dialogic
innovatie • interactie

Internet as a datasource

Results of the consultation round

Commissioned by:
Ministry of Economic Affairs (DGET)

Project:
2008.061

Date:
Utrecht, January 12 2009

Authors:
Reg Brennenraedts
Robbin te Velde
Christiaan Holland

Index

1 Introduction

For many years, the Dutch government has been commissioning research and publishing surveys on the impact that ICT and the Internet has been having on the Dutch economy and society. New Internet applications have been developing so fast that measuring the impact of their growth has been difficult with the traditional statistical methodology.

The lack of recent data was one reason for the Dutch government to commission a piece of fundamental research to examine whether the Internet itself can become a reliable second source of information to signal trends, including their impact on the economy and society. Possible methods to do this have been examined by the research-bureau Dialogic, along with the pros and cons of the various methods available.

Dialogic has reported its experiences and findings in a final report in May 2008.

To test the merits of these rather innovative methods a consultation round has been held during Q3 2008. Via various channels, potential reviewers at home and abroad have been approached.

This report gives a concise overview of the responses we had received by the end of October 2008. The final chapter highlights the recurrent elements in the responses and sketches a way forward.

2 Written responses

The final report has been published for review on the website of the Ministry of Economic Affairs. Further publicity has been given via the network of scientific advisors at the embassies (TWA network). Visitors were invited to give comments on the methods laid out in the report, comment on how useful they will be to implement and what impact they could have on public privacy issues.

2.1 Overview of respondents

A modest number of people has responded to our call. Their affiliations are listed below. The number of contribution from Germany appears to be remarkably high.

- Bayerisches Landesamt für Statistik und Datenverarbeitung, Germany
- Bundesverband der Deutschen Industrie, Germany
- Eurostat, EU
- Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Germany
- Ministère de l'Économie, des Finances et de l'Industrie, France
- Ministry of Economic Affairs, the Netherlands
- Statistisches Bundesamt Destatis, Germany
- Statistisches Landesamt Baden-Württemberg, Germany
- Thüringer Landesamt für Statistik, Germany

For privacy reasons, in the remainder of the report the individual responses have been make anonymous (by means of aggregation).

2.2 Overview of reactions

2.2.1 General comments

Most of the respondents indicate that they have little or no experience with comparable projects. It is a genuinely new topic to them. Because of this they feel they lack background knowledge and experience to give detailed comments. Some state that their comments will be provisional and incomplete.

The quality of the research is regarded as (very) high by the respondents. The research has been thorough and is highly innovative. Most of the respondents also find the research highly interesting. It is considered to be important too since it is felt that statistical methods should be constantly refined and renewed.

2.2.2 Benefits of the method

Most respondents regard the new (IaD) methods as an interesting complement to conventional methods, especially the possibility to detect and measure trends very quickly. In the way, IaD can be used as an 'early warning system'. It could also be used to develop hypotheses that could then be tested with traditional methods. IaD is not seen as a substitute to conventional methods – at least not in the short run. But is does seem to be a useful addition.

In the original research proposal, the possible contribution of the (largely automated) IaD methods toward the further reduction of administrative burgers, was one of the key starting points. The notion is however only mentioned by few respondents.

2.2.3 Disadvantages of the method

Reduction in data quality

Many respondents seem to have a hard time to get a grip on the methodological aspects of the new methods. There is a general feeling of uncertainty about the quality of the data and the possible use of that data. Some respondents argue that the data generated by IaD methods should comply to the same quality levels as data generated by conventional methods. Others think that there is room for a (new) type of indicators that does not meet the existing standards, but has other merits instead. In both causes though, the opinion is that new indicators could never replace existing indicators. Data generated by IaD methods does not met the standard quality demands with regard to accuracy, completeness, transparency etceteras. One respondent even argues that an internet-based sample will never be representative, thus IaD methods should not be used in the first place.

Privacy

There are wide concerns about privacy. Quite many remarks have been made on this topic. It means that a lot of attention should be paid to privacy concerns during the eventual further implementation of IaD methods. One recurrent comment is that citizens might resist the use of IaD methods. It is therefore important to always guarantee anonymity and to communicate this point clearly. There might perhaps also be legal objections against certain aspects of the use of IaD. These potential hurdles have to be clearly mapped prior to further implementation.

Operational objections

A couple of respondents argue that the costs of the new methods will be considerable. They think substantial investments have to be made in knowledge and human capital before the fruits of the methods can be reaped. Furthermore, more insight is needed in the operational costs for the techniques should work in the highly dynamic context of the internet (hence might need constant adjustments).

One respondent specifically refers to its own organisation. At this moment the organisation is very busy with other matters such as the integration of statistical data from various local sources into one generic source. There is hardly any time left to spent on other tasks. There will certainly be no space to experiment with R&D-alike development such as IaD at the near future.

