Building Confidence in the Use of Administrative Data for Statistical Purposes

57th Session of the International Statistical Institute,

Durban, 16-22 August 2009

Richard Laux (UK Statistics Authority), Antonio Baigorri and Walter Radermacher (Eurostat)

Introduction

The production of official statistics in the European Statistical System (ESS) is facing major challenges. There is the need to satisfy an increasing demand from users for more and better statistics, including the faster measurement of new phenomena. At the same time, these users' requests are formulated in an environment requiring the response burden placed on businesses and citizens to be limited. In addition, response rates, especially to household surveys, are declining. Finally, official statistics are, and will be confronted in the future, by a situation of frozen resources that will enforce a systematic increase in productivity and efficiency.

Reacting to this situation requires a re-think of the way that we produce statistics, to be able to cope with present and future demands. The production systems in place should adequately balance the available resources with users' needs, and rely more and more on the intensive use of data collected for non-statistical purposes (secondary data) instead of surveys (primary data). This implies mainly using administrative data or, possibly in the longer term, private sources of data in certain statistical domains such as price statistics. Combining the different sources in a warehouse-type environment and linking data is the next step towards a more efficient way of producing official statistics that satisfy users' demands in a reasonable way, whilst reducing costs and the response burden. Consequently, the use of surveys is expected to decline in the future in favour of a more intensive use of administrative data. Statistical offices will therefore need to adapt to these trends; in the ESS some of them have already moved in this direction.

This new statistical setting also brings challenges of a different nature in relation to:

- the way in which ethical principles related to integrity are applied, in particular professional independence, impartiality, objectivity, equal access for users and respect of confidentiality;
- data quality, mainly arising from the fact that different concepts and methods are used in the collection of administrative data;
- the management of production systems which are more complex than those based on the exploitation of survey data; and
- implementation within the statistical offices, which will require an emphasis on innovation.

Several initiatives are relevant to overcoming the challenges mentioned above. Integrity and quality aspects of the production and dissemination of official statistics are governed by the Statistical Law and by the European Statistics Code of Practice (“the Code”). The integrity and quality principles in the Code are also applicable to administrative data used for statistical purposes. Necessary adaptations of the Code could be considered, in order to strengthen these principles in a production system based on multiple data sources, and their combination. This would provide a more adequate regulatory framework relating to integrity and quality in the production and

dissemination of statistics in this environment.

To incentivise the change needed to move from a production system under direct control, to another that is less controlled but more efficient in addressing users' needs, several elements need to be taken into consideration. These include:

- the capacity to produce statistics in innovative ways beyond traditional approaches;
- the education and training of producers of statistics;
- a reinforced communication between producers and users, and between producers and other stakeholders; and
- the increasing importance of metadata in qualifying and explaining data to users

The following chapters elaborate further how these challenges can be addressed and mainly focus on the use of administrative data for statistical purposes. Some reflections are also provided in relation to the use of private sources of data by statistical offices. Finally some conclusions are presented along with ideas on how to steadily progress towards a system of official statistics that is capable of meeting the potential demands from society in the coming years.

The Code of Practice and administrative data

The Code relates to the production, management and dissemination of European Statistics – essentially, those required to support the European Community's activity as specified in law. European Statistics are produced by national statistical authorities and by Eurostat. National statistical authorities compile their statistics from surveys, and from administrative sources of data, and by combining these two types of source. In different countries the balance between the types of sources differs, according to national arrangements.

The Code is written in such a way as to apply to statistics based on both types of data. In relation to administrative data, it covers elements of *legitimacy* (including accessibility to the underlying data, for statistical purposes), and *quality*.

Legitimacy

The United Nations' Fundamental Principles include the following:"5. Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records."

This principle is well understood. It is implemented by different NSIs and in different statistical domains according to the existence, availability and quality of relevant administrative data. The strengths and limitations of administrative data for statistical purposes are well documented elsewhere, and the potential to reduce the burden otherwise imposed on respondents remains an important driver for the increased use of administrative information.

The Code addresses the issue both directly:

- Principle 2, Indicator 2: The statistical authority is allowed by national legislation to use administrative records for statistical purposes
- Principle 9, Indicator 5: Administrative sources are used whenever possible to avoid duplicating requests for information.

- Principle 10, Indicator 4: Proactive efforts are being made to improve the statistical potential of administrative records and avoid costly direct surveys.

and indirectly:
- Principle 1, Indicator 2: The head of the statistical authority has sufficiently high hierarchical standing to ensure senior-level access to policy authorities and administrative public bodies.

