proposal. This may take the form of a formal business case. An example of a business case outline is given in the Annex, but each country should establish their own template for the process of endorsing data integration projects. The approval process should identify how the integration work will produce or improve official statistics or contribute to related research.

Principle 3

The public benefits of any data integration project should be sufficient to outweigh any privacy or confidentiality concerns about the use of data and/or risks to the integrity of the official statistics system.

Guidelines:

(a) Data integration should occur in a secure environment and in a manner that does not pose risks to the integrity of the official statistical system;

(b) Unless enabled by legislation or provided for in the standard approval process, any direct identifiers associated with the data to be integrated should be removed as soon as possible upon completion of the integration process;

(c) Where appropriate, bodies with responsibility for ensuring that all benefits, privacy concerns and risks are identified and properly considered by the NSO as part of their standard approval process, should be consulted. The list of benefits should include those resulting from any intended long term retention of, or planned extension over time to, the integrated dataset;

(d) In some countries, it will be a legislative requirement that the standard approval process should include a privacy impact assessment;

(e) Where reasonable and practicable, consent should be obtained from the data provider(s);

(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.

Principle 4

Data should not be integrated where any commitment has been given to respondents that would specifically preclude such action.

Guidelines:

(a) The standard approval process (such as a data integration business case) should investigate what undertakings have been made to respondents regarding the purposes for which their data can be used. The Head of the NSO should not approve a data integration proposal if any element of that proposal is incompatible with these undertakings.

Principle 5

Integrated data should only be used for approved statistical or research purposes and any significant variation in the originally approved purposes should result in the submission of a new standard approval process.

Guidelines:

Unless otherwise provided for in legislation, a new approval should be sought whenever:

(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;

(b) The number of characteristics covered by integration is proposed to be expanded significantly;

(c) The number of units covered by the integration process is proposed to be expanded significantly (e.g. extension from a few to all branches of the economy);

(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;

(e) The dataset resulting from integration is proposed to be used for another official statistical purpose or for a research purpose submitted by an outside researcher, which had not been provided for in the original standard approval process.

Principle 6

The number of unit records and data variables to be included in a linked dataset should be no more than required to support the approved purpose(s).

Guidelines:

(a) The 'approved purpose(s)' is that which is approved in the data integration business case. Only data variables necessary to support these purposes should be included in the dataset for the approved data integration work;

(b) The number of unit records to be integrated should be the minimum necessary to support the approved purpose(s) (e.g. consideration should be given to integrating a sample of a full-coverage data source).

Principle 7

NSOs should conduct any data integration in an open and transparent manner.

Guidelines:

(a) The policy of the NSO with respect to data integration, as well as an overview of data integration work being undertaken by an NSO should be published;

(b) The main statistical results of any data integration work should be made publicly available. When data integration work is used to improve the production of official statistics (e.g. through improving quality), the publication of that official statistic meets this requirement. Metadata of statistics published from composite databases should contain information about the original data sources used for data integration;

(c) Unless otherwise enabled by legislation, administrative agencies should, wherever reasonable and practicable, inform respondents that their information may be generally used for statistical or research purposes.

Principle 8

Access to composite unit record data resulting from data integration, but not containing any identifiers, should generally be limited to authorized staff of the NSO. As for other statistical microdata, any proposal to grant access to an external person(s) should have a clear legal basis and be consistent with the purposes of use of data for official statistics. Any person(s) granted such access should provide a legally enforceable institutional and logistical guarantee that their use will be consistent with the approved proposal and that non-authorized persons will have no access to the dataset.

Guidelines:

(a) Composite microdata from integration carried out by an NSO may be used for statistical or related research purposes by other producers of official statistics in the same national statistical system, or if supported by appropriate national legislation, in a supra-national statistical system, providing a business case following guideline 2(b) is approved by the NSO. Approval should include consideration as to whether statistical and related research activity is strictly

separated in organizational terms from any data collection or processing for administrative purposes;

(b) An NSO should not provide information to data providers about any variables in an integrated data that could assist the data provider in carrying out any administrative or regulatory purpose;

(c) External researchers can be granted access to microdata from integrated datasets following the Conference of European Statisticians guidelines “Managing Statistical Confidentiality and Microdata Access”, if a business case is approved by the NSO;

(d) The obligations of the recipient should be laid down in a contract and any infringement of the confidentiality rules by the recipient should fall under a potential sanction stipulated in legislation and be enforceable on the recipient, and where appropriate the sponsoring institution.

ANNEX - EXAMPLE OF A BUSINESS CASE OUTLINE

Principle two provides that a standard approval process should be followed for any new data integration proposal and suggests that this may take the form of a formal business case. If a business case approach is to be used, it is proposed that it should cover the following topic areas:

A. Purpose(s)

The business case should describe the purposes for which the integrated data will be used.

B. Benefit to Official Statistics

The business case should describe how the proposed project will produce or improve official statistics. Improving official statistics could involve improvements to the accuracy, reliability, relevance, timeliness, consistency and coverage of the statistics, the concepts, definitions or methods used to produce the statistics or to reduction of costs or response burden.

C. Other Benefits

The business case should describe who else will benefit, and how they will do so, from the project.

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

E. Retention

The business case should state how long the integrated dataset needs to be retained to support the purposes for use. The retention may be subject to periodic reviews.

F. Data sources

The proposal should describe what sources of data will be used for data integration. This should list the proposed source agencies and describe in general terms the data to be received from each agency.

Any implications of the legislation that source data was collected under to the data integration project should be listed.

G. Alternatives

It should be explained why data integration is preferable to any feasible alternative in terms of cost, quality or minimising compliance burden.

H. Stakeholders

The business case should list all identified key stakeholders (both internal and external) in the data integration project and the results of any consultation with them.

I. Retention of names and addresses

If the data integration project needs to retain personal names and addresses for linking then this should be stated along with how long retention of these details is needed.

J. Frequency of Reviews

The business case should specify the frequency with which reviews of the data integration will be held.

K. Privacy Impact Assessment

A privacy impact assessment should be completed unless a country's legislative and/or relevant NSO policy provides an exemption. It should also be noted that although privacy generally relates to natural persons, it may also relate to a legal person in the case of some businesses or industries. For example, certain unincorporated businesses, such as farms, may generate privacy considerations in some countries.

UNECE, Geneva, 2009