5.8 The Internet as a source for statistical data

Why use the Internet as a source for statistical data?

The Internet has become an indispensable infrastructure for economies and societies. An ever growing share of economic transactions, communication and information supply takes place online. Many of these online actions leave digital "footprints" that can be observed using tools that scan, gather, interpret, filter and organise information from across the Internet, providing a foundation for the use of the Internet as a statistical data source (IaSD). Online data may be of use in combination with, or as a substitute for, data collected by traditional instruments such as statistical surveys or off-line administrative sources. For example, online retailers' websites can be a useful source of information about prices while social media may provide information related to employment, population or societal wellbeing.

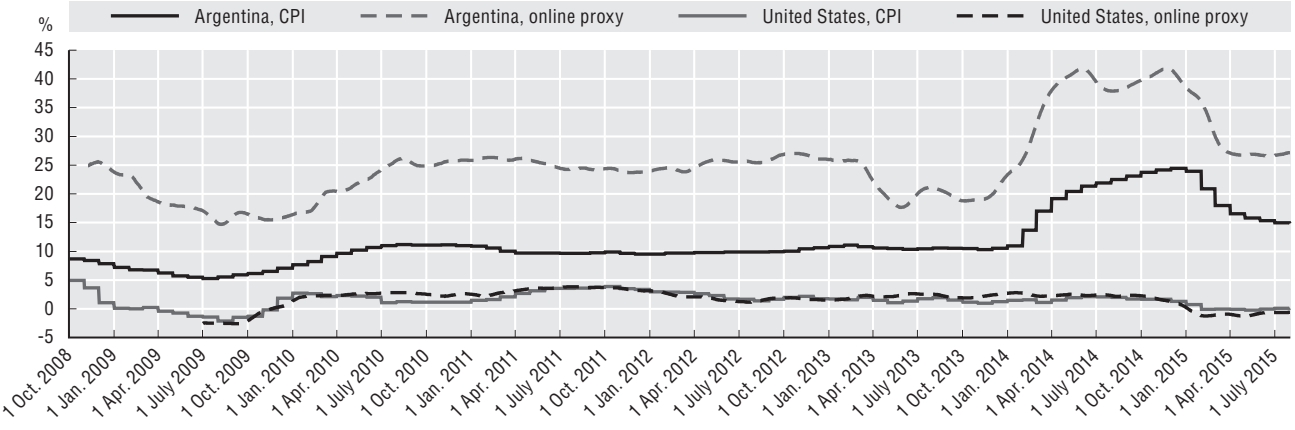
The relatively short history of Internet based social and behavioural research (Hewson et al, 2016) shows that online data can support different elements of statistical activity within national statistical organisations (NSOs) at different steps of the statistical value chain:

- *Identifying and sampling the population of interest*. Internet data can enable efficient updating of registers of statistical units based on Internet presence (e.g. businesses with their own websites or active in online marketplaces), thereby supporting the design of data collection processes.
- *Data collection*. In many instances, web-reading techniques may enable the search for and retrieval of information online that may not otherwise be available with comparable levels of timeliness, detail and exhaustiveness (Bean, 2016). Such data can be timely, especially compared to data collected through traditional survey approaches; Internet search patterns can provide early warning signs about upcoming economic downturns or of health issues emerging in the population, for example. Use of the Internet has the potential to free up NSO resources and reduce response burdens so that surveys can be implemented where they are most effective.
- *Verification / imputation*. Information from the Internet can be used to verify data from other sources, such as surveys. In addition, the use of online information to identify commonalities between respondents and non-respondents may be of use in making imputations to ensure statistics are representative of the target population.
- *Dissemination*. By releasing their statistics online, NSOs also contribute to the enhancement of IaSD for use by expert and interested users, including other NSOs and international organisations.

The use of IaSD is already a reality in many NSOs or is progressively being tested for production environments (e.g. Statistics Canada, US Census Bureau). This opens up avenues to implement subject, object, relationship and network-based measurements (CBS, 2012) that make the most of a vast array of data, including text, images, sound and video files. Of particular interest are data generated in transaction and social media platforms across users through content and service mediation. One example is the "Billion Prices project", an academic initiative aimed at comparing official and alternative, Internet retailer-based measures of inflation, drawing on transaction data. Official data can in some cases be challenged or confirmed, hinting at possible leading indicators.

Comparing official CPI and Internet-based consumer price inflation estimates, 2008-15

Annual Consumer Price Index inflation rates, Argentina and United States, 2008-15



Source: OECD calculations based on Cavallo and Rigobon (2016).

StatLink https://doi.org/10.1787/888933930497

Website metadata, hyperlinks to other sites, logs, cookies and website/subscriber analytics also represent key sources for understanding data flows and network effects. Behavioural data from devices such as smartphones or wearable technology carried by individuals, that record data such as location, physical activity and health status, offer additional