**Economic Commission for Europe**

**Conference of European Statisticians**

**Fifty-ninth plenary session**

Geneva, 14-16 June 2011

Item 7a of the provisional agenda

**Outcome of the in-depth review of selected statistical areas**

**Use of secondary and mixed sources for official statistics: in-depth review by the Bureau of the Conference of European Statisticians**

**Note by the secretariat**

**Summary**

The present note is a revised version of the document that was used as the basis for an in-depth review by the Bureau of the Conference of European Statisticians during 2010. Each year the Bureau reviews selected statistical areas in depth, with the aim of improving the coordination of statistical activities in the region of the United Nations Economic Commission for Europe. The review focuses on strategic issues and highlights conceptual and organisational concerns of statistical organisations.

The document has been updated to take into account the comments made in the Bureau discussions. It also provides information on the outcome of the review and the follow-up actions undertaken.
I. Background

1. The Bureau of the Conference of European Statisticians (CES) regularly conducts in-depth reviews of selected statistical areas. The aim of these reviews is to give an overview of international statistical work in the specified area, identify challenges (such as lack of activities, duplication of work, uncoordinated activities, etc.), and find ways to overcome them.

2. In October 2009, the Bureau selected the use of secondary and mixed sources for official statistics as a topic for an-in-depth review. The UNECE secretariat was asked to prepare a note on this topic to serve as the basis for the in-depth review. This note was presented and discussed at the Bureau meetings in June and November 2010. The note was revised as a result of feedback received, and is now presented to the Conference of European Statisticians.

II. Definitions and scope of the review

3. Data sources for official statistics can be split into two broad categories: data collected primarily for statistical purposes (primary sources), and data collected primarily for non-statistical purposes (secondary sources). Statistical data production that makes use of both types of sources is referred to as “mixed source”.

4. Secondary sources include data from public authorities, often referred to as administrative data, as well as data from private sector and research organisations. This definition is deliberately broad, as in some cases regulatory functions that used to be carried out by government bodies have been transferred to private or semi-private organisations. Typical examples can be found in the health, education or public utilities sectors, where former state monopolies are increasingly being replaced by private companies or non-profit institutions. Secondary sources also include data from the accounting systems of businesses.

5. At the same time, the market for data is rapidly growing, with increasing numbers of private-sector data suppliers emerging to meet the growing demand. The private sector data supply industry started with the development and sale of address lists for marketing purposes, then expanded to cover the provision of credit rating data and business intelligence information, and has now spread to cover virtually all types of data. The private sector has started to realise that data are a very valuable commodity.

6. A relatively recent development has been the emergence of private sector “value-added re-sellers” in the data market. These businesses take existing data from a variety of public and private sector sources, combine them, clean them, and sometimes validate them, and then re-sell them to other organisations. Examples include business data sellers such as Dun and Bradstreet, Bureau van Dijk and Hoppenstedt.

7. This sort of data source can be of interest to official statisticians, as sometimes these private sector data suppliers can actually process and supply data more cheaply than statistical organisations, often simply because they can spread the costs amongst a number of customers. The Eurostat "Eurogroups" project to develop an international statistical register of enterprise groups uses such sources for exactly this reason.

8. As discussed in the previous paragraphs, the potential range of secondary data sources that can be used for statistical purposes is large and growing. The following list is not meant to be exhaustive. Instead it aims to show the range and types of secondary data sources that can be used for official statistics:
(a) Tax data:
   (i) Personal income tax;
   (ii) Sales / Value Added Tax;
   (iii) Business / profits tax;
   (iv) Property tax;
   (v) Import / export duties;
   (vi) Excise duties;
   (vii) Environmental tax;
(b) Social security data (contributions / benefits / pensions);
(c) Health / education records;
(d) Registration systems for persons / businesses / farms / property / vehicles;
(e) Regulatory bodies for specific activities;
(f) Identity cards / passports / driving licenses;
(g) Electoral registers;
(h) Building permits;
(i) Licensing systems e.g. for televisions, or sales of restricted goods;
(j) Data from private or academic research projects;
(k) Published business accounts;
(l) Accounting data held by businesses;
(m) Commercial data from private businesses:
   (i) Credit information;
   (ii) Business analysis;
   (iii) Utility company records;
   (iv) Telephone directories;
   (v) Store card / till roll / scanner data etc. held by retailers.