Quality

Because administrative data exist as a by-product of systems designed for other primary purposes, it is particularly important for statisticians to ensure that the resulting statistics meet the quality needs of users. The Code of Practice addresses quality issues (especially relating to accuracy and comparability) of statistics based on administrative data as follows:

- Principle 8, Indicator 1: Where European Statistics are based on administrative data, the definitions and concepts used for the administrative purpose must be a good approximation to those required for statistical purposes
- Principle 12 Indicator 2: Sampling errors and non-sampling errors are measured and systematically documented ….
- Principle 14 Indicator 4: Statistics from the different surveys and sources are compared and reconciled.

Some proposals for developing the Code

The ESS believes that there should be a clear and convincing rationale for considering any additions to the Code. The longer the list of indicators in the Code, the greater the difficulty for statistical authorities in observing it – the key is to ensure that the Code evolves in such a way as to deliver, and to be seen to deliver, improving good practice.

In the current context, further developments to the Code should only be considered in relation to issues that are specific to administrative (as opposed to survey) data. Hence good practice in relation to, for example, confidentiality, or the provision to researchers of access to micro-data, does not need to distinguish between survey and administrative data in the context of indicators in the Code. But there do seem to be areas in which it might usefully be developed, in relation to: ownership and governance of administrative data; the quality of the underlying data, data linkage, and response burdens.

Ownership and governance of administrative data

Shared commitment

In order to help ensure that administrative data are fully exploited, statistical authorities might consider establishing protocols with providers of administrative data used for the production of European Statistics. Such protocols could cover details of the administrative sources, procedures to ensure that the NSI is consulted and involved in any changes to the administrative system and data collection, information on related administrative sources and systems with statistical potential, access arrangements, quality reviewing arrangements, security arrangements, and so on. The Code could require that: “Statistical authorities should publish agreements with owners of administrative records which set out their shared commitment to the use of these records for statistical purposes and ways in which this Code of Practice will be applied”. It could also require that: “Statistical authorities should be involved in the design and configuration of administrative

data in order to make administrative data more suitable for statistical purposes”

Information about those with access to the underlying administrative data
Most ministerial departments operate administrative systems. Usually these systems generate (administrative) data, some of which are subsequently published as official statistics. Ministers need to be aware of latest trends that may require their executive action, if they are to be able to act in the public interest in fulfilling their responsibilities. Because their being made aware of latest trends amounts to a form of pre-release access, they are at risk of inadvertently leaking statistical information, and their impartiality being criticised. Transparent arrangements could be a powerful tool in this instance. The Code could require that: “For all statistical outputs produced from administrative data sources, summary details should be published describing the nature of any access to the data and derived statistics ahead of the publication of the statistics themselves”.

The quality of the underlying data
Whilst the Code already touches on some aspects of quality, as noted earlier, it could be modified to ensure that suitable information is provided to users. It could require that: “The strengths and limitations of administrative sources used for statistical purposes should be published”.

Data linkage
Statistical authorities should be supported in their efforts to link different data sources (where the linked source will be used only for statistical purposes). The Code could require that: “Linking data sources to reduce burdens and to add analytical value should be encouraged. Obstacles to linkage, such as inadequate identifiers, should be tackled”.

Response burdens
The procedures to collect data from citizens and businesses need to be more integrated. The collection of administrative data should aim to simultaneously focus on both uses - administrative *and* statistical purposes. In addition, before launching a new survey, statistical offices should be certain that the request cannot be satisfied from existing administrative data. The Code could introduce both requirements in the context of the existing principle of the Code relating to the non-excessive burden on respondents by strengthening or extending the wording of the indicator “administrative sources are used whenever possible to avoid duplicating requests for information”. For example, the UK’s Code of Practice for Official Statistics requires producers to “evaluate existing data sources and estimation techniques before undertaking new surveys”.

What is beyond the Code?

Adapting the Code will not be enough in itself to enable statisticians to be able to exploit administrative data more effectively; it is necessary but not sufficient. A number of complementary activities are required, primarily:

- reinforced communication with main stakeholders (i.e. users of statistics including the general public and the owners of administrative systems); and
- a strategic investment by statistical authorities to facilitate the implementation of this complex production setting, requiring the alignment of arrangements for governance, organisation, resource prioritisation, and planning, complemented by a policy framework to support the exploitation of administrative data

Communication activities

a) More interactive communication between producers and users

It is important to provide users with a precise and clear message, informing them that official statistics are compiled in compliance with the Code of Practice. A quality mark for European statistics would communicate to users that both, product and process quality, as well as the institutional environment in which the statistical authority operates, have been scrutinized and found to be compliant with the principles and indicators of the Code. The concrete design of the quality marking, the procedure and the quality mark itself, could take various forms (for example, it could be one-dimensional or multi-dimensional, and cover all or only a subset of the statistics produced, and so on) and will depend on the objectives/benefits intended by this quality mark. In principle multidimensional quality marking is more informative, in particular for experienced users but fundamentally such a mark should communicate trust, inform users and provide guidance for the use of statistics.

