Internet as a new source of information for the production of official statistics. Experiences of Statistics Netherlands.

Nicolaes Heerschap
Statistics Netherlands, The Hague, Netherlands
nm.heerschap@cbs.nl

Abstract

It is likely that Internet as a new data source (*IaD*) will play a role in the production of official statistics at Statistics Netherlands (*SN*) in the future. In spite of the fact that obstacles remain to be overcome, the results of the IaD projects conducted in 2011 and 2012 indicate that there are good opportunities for integrating IaD into the statistical framework. Various benefits can be achieved, either for new or quicker statistics; either for more detail or lesser survey burden or as auxiliary information. Benefits, which fit the strategy of SN. Opportunities to implement IaD will not as much depend on the limitations of the technology and the collection of data, but rather on the possibilities of solving the methodological issues that crop up during the subsequent phases. The use of IaD does not need to be restricted to new statistics. Especially the combination of IaD with existing statistics or the use of IaD as a supplement to existing statistics (e.g. beta-indicators) appear to offer good opportunities. Finally, to fully exploit the potential of IaD a different way of thinking, than the current culture of producing statistics, is needed.

Keywords: IaD, internet crawlers, smartphones, internet, big data, new environment

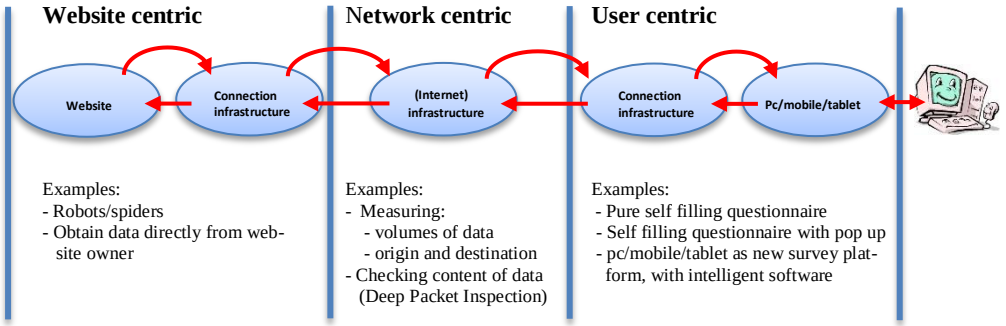
1. Introduction

The internet has become an indispensable infrastructure for society. An ever growing proportion of our supply of information, communication and transactions are now also conducted via the internet. The internet has been fully integrated into our society.

This does not only mean that Statistics Netherlands (*SN*) needs to produce statistical indicators about the nature and impact of internet. It also means that it has become necessary to investigate to what extent the internet itself and related techniques, such as smartphone information and track and trace technology, can be used as a data source for the production of statistics (*IaD*). The available information does not need to be related only to the internet as a phenomenon, but also pertains to data about the "old economy" or in fact about almost "everything". This is possible because activities on the internet leave behind *digital footprints* which can be measured real-time, 24 hours a day, 7 days a week.

This paper gives an overview of the experiences and results of several IaD projects, which were carried out in 2011 and 2012, by SN in a two year stimulation program.

Figure 1. Different types of IaD techniques.



2. New data environment

Recently, more and more data is becoming available for research and statistics. Firstly, this is due to the growing technological possibilities to digitalise data in ever greater volumes at lower costs. This not only involves large administrative sources, but also data sets generated by, for example, the use of sensors, cameras, telephony, gps and rfid-tags. Secondly, the development of internet has played an important role, especially the breakthrough of social media. This is re-inforced by the introduction of mobile internet, supported by mobile devices, such as smartphones and tablets ("portable revolution"), and the emergence of cloud-based storage and processing. It seems less and less a matter of finding data, but rather the question of what to do with it (the research question) and what should be a sound way of processing the data (methodology)?

For an organisation like SN the increasing availability of large digitalised datasets and the rise of the internet as a new data source raise a number of questions:

- to what extent is the internet as a phenomenon well represented in the existing system of statistics and their different frameworks and methodologies?
- to what extent is it possible to use big data and internet as new data source to produce existing statistics faster, better and cheaper with less administrative burden?
- to what extent do these new data sources offer new opportunities for creating new statistical information or develop rapid (early warning) indicators?

Behind these questions, however, another question is hidden, that is how, in the long run, SN can position itself in such a new data environment? A data environment where data, at least from the internet, is available for everybody and where there is more and more competition between private and public organisations for the ever growing amount of big data sets, including the eventual production of valuable statistics. An environment where also the expectations of customers and data users, in terms of faster, more detailed and new statistics against lower costs and a minimum of survey burden, are ever getting higher.