Side effects

Some respondents mention certain side effects of the use of IaD methods. One effect could be that citizens will become more cautious about the traces they leave on the internet. As a result IaD methods might become less usable. On the other hand, one questions whether it is possible at all to wipe out traces. Taking the first argument somewhat further, it is possible that only *certain types of citizens* (e.g., highly experienced internet users) will leave less traces on the internet. Thus the structural bias of future measurements will increase. In the particular case of the Netherlands, it is official policy of the Ministry of Economic Affairs to increase the 'digital awareness' of citizens, that is, to teach them to leave less data on the internet.

Follow-up

Several respondents have given concrete advises about the possible follow-up on the study. One option is to bring together national statistical agencies and to encourage them to explore the possibilities of IaD. Another option is that the OECD should conduct an international exploratory study on the topic. In any concrete follow-up, it is important to keep commercial and public sector applications strictly separated.

3 Presentations

3.1 OECD|WPIIS (Paris, April 2008)

Prior to the publication of the final report, the interim results of the study have been presented during the plenary session of the OECD Working Party on Indicators for the Information Society (WPIIS).[1]

The OECD has made the following report of the meeting: [2]

Internet as a data source: a project by the Dutch Ministry of Economic Affairs

1. Mr. **Pim den Hertog**, Mr. **Robbin te Velde**, Dialogic, and **Mr. Henrik Schulze**, Ipoque, presented a report on "Internet as a data source" prepared for the Dutch Ministry of Economic Affairs. The report identifies new data and indicators derived directly from the Internet (both primary data using e.g. web crawlers and secondary data) and describe the methodologies used to generate these data.[3]

2. Eurostat expressed strong interest for this project and suggested that NSOs should be involved.

3. Turkey informed that some preliminary indicators of this type would become available in Turkey soon.

4. Canada observed that privacy and legal issues prevent the NSO to carry out this type of data collection. However, the delegate observed that NSOs would be allowed to measure traffic and flows over the Internet, although not their content or source.

5. The delegate from **Denmark** found the approach interesting but he observed that some barriers that need to be taken into account. How to avoid being dependent on certain websites? How not to mention certain websites? Further, he questioned the representativeness of Internet-based statistics and whether the data collected were weighed.

6. Mr. **te Velde** acknowledged that a number of issues are still open but he invited to assess the pros and cons of Internet-based statistics against more traditional statistics.

7. Further to a question by the OECD Secretariat, **the Netherlands** informed that Statistics Netherlands was involved in this project but they considered that the statistical quality of the data collected was not sufficient for publication.

8. The **Chair** invited the Delegates to their comments on the report to the website of the Dutch Ministry of Economic Affairs.

The OECD has later given a lengthy summary and review of the final report. The author noted that the general opinion of the representatives of national statistical offices was that IaD "[is] very interesting but current statistical regulation does not allow us to collect data on the Internet."

---

[1] The original presentation (PPT) can be downloaded at from Dialogic's website:

http://www.dialogic.nl/modules/document/bestand.aspx?BID=326

[2] DSTI ICCP ISS M(2008)

[3] The main Report and Annexes report can be downloaded from:

http://www.ez.nl/Onderwerpen/Betrouwbare_telecom/ICT_beleid/Consultatie_rapport_Go_with_the_Dataflow/Go_with_the_dataflow_Main_report

The natural role of the OECD is to see whether there are possibilities to adapt legislation in such a way that the collection of internet-based data could be done will safeguarding the privacy and security of citizens.

These issues are typically dealt with by another OECD workgroup, the *Working Party on Information Security and Privacy* (WISP). The author therefore proposed to submit his paper – with additional comments from the Ministry and Dialogic – to the WISP and ask them for advices about eventual changes to legislation.

3.2 SIREN (Amsterdam, September 2008)

On 29 September Dialogic has been present with a poster presentation at the annual Scientific ICT-Research Event Netherlands (SIREN) at the Free University in Amsterdam.[4]

Most of the discussions centred around privacy issues and the quality of data. A general concern was that the government looked (too) deep into the lives of citizens. At the same time, it would be even less desirable of commercial parties would do this – and they *will*, if not are already doing it.

With regard to the quality of the data, possible problems with representativity and validity were again mentioned. Yet a possible trade-off might occur between quality and timeliness – outdated exact data versus up to date coarse data. The general conclusion was that IaD could only be used as a complement not as a substitute to traditional data. Also, the data does not (yet) met the conventional quality standard.

3.3 Eurostat (Luxembourg, October 2008)

At October 7 Dialogic was invited by Eurostat to present the results of the study during the annual plenary session of the European Information Society Statistics/ ICT Sector and ICT Investments Working Group.[5]

The study seem to be somewhat less received than at OECD WPIIS. Representatives from outside national statistical offices were more interested, e.g., in the possibilities of establishing an early warning system and in combining IaD with traditional methods. In the discussion after the presentation, the known objections of privacy, security (spyware) and the need for a clear legal framework were brought forward.

