What are Administrative and Secondary Sources?

1.1 Introduction

Before we start to consider the practicalities of using data from administrative and secondary sources, it is worth just taking some time to clearly define what these terms mean. Several definitions exist in the literature currently available, the most relevant of which are examined in this chapter. The chapter ends by proposing a relatively simple and broad definition, which is then used as the basis for the remainder of this handbook.

1.2 Traditional Definitions

Administrative sources have traditionally been defined as collections of data held by other parts of government, collected and used for the purposes of administering taxes, benefits or services. Perhaps the most comprehensive of the traditional definitions was set out by Gordon Brackstone of Statistics Canada in his 1987 paper “Statistical Issues of Administrative Data: Issues and Challenges”[1]. Brackstone identified four distinguishing features of administrative data:

1. The agent that supplies the data to the statistical agency and the unit to which the data relate are different (in contrast to most statistical surveys);
2. The data were originally collected for a definite non-statistical purpose that might affect the treatment of the source unit;
3. Complete coverage of the target population is the aim;
4. Control of the methods by which the administrative data are collected and processed rests with the administrative agency.

This definition is broadly in line with that proposed by the Statistical Data and Metadata eXchange (SDMX) initiative[2]:

“A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations.”

During 1996-97 an internal Eurostat task force examined ways to better coordinate work relating to the use of administrative sources across different domains of statistics. This task force used a simple typology of data sources to consider how administrative sources should be defined. Firstly all data sources were divided into primary sources (data collected for statistical purposes) and secondary sources (all other data). A traditional or “narrow” definition of administrative sources comprises just public sector non-statistical sources, whereas a wider definition would also include private sector sources.

The wider approach is consistent with the definition of administrative data adopted by the Conference of European Statisticians in the publication “Terminology on Statistical Metadata”[3]:

“Data collected by sources external to statistical offices.”

The narrow and wider definitions can be shown graphically as follows:

<table>
<thead>
<tr>
<th>Figure 1.1 - Narrow definition</th>
<th>Figure 1.2 - Wider definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Diagram of narrow definition" /></td>
<td><img src="image2" alt="Diagram of wider definition" /></td>
</tr>
</tbody>
</table>

Thus under the narrow definition, administrative sources are a sub-set of secondary sources, whilst under the wider definition these terms are synonyms.

There are a growing number of reasons for favouring the wider definition, including:

- **Increasing privatisation of government functions:**
  
  In several countries, regulatory functions that used to be carried out by government departments or agencies are being transferred to private or semi-private organisations. Typical examples are usually in the health, education or public utilities sectors, where former state monopolies are increasingly being replaced by private companies or non-profit institutions.

  Registration functions, including the operation of administrative registers on behalf of government departments are also under consideration for privatisation in several countries. This means that the traditional distinctions between public and private sector functions are becoming increasingly blurred, and that the traditional or “narrow” definition of administrative sources is becoming too restrictive.

- **Growth of private sector data and “value-added re-sellers”:**
The amount of digital information in the world is growing exponentially, increasing by a factor of ten approximately every 5 years. Even if only a tiny fraction of this “data deluge” is of interest for official statistics, the volumes of data, and the range of topics they cover are still huge.

At the same time, the commercial value of data is starting to become apparent, and the market for data is rapidly increasing within the private sector. This started with the development and sale of address lists for marketing purposes, it expanded to cover the provision of credit rating data and business intelligence information, and has now spread to cover virtually all types of data. As the size of this market has increased, so has the number of businesses seeking to profit from it. The private sector realises that data are a very valuable commodity.

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 Bonnier.

This sort of data source can be of interest to official statistics providers, as it may be the case that 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 “Eurogroups” project to develop a European statistical register of enterprise groups uses such sources for exactly this reason.

An alternative to direct use of micro-data from such sources can be the use of aggregates for benchmarking purposes, comparing the coverage of target populations between private sources and official statistical registers. An exercise to compare the coverage of the UK statistical business register with that of leading private sector sources revealed statistical under-coverage of business activities in inner-city and holiday resort areas, illustrating the difficulties associated with covering marginal and seasonal activities in official statistics, as well as giving clear indications of the scale of this sort of under-coverage.

User interest in new types of data

Users of official statistics are constantly requesting new types of data. Pressures to reduce costs and burdens on respondents to statistical surveys make it difficult to launch new surveys to meet these demands, so statisticians increasingly need to look for alternative solutions. As the volume, content and coverage of private sector sources grows, so does their attractiveness as an alternative to statistical surveys.

1.3 Types of Administrative Sources

As discussed in the previous paragraphs, the potential range of administrative sources that could be used for statistical purposes is large and growing. The following list is not meant to be exhaustive; instead it aims to show range and types of potential data sources, as the final step towards arriving at an operational definition of administrative sources.

- Tax data
  - Personal income tax
  - Value Added Tax (VAT)
  - Business / profits tax
  - Property taxes
  - Import / export duties

- Social security data
  - Contributions
  - Benefits
  - Pensions

- Health / education records
- Registration systems for persons / businesses / property / vehicles
- Identity cards / passports / driving licenses
- Electoral registers
- Register of farms
- Local council registers
- Building permits
- Licensing systems e.g. television, sale of restricted goods
- Published business accounts
- Internal accounting data held by businesses
- Private businesses with data holdings:
  - Credit agencies
  - Business analysts
  - Utility companies
  - Telephone directories
  - Retailers with store cards etc.

1.4 Summary

In conclusion, this chapter argues the case for a wide definition of administrative and secondary sources. It also highlights the need for imaginative assessments of the potential value of new types of data sources. For these reasons, the definition of administrative and secondary sources should not place any artificial restrictions on statisticians, and should be as wide as possible. As the terms “administrative sources” and “secondary sources” are therefore considered to be synonyms, this handbook will henceforth just use the term “administrative sources”, to cover both concepts.

The definition proposed is therefore:

Administrative sources are data holdings containing information which is not primarily collected for statistical purposes.

This definition is used as the basis for the contents of the rest of this handbook.
Box 1.1 – Looking to the Future: Store Cards – a Potential Data Source?

Store cards are a typical example of a new type of private sector data source. In return for benefits such as discounts and exclusive special offers, users of store cards give the stores a lot of data every time they use them. If you have a store card, the store knows or can derive the following data about you:

- Name, address, sex, age
- Family circumstances (e.g., if you regularly buy baby products, toys, pet food, or products such as meat in a certain quantity or size, it is easy to estimate the composition of your household)
- Indicators of work status and income (e.g., the time at which you shop can indicate whether or not you work, and the type of goods purchased can indicate disposable income)
- Other household indicators, such as car ownership (purchases of petrol and car care products), religion (purchase of goods linked to a particular religion, e.g., halal or kosher meat), etc.

This may seem a rather extreme example of a potential source, and one that is unlikely to be considered for the purposes of official statistics in the near future. However, several countries have considered the use of till roll data from major retailers as a source of data on retail sales and prices, and Statistics New Zealand has produced an experimental data series using electronic card transaction data[5].

The use of store card data could be seen as the next logical step, particularly if coverage can be improved by linking data from different store card schemes, as well as data from other commercial sources. If this sort of administrative data source is ignored by official statisticians, how long will it be before private sector businesses with access to these data, start to offer plausible, and more cost effective alternatives to key official statistical outputs such as population census data?