III. Activities relating to the use of secondary and mixed sources for official statistics

9. There is currently no specific on-going international expert group dealing with the use of secondary and mixed sources across all domains of official statistics. Instead this topic tends to be treated either as part of the work of domain-specific groups, or through ad-hoc projects, seminars and conferences. There is, therefore, a lack of coherence and continuity of statistical work on this topic.

10. The current situation is highlighted in the Database of International Statistical Activities¹, where activities for 2010 reported under category 4.3.5 “Other administrative

¹ http://www1.unece.org/stat/platform/display/disaarchive
and non-survey sources” comprised just a short list of rather specific activities in the domains of health, financial reporting and vital statistics. However, on closer examination, this apparent lack of activity is misleading, as there are many other references to the use of secondary sources under other categories of the DISA classification.

11. Information presented at the 2010 Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses showed that, for the first time, a majority of European Union countries will use secondary or mixed sources for the 2010 round of population censuses. Other major initiatives demonstrating that there is a strong interest in further exploring the use of secondary and mixed sources include:

(a) The European Union “ESSnet” project on the use of administrative and accounts data in the production of business statistics;

(b) The “Eurogroups” project mentioned in the previous section;

(c) The development of projects and tools for data matching and integration in several statistical organisations;

(d) Increasing focus on the quality of data derived from secondary and mixed sources, for example in the work of the UN Expert Group on National Quality Assurance Frameworks and the Eurostat Quality Working Group.

12. This interest is further confirmed by the long list of ad-hoc events and conferences fully or partially devoted to the use of secondary and mixed sources, of which the following are a few examples:

(a) Statistics Canada 2010 International Methodology Symposium “Social Statistics: The Interplay among Censuses, Surveys and Administrative Data”, Ottawa, October 2010;

(b) Various sessions during the Q2010 Conference on Quality in Official Statistics, Helsinki, May 2010 (and previous conferences in this series);

(c) Istat Seminar on Using Administrative Data in the Production of Business Statistics - Member States Experiences, Rome, March 2010 (a follow-up meeting is planned for later 2010 in Vilnius);

(d) Various invited and contributed paper meetings during the 57th Session of the International Statistical Institute, Durban, August 2009 (and previous ISI sessions);

(e) The IAOS Conference on Reshaping Official Statistics, Shanghai, October 2008;

(f) Seminar on the Use of Administrative Data Sources (side-event to the 39th session of the UN Statistical Commission), February 2008;

(g) Statistics Finland Seminar on Registers in Statistics - Methodology and Quality, Helsinki, May 2007;

(h) United Nations Statistical Division Regional Workshop on the Use of Administrative Data in Economic Statistics, Moscow, November 2006;


(j) CES work sessions on registers and administrative records in social and demographic statistics, several sessions up to 2002;

13. The use of secondary and mixed sources seems to be gaining more attention as statistical organisations increasingly consider a more process-oriented approach to statistical production. This is closely tied to the need to make statistical production more efficient and decrease the burden on respondents. However, as much of this work is done in domain-specific “silos”, or within an organisation-specific context, the scope for cross-fertilisation of ideas is rather limited at present.

IV. Other resources relating to the use of secondary and mixed sources for official statistics

14. The conferences and events listed in the previous section have generated quite a large volume of papers, mostly describing national practices. In most cases, these papers are available on meeting web sites, but there is currently no single entry point or means of finding and categorising such papers, so those searching for information on this topic do not always find what they are looking for.

15. In addition to these papers, several publications and other resources exist on the topic of the use of secondary and mixed sources for official statistics. These include:


(b) Using Administrative Sources for Official Statistics - A Handbook of Principles and Practices, UNECE, draft distributed to participants of the UNECE / Statistics Finland training courses on the use of administrative data since 2007;

(c) Register-based Statistics: Administrative Data for Statistical Purposes, Anders and Britt Wallgren, 2007;

(d) Use of Registers and Administrative Data Sources for Statistical Purposes – Best Practices of Statistics Finland, Statistics Finland, 2004;

(e) Quality Assessment of Administrative Data for Statistical Purposes, Eurostat, 2003


(g) Use of Administrative Sources for Statistical Purposes - Seminar Proceedings, Eurostat, 1997;


16. Other publications not directly concerned with the use of secondary or mixed sources, can nevertheless contain useful information on specific issues, for example “Methods for Automatic Record Matching and Linking and their use in National Statistics”, United Kingdom Office for National Statistics, 2001.

