Challenges for statistical disclosure control in a world with big data and open data

Peter-Paul de Wolf*
Statistics Netherlands, The Hague, Netherlands – pp.dewolf@cbs.nl

Kees Zeelenberg*
Statistics Netherlands, The Hague, Netherlands – k.zeelenberg@cbs.nl

Traditionally, statistical disclosure control by national statistical institutes (NSIs) has focused on tables and microdata collected and produced by the NSI itself. In the last decade NSIs have increasingly also been using administrative data, collected by government agencies; and usually, the individual administrative data have not been published and have not been used by other government agencies or private companies.
However, in recent years this context of an almost monopoly for the NSIs has changed. Governments have been actively pursuing an *open-data policy* with easy access to government data, not only as a service to the general public, but also with the express purpose of stimulating the market for data. NSIs are also beginning to offer their tabular data as open data, by providing a public and easy way to download these data in an automatic way. This market for data has been further stimulated by the advent of *big data*, data that come in large amounts, sometimes in millions or even billions of records per day; some examples are telephone records, traffic data, and banking transactions. These big data have made it possible for private data companies to assemble databases with very detailed information on individuals and enterprises, with data coming from many sources, and linked sometimes deterministically but often also probabilistically.
This abundance of data poses new problems for statistical disclosure control, both strategic and ethical problems as well as methodological problems. For example, should NSIs, when protecting tables or data files against disclosure, take into account the possibility, or even the absolute certain fact, that government agencies and private companies will use the NSI data to enrich their own databases and so get to know more about their citizens or customers? And what if these enriched data are used to profile citizens so that they may come under suspicion or are denied access to certain services such as loans? And should we use another methodological paradigm than we have used so far? Should we consider different disclosure scenarios? Does the changing attitude towards privacy influence the way we should treat our published data? How much existing as well as future data should we take into account when assessing disclosure risks?

Keywords: disclosure control; open data; big data.