3. Actual practice

3.1. Completed and on-going projects

In 2009 SN began experimenting with IaD. This involved the observation of prices of airline tickets and petrol prices using internet robots. This research on internet prices is slowly elaborated to the domains of clothing, cinemas, driving schools and annual business reports. In 2010 a study was started to investigate the possibilities of the use of mobile positioning data for, among others, mobility statistics, tourism and daytime population. In the same year a pilot was set up to see if Twitter text messages could be used to produce statistics. In the meanwhile this research has evolved in a sentiment index, which reflects the consumer index quite well.

In the beginning of 2011 a two year stimulation program was set up to especially promote projects at SN which investigated the possibilities of IaD and related techniques, such as smartphones. Table 1 gives an overview of the IaD projects, which in the last two years were carried out under the auspices of this program. The program also included two projects which were not directly related to IaD, that is: the impact of social media on society and economy and the construction of a framework of the internet economy. May 2013 the program will end. Results and on-going projects are or will be transferred to the relevant departments or the Knowledge and Innovation Program (*KEI*) which was newly set up by SN in 2012. With the aid of KEI recently projects were initiated on the possibilities to use data from electronic road loops and the interpretation of aerial photographs for the estimation of crop production. Furthermore a separate program was set up for Big Data.

Table 1. Results from the IaD projects carried out by Statistics Netherlands.

Project	POC successful?	Continued after 2012?	Candidate for implementation?	Specific issues
1. Mobile phone and smartphone studies				
1A. Smartphone measurements (as a self filling questionnaire)	Yes	Not sure	+++	Main challenges: recruitment of respondents and funds to continue the research.
1B. Smartphones as a new (intelligent) survey tool	Yes	Yes	++	Main challenges: development of software and carry out a test in the survey on mobility (Ovin).
1C. Mobile positioning data	Not sure	Yes	Unknown	Much time was spent on the negotiations with the data provider. Research is just started.
2. Infrastructure and broadband access				
2A. Broadband access (self filling "app" on a desktop)	Yes/No	No	No	Five years of data from external party. Technology deployable. Panel that was used not representative.
3. Internet robots				
3A. Price observation	Yes	Yes	++	Own internet robots deployed. Still in research phase.
3B. Housing market	Yes	Yes	++++	Own internet robots deployed. Talks with website owners if they are allowing SN to spider their website.
3C. Job Vacancies	Partly	Yes	+++	Data from external party. Challenge: representativeness. Survey burden can be decreased substantially.
3D. Tourism	Yes	Yes	+++	Own internet robots deployed. Especially used to look at the population of smaller tourism accommodations.
3E. Web shops	Yes	No	-	Own internet robots deployed. One off research to test the quality of the population of web shops in SN business register.
4. C-to-C market				
4A. Marktplaats data (= Dutch Ebay)	Yes	Maybe	++	Six years of data from external party. Challenge: representativeness of the data
5. Track and trace technology				
5A. Track and trace	Not sure	Yes	Unknown	This concerned a research project, not a proof of concept
6. Web panel				
6A. Web panel	Yes	Probably	+++	Decision about continuation taken end of April 2013.

POC = Proof of concept

3.2. General conclusions

In this section the main conclusions are presented of the work, which was carried out in the IaD projects in 2011 and 2012. There is no question of completeness.

Opportunities and threats

1. Results show that there are good opportunities to implement IAD in the statistical processes of SN, even in the short term. There appear to be good opportunities for the deployment of internet robots (e.g. price observation, housing market, tourism and to a lesser extent job vacancies), a web panel and the use of smartphones (e.g. mobile services, internet surfing and mobility). Also track and trace technology appear to have potential. Benefits range from new and quicker statistics to possibilities to decrease the survey burden.
2. The implementation of IaD is not necessarily restricted to new statistics. There are also good opportunities by combining existing statistics with internet data. For example, the internet data can provide auxiliary information for producing greater detail (e.g. regional level), improve the quality of statistics and reduce the sample size. In addition, internet data can be used to compose fast indicators for tracking developments and trends ('beta-indicators'). The possibilities to link internet data to existing statistics or delineated population groups significantly increase the probability of successful IaD applications.
3. However, the experience also shows that actual practice can be unruly. Sometimes, at the start of a project it would seem that the benefits are ripe for the picking, while in actual practice this turns out to be more difficult to realise. Generally this is not due to the technology, but much more so due to the processes required to ultimately transform the raw internet or smartphone data into sound statistics. For various reasons, IaD tend to linger in the pilot phase. Much extra effort is needed to make the step from a successful proof of concept to the actual implementation. The development time of projects easily exceeds the two years.
4. The emergence of Big data and IaD will change the data environment for statistical offices, that is the availability of data and the main players in the domain of