V. Issues and challenges relating to the use of secondary and mixed sources for official statistics

17. The use of secondary and mixed sources for official statistics creates a number of challenges, which can be grouped according to the following types, and are examined in more detail in the following paragraphs:
(a) Legal issues – related to the access to and the use of data;
(b) Public relations issues – managing public trust and acceptance of data linkage and re-use, and maintaining the profile of the statistical organisation;
(c) Financial issues – private sector sources are rarely free of charge, and there may also be costs relating to the use of public sector data;
(d) Management issues – re-designing the statistical production process and reallocating resources;
(e) Methodological issues – data processing methods should aim to achieve the optimal level of quality with the resources available, and explain any changes in quality due to changes in data sources.

18. In general, the legal framework for the use of secondary and mixed data sources is country-specific. The main restrictions are usually on access to and use of public sector administrative data, and there may also be restrictions on data linkage. However, in the case of the European Union, there is also a European dimension to the legal framework, providing a limited degree of harmonisation. Article 24 of the new “Statistical Law” states that:

“In order to reduce the burden on respondents, the NSIs and other national authorities and the Commission (Eurostat) shall have access to administrative data sources, from within their respective public administrative system, to the extent that these data are necessary for the development, production and dissemination of European statistics.”

19. However, this broad and encouraging statement is then qualified to some extent by the following paragraph:

“The practical arrangements and the conditions for achieving effective access shall be determined where necessary by each Member State and the Commission, within their respective spheres of competence.”

20. Whilst not providing a legal framework, the Fundamental Principles of Official Statistics give useful guidance on this topic. Principle 5 can be used as justification for improved access to secondary data sources:

“Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.”

21. However, even where a strong legal framework exists, there are often difficulties in using it in practice. Those responsible for public sector secondary sources may see the supply of data to statistical organisations as an extra burden with no benefit to them. They may not understand or accept the principles of statistical confidentiality and the one-way flow of data. Legal frameworks therefore need to be backed-up by government policy and inter-organisational agreements.

22. Often closely related to legal rights and responsibilities is the issue of public acceptance of data sharing, linking and re-use, particularly between government agencies. Public opinion varies from positive support, and even encouragement in some countries, to hostility and mistrust in others. The benefits and efficiencies of using secondary and mixed

---

data sources are weighed against fears that governments are “spying” on their subjects. Statistical organisations therefore need to find ways to build and maintain public trust if they wish to extend the use of secondary sources. More pro-active “marketing” of the positive impacts (particularly the cost and response burden reductions) of using these sources may also help to influence public opinion.

23. Another side of the public relations issue is that when statistical organisations collect data through statistical surveys, they are interacting directly with the public. If they switch to secondary sources, that contact may be lost, and the statistical organisation may need to invest more in retaining its public profile.

24. There may be a financial cost associated with using a secondary source. Private sector data suppliers are, by definition, seeking to make a profit, and public sector suppliers may require the statistical organisation to at least cover the costs of extracting and transferring the data. There will often also be costs associated with ensuring the data are fit for purpose, and making any necessary transformations. These costs have to be carefully weighed against the benefits, and it is often necessary to accept a transition period during which costs increase as systems are adapted, before they start to decrease as the benefits are realised.

25. Any change in statistical data sources will have an impact on how the statistical production process is managed. When multiple data sources are used, it is often more appropriate to think in terms of data requirements and outputs, rather than traditional surveys. This means taking a broader view of data availability and use, and ensuring re-use wherever possible. The logical conclusion is then to only use statistical surveys as a last resort, to collect data items not available elsewhere. This can lead to the possibility of questionnaires that are tailored to the specific data requirements from each respondent, rather than standard survey questionnaires for all units.

26. The move to the use of secondary and mixed sources may also increase the attractiveness of a more process-oriented approach to statistical production, grouping activities by function rather than statistical domain. Several national statistical organisations are moving in this direction, particularly with regards to consolidating data collection and acquisition in a single organisational unit. The resource implications of such a strategy can be significant, but so are the efficiency savings and quality improvements claimed by some organisations that have moved in this direction. In many cases, a transition period is needed, during which data are collected and processed according to both the old and new models, so that the impacts can be properly assessed.

