Confidentiality in an era of Big Data: an official statistics perspective

Stephen John PENNECK

ISI Vice-President

Former DG, Office of National Statistics, UK

Abstract

The presentation will emphasise the importance of confidentiality in maintaining trust in official statistics. It is encompassed in principle 6 of the UN Fundamental Principles of Official Statistics; is one of the ethical principles and the shared professional values of the ISI's Declaration on Professional Ethics; and is principle 5 of the EU Code of Practice. The requirement is enshrined in the statistical law of many countries as a legal requirement for national statistical offices.

In official statistics, data collection may be undertaken voluntarily, as with many household surveys, or may be collected under legal powers. Business surveys and the census of population are typically collected with legal sanctions for refusal to provide information, and an associated legal requirement on the statistics office to keep such data secure. Where data collection is voluntary, 'informed consent' is sought from data subjects to the use of their data. Where data is collection under legal powers, data subjects will be told that their data is protected by the law.

The world of statistics is changing and Big Data brings new challenges. Big data say a lot about who we are, what we do, who with and where. They tend to be personal data, with an enormous potential for use in official statistical systems to tell us more about society.

It has been said that Big Data raise no new ethical issues for statisticians, but they do add a great deal of complexity. Data subjects may not understand the use to which their data may be put. Most national privacy legislation secures the right of individuals to control what information about them may be disclosed. With Big Data, data subjects may be unaware they are generating data and what it can be used for, despite the efforts of the social media platforms in this respect. One of the challenges is to manage the acceptance of data re-use and data linkage, which would not necessarily be expected by data subjects.

Big Data are valuable assets for the companies that hold them. The data

are generally not publicly available and to gain access, statistics offices need to negotiate agreements or gain legal access rights. Data need to be obtained in an identifiable form to enable them to be used statistically – eg to enable them to be linked to a random data set for estimation purposes, or for records to be linked to produce wider analytical value: but the resulting published results need to be non-identifiable.

Much Big Data reach across international boundaries, and there have been a number of international initiatives to look at the implications of Big Data for official statistics. A lot of this looks at some of the technical issues, developing the tools and skills needed to resolve the methodological questions, rather than the ethical questions. Given many official statistics are collected using legal powers, much emphasis has been on the legal protections required. Official statistics in the EU is governed by the European Statistical Law, and in 2015 it was revised and now gives national statistics offices the right to access administrative records held by public bodies.

In the UK the Digital Economy Act 2017 facilitates the linking and sharing of public sector data sets for research purposes. It also gives the UK Statistics Authority the power to require private sector companies to provide digital data for official statistics purposes. The Act requires the UK Statistics Authority to prepare a Code of Practice concerning the holding, processing, disclosure and use of personal information under the Act. Consultation on this code is currently under way [to be updated]. The draft Code of Practice includes seven principles including:

- Confidentiality: all processing and disclosure of data must minimise the risk of compromising the confidentiality of personal information
- Onward disclosure: any data released to researchers must undergo disclosure control to minimise the risk of re-identification.

In Europe a new regulation, the General Data Protection regulation, coming into force next year, takes data protection further. It brings pseudonymised data within the scope of data protection. Guidance on this is still being developed.

The EU has adopted a road map which gives a way forward for European NSIs. In the long term they look for legislation to be adapted so that it is compatible with the ethical use of Big Data in official statistics.

So, developing a legal framework is the route that many statistics offices are following, especially in Europe, but the use of legal powers does not negate the need for public debate and acceptance of how data are used. In a democracy, the legal framework can only advance in line with public opinion, and legislatures will be wary of being out of step with this.