Standardized metadata is a key element in informing users about the integrity, quality and other aspects of the data. Principle 15 of the Code requires producers to disseminate data with supporting metadata and guidance, using standard metadata systems. This includes the sources used - less standardization of data sources will imply a greater effort in explaining them- and how the statistics are compiled. The Euro SDMX Metadata Structure (ESMS) provides a harmonized platform to document statistical data in a structured way and provides information to assess the production processes as well as product quality according to the ESS quality criteria. ESMS also incorporates concepts suitable for describing complex production systems, including elements such as the type of sources used, the data collection methods applied, and the data compilation methods used to derive new information.

This reinforced communication should not only be from producers to users. Adequate channels to facilitate interaction need to be opened and proper tools (such as Web2.0 based approaches) to support them put in place, to enable users to offer their views about the quality and relevance of sets of statistics for their own uses, as well as to receive information about statistics..

b) Communication with the general public

The production environment under consideration might be considered to lack transparency and may create a sense of mistrust amongst citizens in relation to the use of administrative data for other purposes, leading to a kind of 'Big Brother' effect in society. For this reason statistical offices need to dedicate efforts to communicate with citizens and assure them that the administrative data provided are only further used for legitimate statistical purposes, and that is done in line with the ethical considerations promoted in the Code.

In this context and consistent with the Code it is also worth clarifying to the general public those uses of administrative data which are very close to statistical purposes. This is the case for example for administrative data on registered unemployment/employment. These kinds of data appear frequently in the media, and statistical offices therefore need to provide basic information on the concepts used, to avoid confusion and misunderstandings with the data presented in statistical releases.

c) Communication and coordination with the owners of administrative data

When combining administrative and survey data special attention needs to be given to the

comparability and timeliness of statistics as both could be negatively affected. Comparability could be influenced by the use of concepts and definitions in administrative data that differ from those used in statistical surveys. The timeliness of statistics may not be fully under the control of statisticians because of different or non harmonized data collection schedules. Strengthened coordination between statistical and administrative authorities is necessary to address these potential shortcomings, and the protocols proposed above could act as a suitable instrument – by formalising and institutionalising coordination.

In addition the simultaneous data collection proposed above will require an intensified collaboration between the public administrations involved, as well as the use of standardized tools to facilitate data transmissions. It is worth mentioning, as good practice, the experiences from Portugal and the Netherlands regarding the single transmission of accounting data to serve regulatory, tax and statistical uses.

Factors to mitigate resistance to change and overcome barriers

Statisticians are, rightly, keen to ensure that when using administrative data they have the information needed to understand (and can explain to users) issues relating to data quality, to integrity, to confidentiality, and so on. This undoubtedly introduces challenges. Focusing on 'quality', for example, it is likely that additional indicators of the ESS quality criteria – relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability, and coherence – will be needed in future. The relevance of statistics should involve offering better targeted datasets to more users and the coherence of the statistics disseminated should be much better in a warehouse type production setting. At the same time the accuracy criterion will need to be reviewed. Having a common methodological reference for sample surveys has contributed in the past to the development of a detailed set of quality indicators related to accuracy. This common reference is not systematically applicable to administrative data combined with surveys. Error estimation becomes much more complex, and a revised set of quality measures needs to be further established. In addition the combination will introduce a new factor relating the quality of the outputs to the quality of several inputs.

Without strong leadership, an emphasis upon innovation, and a willingness to learn from others' best practices, the implementation of more efficient approaches to our business may be slowed down. At the same time, governance and organizational structures, and appropriate policies on training and on human and financial resources, need to be aligned with the new multi-source production setting that is required to respond to the challenges described at the beginning of this paper. Investment must be targeted on specific activities that will play the greatest role in facilitating the change within statistical offices – such as: the promotion of harmonized and standard processes; common IT tools; common dissemination platforms; and so on. These investments at the level of the ESS could be accommodated within different programs providing financial support from the Commission on the combination of survey and administrative data such as the future ESS-net on Data Integration and the Modernisation of European Enterprise and Trade Statistics (MEETS) program

The research community can also play an important role in facilitating the implementation by statistical offices of this complex production setting. In the past, research in official statistics has been mainly focused on compilation methods for primary data; this focus may need to be revised to align research work with the support required by statistical offices. Research on areas such as

integration and data linking, the adaptation of the framework for quality, and so on, could identify potential implementation problems in advance. The Seventh Framework Programme for research and technological development (FP7) is the European Union main instrument for funding research over the period 2007 to 2013 and can still play a key role in supporting research, in the direction mentioned above.

The role of external bodies

To build confidence in the use of administrative data for statistical purposes, external advisory bodies composed of high level representatives with national and international professional experience in the field of statistics, can play an active role and convey an independent view on two aspects:

- On how the Code is implemented by statistical authorities in relation to the use of administrative data for statistical purposes
- On advising how to address new demands on statistics from existing administrative data, as well as on communication with the different type of users of statistics.

