

B. Integrating new data sources (such as big data) and traditional sources

96. In recent years, the official statistics community has acknowledged the value of big data and has been exploring the use of diverse sources in several domains. Many different types of data sources fit under the umbrella of big data. One example is scanner data on prices, coming from scanner transactions in supermarkets and often provided to statistical organisations by private companies working in marketing. Another example is data scraped from the internet. Official statistics has acknowledged the value of Internet-scraped data and has been exploring their use in several domains (for instance in statistics on ICT use in enterprises and tourism). Data can be scraped directly from individual websites, but this approach requires first identifying the websites and then dealing with different queries and different formats obtained from each website. Alternatively, data can be scraped from "hub" websites describing a plurality of units (for instance hotels data), although the information available may be summarised.

97. It is quite common to state that big data provides information useful for statistical purposes in a way that is substantially cheaper, faster, more timely than survey and administrative data. However, it is not always recognised that the relevance of data coming from the new sources should be investigated first. Moreover, the introduction of big data approaches, i.e. data provider agreements, new IT tools and capabilities, can also be very expensive and time consuming, jeopardizing at a first step the advantages of the big data usages.

98. One significant opportunity arises from the global nature of some big data: the opportunity for statistical organisations to collaborate on crafting global data agreements and global partnerships with big data providers.

99. Consequently, the constraints/limitations in the usage of big data should be properly understood, at the design stage of the data integration activities. As usual, innovation requires acceptance of some risks, but those should be clearly understood, stated and managed to mitigate them.

100. A way to mitigate risk could be to focus on expectations, making them as clear and reasonable as possible. Moreover, the use of big data requires flexibility agile approaches, due to often unexpected changes in the source data. For example, in web scraping website changes and data layout changes can occur without warning. Good relationships and agreements with data providers may help in managing these situations; however, it is important to consider in advance what might go wrong and how to react.

101. For some of these external sources, the reported objects can be easily associated with statistical units of the target population. On the contrary, there are cases in which the big data objects need to be elaborated in order to be compared to statistical units. In most cases, when big data are not directly comparable with data collected and organized by statistical organisations, a lot of work is needed to create integrated data. Finally, sometimes the big data offers information on topics that are not well covered by traditional surveys - in this case the advantages in their use is unquestionable.

102. In addition, it is likely that the big data are not standardized or codified. In some sense, official statistics are already prepared to face this kind of problem, but the pre-processing phase is a very time-consuming process and a lot of work is needed to identify models that can easily support data reconciliation, management of the complexity and to allow the data integration step. In order to integrate big data in the statistical production process, a system is required for data ingestion and reconciliation that allows managing a big data volume of data coming from a variety of sources. The statistical production system needs to produce the ontology and the big data architecture, and the mechanisms for the data verification, reconciliation and validation.

103. In the cases of coincidence or harmonization between big data objects and statistical units, if a unit identifier is available and shared with the statistical organisation, the big data can be integrated with existing statistical data at micro-level, so to enlarge the content, the coverage, the accuracy and the timeliness of official statistics. When identifiers are not available, big data can be used in combination with other sources at aggregated levels.

104. When using and integrating new data sources, new methods may need to be developed and integrated with the existing ones. The opportunity to study and develop new methods requires some patience to allow them to evolve and to become stable. In this spirit, it is important to not leave research works in the drawer when they don't produce positive or expected outcomes, so that other groups don't replicate unsuccessful experiences.

105. It is important to collaborate with developers/administrators of sandboxes and big data technologies. The IT sector is strongly involved in the modernization program looking at the use of big data. Due to their characteristics, big data are often difficult to integrate into existing systems so costly changes to IT infrastructures may be necessary.

106. The usage and integration of new data sources require a composite team of skills and professionalisms. The best would be a team composed by experts from methodology, IT, social-media, subject matter, new tools (e.g. web scraping, visualization)

107. As for administrative data, it is important to assess the quality of the input, the throughput and the output, however, sometimes the input it is not under full control of the statistical organisations as well as the procedures in the data processing steps are not fully understandable by traditional skills. In these cases, a good relationship with the data providers are important to understand the data, so as definitions and concepts behind the data will help in evaluating their quality.

108. A framework for quality assessment when using big data seems to be produced in compliance with already existing quality framework, e.g. in sample survey or in combined use of administrative data. As far as the quality dimensions, it is often noticed that the big data may suffer of coverage issue, being not representative of the target statistical population

109. When using and integrating big data in the statistical production system, it is necessary to proceed in steps, starting with clearly stating the focus of the analysis in which the big data are involved. Consultation with experts around the world, contacts with others approaching the same tasks, learning lessons from other experiences are important points in order to design further steps, as well as understanding risks, and fix what is necessary in terms of skills, IT capabilities and methods.

110. The good relationship and agreement with data providers needs to be established, when the data are not directly collected by the statistical organisations. Then an experimental stage should follow, in which a sample test data should be requested, as well as several new skills, IT tools and methods should be experimented in a common setting, together with well-established tools and methods. The experimental step will require flexibility and patience in evaluating the potentialities of the outcome. Some efforts should be devoted to the quality evaluation and to share also intermediate results with others.

111. At the end, after setting some conditions on how much research should be done (e.g. what kind of quality is acceptable), there will be possible to introduce the new data, tools and methods into the production system, also via comparison with existing results coming from traditional data sources.