[4] The poster is publicly available. Enquiries can be made to: brennenraeds@dialogic.nl

[5] The original presentation can be obtained from Robbin te Velde: tevelde@dialogic.nl

4 Conclusions

The number of responses that has been received is fairly limited. Seemingly the use of IaD methods is new to most of the people in the field. The majority of respondents has indicated that they are not aware of any comparable initiatives. This is one of the reasons that they do not (yet) have a clear opinion about IaD. In general people seem to be interested or even intrigued by the possibilities of directly collecting data from the internet. Yet at the same time, as the Dutch saying goes, “to be unknown is to be unloved”.

Respondents appreciate most the *speed* of identification and measurements of trends. Reduction of administrative burden is hardly mentioned by individuals but is an important issue for institutions (such as OECD and Eurostat). This can be explained by the dominant view on IaD methods: they are an interesting supplement to current methods but will certainly not replace them. Boldly speaking one could say that both the *possibilities* of IaD and the *limitations* of current methods are generally underrated. Hardly nobody seems to realize that current methods are increasingly becoming useless – especially for the most dynamic parts of the economy and society.

As for the limitations of IaD methods, respondents are especially concerned about a decline in data quality and about possible privacy breaches.

Decline in data quality

Many respondents question the quality of the data that can be generated with the help of IaD methods. These objections are partly justified. A detailed discussion on the pros and cons of IaD methods has been included in the final report. The key issue is the trade-off betweeb the (dis)advantages of non-invasive, spontaneous (IaD) methods and invasive, mediated (traditional) methods. The first type of methods is generally *more* reliable than the second type because less distortion due to mediation occurs. At the same time the data generated by the first type is much more difficult to interpret. Thus the actual challenges with regard to data quality do not so much concern the *collection* of data – there are still real problems at this moment but these can largely by circumvented – but more so the *interpretation* of the data. This applies especially to the linking of data to behaviour patterns of firms and households.

Some respondents demand the same quality from IaD data as from conventional data and therefore reject the very use of IaD methods. This argument does not take into account the early stage of IaD methods. One cannot compare a beta or even a gamma version with a stable version. If no room for experiments is given, there is no space to improve the quality either.

Privacy

The challenges with regard to privacy are inherent to the non-invasive nature of IaD methods. Respondents are not bothered because data is automatically gathered. The drawback is that they do not automatically know data is being collected. The problems are real and should be addressed in a firm and proper manner. Otherwise there will probably be considerable resistance from households and firms (esp. ISPs) against the use of these new methods.

Due to the 'fear for the unknown' the privacy threat is however grossly overestimated. From a technical point of view collection data *at the level of individuals* is probably not even feasible. At the same time, organisational-wise it is very well possible to limit the collection of data to the *aggregrate level*.

It is important to clearly distinguish both levels. For example, some respondents have – rightfully – pointed at the indirect effects (at the aggregate level) of the individual responses of people. With a wider spread of IaD methods, esp, sophisticated users will mask their digital traces. Nevertheless, for statistical purposes only data at the aggregate level is of interest. Data can certainly be protected at the individual level while the aggregated data is being used for policy purposes. This is already common practise for traditional methods.

Operational challenges

National statistical offices mostly have substantial operational objectives against the use of IaD. These agencies currently often face considerable pressure to make the use of traditional methods more efficient. Because of this, there is little or no space to experiment with these type of highly innovative methods. They are *nice to have* but certainly not necessary and arrive at the wrong time. The investments that go hand in hand with the use of IaD – both in terms of training and education as in terms of capital – are also thought to be relatively high.

Given the current position of the national statistical offices these arguments do make sense. On the other hand they carry probably much less weight than the data quality and privacy issues. The rather conservative attitude of the agencies might in the longer run prove to be somewhat short-sighted and might even undermine their own position. It is a well-known fact in innovation literature that radical innovations usually come from outside the industry.

It has been suggested that the national statistical offices should play a central role in the further development of IaD. The agencies are indeed probably well-positioned to deal with the data quality and privacy issues. At the same time, it are exactly the national statistical agencies that are quite conservative in their way of working and they might even regard IaD as a threat. The happy mean would perhaps be to actively involve the agencies in the roll-out of IaD while simultaneously strive to take away their reservations. One suggestion would be to:

1. Make the agencies aware of the fact that the traditional methods are also not without flaws – and that the methods are increasingly covering less of the dynamic parts of the economy and society. The new methods are many times more cheaper and in some respects (especially speed) much more effective;
2. Facilitate the agencies in coping with the changing circumstances. Support the building up of knowledge and give ample space for experimenting with IaD;
3. Help the agencies to regain their position vis-à-vis commercial market research firms in the field of new methods. National statistical agencies have a legitimate and important role in the field of IaD. The agencies are best in collection official statistics. This should not be left to the market.

Contact:
Dialogic
Hooghiemstraplein 33-36
3514 AX Utrecht
Tel. +31 (0)30 215 05 80
Fax +31 (0)30 215 05 95
www.dialogic.nl