27. The use of secondary and mixed sources, however, introduces a certain dependency on supplying organisations. This introduces new risks related to changes in the scope, definitions and timeliness of the source data, which require careful management, for example through contingency planning and systems to manage and coordinate different sources.

28. Another key management issue often faced is resistance to change. Statisticians generally only trust data that they have collected themselves. They may have an inflated opinion of the quality of the data they have collected, often trusting that respondents read their pages of notes and definitions, and diligently apply them when completing survey forms. Any differences between results derived from primary and secondary sources are often seen as evidence of the inadequacies of the latter. All of this can make it difficult to convince subject-matter statisticians that secondary and mixed sources have their place in the statistical production process.

29. There are many methodological issues associated with the use of secondary and mixed data sources, and particularly when considering changing from a primary source. The types of issues faced will vary according to the statistical domain, the specific source,
and various other factors. A detailed description of these issues, and possible solutions, is clearly beyond the scope of this paper, however, the following non-exhaustive list includes some of the most common methodological issues:

(a) Units – the units in secondary data sources do not always comply with statistical unit definitions. A common example is that secondary sources of business data often use the legal unit, which does not necessarily fit with the statistical definition of the enterprise;

(b) Definitions of variables – the variables in secondary data sources are often defined according to administrative or policy-related requirements, rather than statistical standards. For example employment status in secondary sources rarely complies perfectly with the International Labour Organisation definition, and business accounting definitions do not always match statistical definitions;

(c) Classification systems – as for definitions, classification systems in secondary sources tend to be designed and applied in different ways depending on the main purpose of the source;

(d) Timeliness – can be both an advantage and a disadvantage of secondary sources, depending largely on the frequency of the statistical output. Modelling and estimation techniques may sometimes need to be applied to secondary sources to get reliable short-period estimates;

(e) Inconsistency between sources – when mixed sources are used, it is almost inevitable that there will be inconsistencies, either for the same variable, or across logically related variables. In-depth knowledge of sources can often help to resolve conflicts through the application of automatic priority algorithms, and data confrontation techniques can also be useful;

(f) Missing data – this problem is not unique to secondary sources. Often techniques to deal with unit and item non-response in survey data can be successfully adapted to secondary sources. However the patterns of, and reasons for, missing data may be different, and could alter biases in the data. The impact of missing data should therefore be considered carefully;

(g) Quality – it is important to remember that quality is a multi-faceted concept, and that different sources will have different quality profiles. Data from secondary sources could, for example, help to improve timeliness, but might reduce comparability or accuracy. Focussing on just one of these dimensions will give an unbalanced view of the relative quality of sources, so it is important that an holistic approach to quality measurement is used when comparing sources. The reliability of secondary sources can be a serious issue in some countries or statistical domains, however the relative reliability of primary sources should also be considered. It is necessary to develop an in-depth understanding of both primary and secondary sources to be able to accurately judge their quality and reliability, taking into account any possible incentives for misreporting.

30. The previous paragraphs perhaps give a rather gloomy view of the potential to use secondary sources. However, it should be noted that in many cases, statistical organisations have found creative solutions to solve, or at least mitigate these problems, highlighting the importance of mechanisms to share ideas and experiences in this field. It is probably also true to say that the main barriers to the use of secondary and mixed sources are now more often legal and managerial rather than methodological.

31. One of the best ways to manage the issues above is to try to maintain very close cooperation with the providers of secondary data sources, to understand their constraints, and to help them understand statistical requirements. This often requires a very pro-active approach by the statistical organisation, particularly when starting to use a new source.
32. The final point for this section is that it should be remembered that despite all of the issues, the benefits of using secondary and mixed sources are often much greater than the costs.

VI. Conclusions of the in-depth review paper

33. There is no escaping the fact that secondary and mixed data sources will play an increasingly important role in the production of official statistics in the future. This will be facilitated by technological and methodological advances, and driven by cost and response burden considerations. Distinctions will increasingly be made between data sources that meet or do not meet the requirements for official statistics outputs, rather than between official and non-official sources. Therefore developing methods and metadata to help judge the value of data from different sources, as well as actions to improve source quality and fitness for purpose, will become priorities.

34. In countries that still conduct traditional statistical surveys and censuses, the costs of such operations are becoming an increasing issue with every iteration. Political pressure to reduce costs, combined with the growing availability of private-sector, mixed source data sets should not be ignored. The potential for private sector data providers to develop outputs in competition with official statistics (for example Google price indices) is likely to increase.