statistics. IaD is in fact available for everybody. More and more organisations emerge which have in fact more knowledge of these new technologies and resource than statistical offices. In addition, organisations that generate big data are becoming more and more aware of the value of their data. So, it must be expected that competition from private parties in the field of statistics will increase, especially as big data and IaD are seen as the next big thing. Strengths of statistical offices remain, among others, the possibility to link internet or smartphone data to other sources, time series, (international) comparability, the focus on methodology and the feel for quality of the output and privacy of the data.

Methodology

5. As stated, technology and the actual collection of internet (or smartphone) data are not the real problem. Direct and rapid (new) web statistics can be made, based on the collected internet data (e.g. beta-indicators). The real challenge lies in the methodology to transform the raw internet data into representative en quality statistics, in the phases of processing, analysis and visualisation. This includes among others: translation of unstructured data to statistically useful information (text mining), de-duplication, classification and categorization, linking unstructured sources, volatility of the data and sampling. But especially the representativeness and the noise of not related influences in internet data can be a problem. It appears difficult to determine and quantify these spurious influences. IaD of course also relates to Big data and its specific problems. Furthermore, all kinds of substantive statistical matters pop up, such as estimates based on proxies, estimates based on trends, the development of beta-indicators and the development of new statistics and methodology.
6. IaD offers good opportunities for improving the quality of (existing) statistics. For example, it enables actual behaviour to be measured rather than inquiring about perceived behaviour after the fact. By relating internet data to existing statistics, it is also possible to improve these statistics qualitatively. On the other hand, as indicated, a fair bit of processing of the collected internet data is required to come up with acceptable statistics. This affects the quality of the ultimate statistics. For example, during the production of beta-indicators concessions need to be made in terms of the quality of the output in favour of speed.

The survey burden

7. IaD can be used to reduce the survey burden on companies as well as on persons. It is obvious that in the design of new IaD based statistics no survey burden is created. But in replacing traditional data collection of existing statistics with IaD the survey burden also can be reduced, sometimes considerably. There are, for example, good opportunities with the deployment of internet robots for job vacancies and prices and the use of track and trace technologies in the domain of the transportation, where the survey burden is high. For some cases it was not easy to find a proper balance, that is a positive financial business case, between the savings made on the data collection on the one side and the needed level of structural investments to obtain the IaD data from a third party on the other side.

Legal issues

8. Legal issues continuously play a role in the implementation of IaD projects, especially when it comes to personal data. In fact all the necessary legal steps must be taken care of before the project, including pilots, can start at all. This also must include the physical storage and security of the data. And last but not least one should always be aware of reputational risks, rightly or wrongly. If taken proper care of, legal issues are not a showstopper to implement IaD.
9. Network centric methods, in which passing data is collected somewhere in the infrastructure, can cause legal issues, because it is simply not possible to request

consent from all the owners of the data. A technique like "deep packet inspection" seems to be outside the scope of statistical offices. A possible way out here is to work with anonimysed or aggregated data. With network centric methods the database law and ownership rights play a crucial role.

Organisational issues

10. What one sees happening is that in the beginning in various parts of the organisation people start to set up IaD projects. Most of the time they are not really aware of what others are doing (experimental phase). After a while when the number of IaD projects is growing there is an increased call for coordination and central management with a clear vision and strategy (coordination phase). In whatever direction the coordination phase will evolve depends on the choices made. This could be a more decentralized structure within the framework of a clear centralised strategy and vision how to set up and implement IaD projects.
11. An exception to such a decentralized approach is the use of internet robots. This requires, among others, a specific infrastructure, a clear and recognizable way in which the internet robots are deployed on the internet and specific knowledge for the building and maintenance, including monitoring, of these internet robots. A suggestion could be to establish a Central Data Service Centre.
12. Depending on the specific situation of a IaD project, often the following choices could pass by:
    a. whether the necessary knowledge is developed within the statistical office, the knowledge is temporary hired or that a part of the data collection and other steps in the process are outsourced to a third party?
    b. whether one is prepared to pay money for data from third parties and, if yes, how much? Or if one is prepared to pay money to respondents, so they are more willing to participate in, for example, smartphone studies?
    c. what kind of output must be produced? Because of growing volumes of data the opportunities to produce all kinds of statistics are often unlimited. Choices have to be made on the basis of user needs.