The European Statistical Governance Advisory Board (ESGAB) and the European Statistical Advisory Committee (ESAC) fulfil these roles, in the European Statistical System. The scope of their mandate goes beyond administrative data and refers to the production and dissemination of European Statistics:

- The purpose of the ESGAB shall be to provide an independent overview of the European Statistical System as regards the implementation of the European Statistics Code of Practice

- The ESAC mandate request this Committee to: assist the European Institutions in ensuring that users requirements and the costs borne by information providers and producers are taken into account in coordinating the strategic objectives and priorities of the Community's statistical information policy.

Similar kind of advisory bodies have or are being implemented in a number of Members States. In the UK, for example, the Statistics Authority has published a Code of Practice for Official Statistics which explicitly addresses the use of administrative sources for statistical purposes, requiring that “administrative sources should be fully exploited for statistical purposes, subject to adherence to appropriate safeguards”. The Authority is responsible for the assessment of producers of official statistics against this Code, and so is in a strong position to monitor and support the use of administrative data.

Some reflections on private sources of data

In the context of reducing response burdens and better satisfying users' demands it is also worth mentioning the use of private sources of data for statistical purposes. These are data, outside the scope of the mandate for data collection (Principle 2 of the Code), which are produced by entities either as being part of their core business or to support their activities and management. This category may include, for example data compiled to be commercialized, accounting data (beyond principle 2 of the Code) and other data of more general nature, such as databases used to manage enterprises or other organizations (for example prices, product consumption, and so on).

The use of private data for statistical purposes may bring quality problems, in particular related to coverage and possibly introducing bias. If this is the case, their use should first be limited to producing experimental statistics with the purpose of quality improving; filling gaps; or to complement information (for example, with additional breakdowns) in official statistics. The use of these private sources of data may be hampered by confidentiality issues particularly referring to household or persons. For this reason it may be more appropriate to start with enterprise data than household/personal data. In addition the transmission of these data to statistical offices need to be facilitated, for example by moving from a “push” to a ”pull” model (online data collection from business accounting, collaboration with providers of ERP-software) in order to reduce the burden on private companies.

Within the European Statistical System, an initiative to acquire data from different data suppliers has been undertaken to build the Euro Groups register (EGR), which includes multinational enterprises. This register should contribute to a general improvement of the quality of business statistics and in particular of Balance of Payments (BoP) and Foreign Affiliates (FATS) statistics.

Conclusions

This paper argues that building confidence in the use of administrative data for statistical purposes requires statisticians to seek legitimacy from data subjects and owners of the administrative data. But it also requires statisticians to demonstrate to a wide range of stakeholders, especially users, that they are competent and forward working. Within this context, developments of the Code of Practice are a necessary, but not a sufficient condition. In order to complement the development of the Code, the statistical community should:

1. develop approaches to communicating with the general public, users, owners of administrative data and politicians, that reflect their various perspectives, needs and interests;
2. review their governance and organisational structures to support leadership and delivery in areas as diverse as cultural change, quality management, IT systems, and training - all of which should be aligned with the business imperative of making more effective use of administrative sources;
3. make use of Commission co-ordinated research programs such as the MEETS or the Research Framework Programs to accommodate new development work related to methodology and quality;
4. explore ways of making greater use of the largely untapped potential of private sources of (administrative) information – for example, from business and local government – to develop new or improved statistical products, to understand better the quality of our statistics and to reduce burdens on data supplies.

Such a multi-faceted strategy seeing statisticians take ownership of the challenges associated with making greater use of administrative data, would demonstrate our credentials, our legitimacy and our competence – and would therefore build confidence in our use of administrative data.

REFERENCES (RÉFERENCES)

Baigorri A., Hahn M. (2007) Eurostat communication with users on quality of statistics Eurostat 2007 Conference "Modern Statistics for Modern Society"
De Leeuw E. D. (2005) To Mix or Not to Mix Data Collection Modes in Surveys The Journal of Official Statistics, 21(2), 233-255
Eurostat (2008) The Implementation of the Code of Practice and its relation to independence and ethical issues. Bi-annual IAOS Conference. Shanghai
Göttgens R., Snijkers G. (2007) Die Strategie der Datenerhebung bei Unternehmensstatistiken in den Niederlanden. Wirt Sozialstat Archiv 1: 135–143
Körner T., Radermacher W. (2006) Data collection under pressure: towards a strategy of mixed data sources. Eurostat Quality Conference. Cardiff
SnijkersG. (2009) Getting Data for (Business) Statistics: What's new? What's next?. NTTS Conference
UK Statistics Authority, Code of Practice for Official Statistics (2009) (http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html)
Walgren A., Walgren B. (2007) Register-based Statistics - Administrative Data for Statistical Purposes John Wiley & Sons, Ltd