35. The official statistics community is, of course, already starting to respond to these challenges, and a lot of useful and interesting work is going on related to the use of secondary and mixed sources in different statistical domains. However, a key theme highlighted in this paper is that there is little coordination of this work. Whilst statistical production is increasingly moving from domain-based (or “stove-pipe”) to process-based models, international statistical coordination is perhaps lagging behind. Topics such as the use of secondary and mixed sources, which cut across statistical domains, are becoming more important, and deserving of greater attention.

VII. Summary of the discussions at the CES Bureau meetings

36. The Bureau discussed the use of secondary and mixed sources for official statistics in June and November 2010. These discussions were based on a paper by the UNECE secretariat, and additional notes by Slovenia and the Statistical Commission Friends of the Chair Group on Integrated Economic Statistics. The final paper (reproduced above) has benefited greatly from inputs from these discussions.

37. In addition to specific comments relating to the paper, Bureau members made the following general points during the discussions:

   (a) The topic is very actual and important. There are many common challenges for statistical organisations in this area;

   (b) Even though practices and legal arrangements differ between countries, it would be useful to discuss common strategic and management level issues in using secondary sources at the international level. For example, managing relations with other agencies, responsibilities for data maintenance and quality, data sharing, efficiency gains, etc.;

   (c) Guidelines on using secondary sources could be helpful for developing countries where there are problems with the low quality and reliability of these sources;
(d) International guidelines could be useful for defending the position of a statistical organization in accordance with the Fundamental Principles of Official Statistics. For example to counter requests to provide statistical microdata for administrative purposes, clarifying responsibilities for data maintenance, etc.;

(e) There are specific issues in this area at the national level but it is not clear whether international coordination would provide an added value;

(f) A knowledge base providing an overview of the topic would be necessary to clarify whether there is a need for any further work in this area at international level;

(g) The knowledge base could list the managerial level problems in this area and complement these with good practice examples from countries; the first step could be to identify the main issues;

(h) An informal group/network could be helpful in bringing together best practices in this area;

(i) The Statistical Commission Friends of the Chair Group on Integrated Economic Statistics is working on a handbook that deals with some issues related to the use of secondary sources in economic statistics. However, after discussing with the chair of that group, the Bureau noted that there is no overlap with this work, as the problems and issues considered are different;

(j) The 2011 CES seminar on organization of data collection and data sharing could provide some information on best practices of data use from secondary sources;

(k) The main issue is not to distinguish between official and non-official data, but between data that meet and data that do not meet the quality standards of official statistics. It is also necessary to have enough information to be able to judge about the data quality. The official statistical system can help to improve the quality of the non-official statistics, e.g. by providing quality guidelines. Official statisticians could establish a dialogue with the producers of statistics outside the system to improve the quality of their data.

38. The Bureau identified the following main conclusions from the discussions:

(a) The secretariat will create and maintain a knowledge base comprising the already existing information and containing links to relevant training materials, papers, conferences and publications. The knowledge base will take the form of a wiki on the UNECE Statistical Division’s website, and should have a logical structure to facilitate information exchange;

(b) For the time being, no group should be created to coordinate international work on the use of secondary and mixed sources;

(c) The secretariat will develop the existing training materials on “Using administrative sources for official statistics”, taking into account comments from Bureau members.

VIII. Follow-up action

39. The in-depth review paper has been updated to reflect the comments and conclusions of the Bureau.

40. The UNECE secretariat has created a knowledge base in the form of a public wiki. This knowledge base is called ASSIST (Administrative and Secondary Sources In STatistics), and contains links to relevant training materials, papers, conferences and
publications\(^3\). For ease of use, papers are classified by statistical domain using the list of domains from the Database of International Statistical Activities. The wiki approach allows the possibility of editors from other organisations, as well as comments and discussion forums. This knowledge base should be viewed as “work in progress” rather than a finished product, as there will always be further materials that could be added. Suggestions for additions and improvements to the knowledge base are therefore very welcome.

41. The manual produced to accompany the UNECE / Statistics Finland training courses on the use of administrative sources\(^4\) will be further developed as requested by the Bureau. Additional inputs are welcome.

\(^3\) http://www1.unece.org/stat/platform/display/adso/ASSIST  
\(^4\) http://www1.unece.org/stat/platform/download/attachments/8193015/Using+Administrative+Sources+v1.2.pdf?version=1