Statistical culture

13. Finally, it should be noted that the traditional method of preparing statistics does not always go hand in hand with the implementation of IaD. The use of IaD requires a different way of thinking (statistical culture). The starting point is the unlimited quantity of data on the internet or from other related techniques from which the desired statistical information is to be extracted to produce the output. This could require concessions to be made in terms of representativeness and quality of the data, the use of proxy's because the data does not exactly match the variables defined for the output and the use of developments rather than measuring levels. An advantage of IaD is that it is much easier to obtain data on flows than with traditional surveys. For example, following the main routes of foreign tourists visiting the Netherlands instead of only the number of nights stayed.

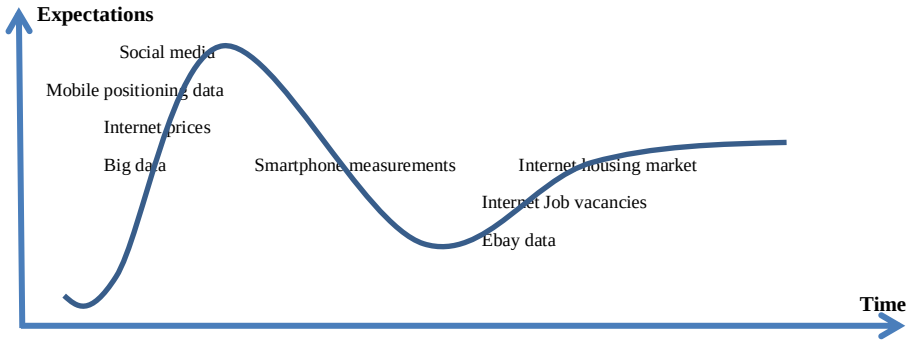
4. Concluding remarks

So the main conclusion is that IaD and IaD-related technologies will likely play a role in the future in the production official statistics of SN. The results of the IaD projects that were carried out in the last two years indicate that there are good opportunities for integrating IaD into the statistical processes of SN. The benefits of IaD fit the strategy of SN. Opportunities to implement IaD in the statistical process will not as much depend on the limitations of the technology and the collection of internet or smartphone data itself, but rather on the possibilities of solving the methodological issues that crop up during the subsequent phases. The use of IaD does not need to be restricted to new statistics. It is especially the combination of IaD with existing statistics or the use of IaD as a supplement to existing statistics that appear to offer good opportunities. Fi-

nally, to fully exploit the potential of IaD a different way of thinking. than the current culture of producing statistics, is needed.

It becomes increasingly clear that the growing importance of Big data and IaD will change the environment for statistics. The availability of data and statistical knowledge will be distributed differently over the players in this domain than it is today. IaD is in principal accessible for everybody. The threshold and investments to collect internet data, build up a database and offer more structured data to third parties or even produce statistics have become very low. While before this was to certain degree the sole domain of research companies and statistical offices. In addition, organisations, which generate big data sets, are becoming more and more aware of the value of their data. This will lead to more competition for data and possibly higher costs for statistical offices to obtain these data. On the other hand the expectations of customers of statistical offices, like policymakers, are ever getting higher, because of the potential possibilities of Big data and IaD, like better, faster en newer statistics with more detail against lower prices and with a minimum of survey burden. In business language: the customers expect shorter product cycles and faster time-to-market processes. This means that the current position of statistical offices cannot be taken for granted anymore, also because the growing use of open data weakens their monopoly position with respect to government data. For statistical offices it is therefore important to evaluate their position and to be ready to adapt and align themselves to the new statistical environment. The question for statistical offices is not whether they should meet these new developments around Big data and IaD, but rather how good they will be to deploy and integrate these developments into their statistical framework. Strengths remain the possibility to link data with other data sources, sound time series, (international) comparable statistics, standardisation, methodology and the feel for quality and ensuring privacy en security of the data. Some argue for a change in the statistical law, by which it becomes easier for statistical offices to get grasp of big data. It is a question if this is a realistic option?

Figure 2: Gartner's hype curve



The future will show whether these new developments will consist of a gradual change or a 'revolution'. The developments are still in their infancy. Real success stories in the domain of official statistics are still scares. This fits Gartner's hype curve. This curve states that with the introduction of every new technology the expected benefits are high. However after some time disappointment takes the upper hand, because implementation is more difficult than expected. Only in the next stage it becomes clear what the real added value is of the new technology. And, finally, of course not all statistics will and can be based on these new data sources. A combination of traditional statistics and the integration of Big data and IaD sources seems the most plausible future.