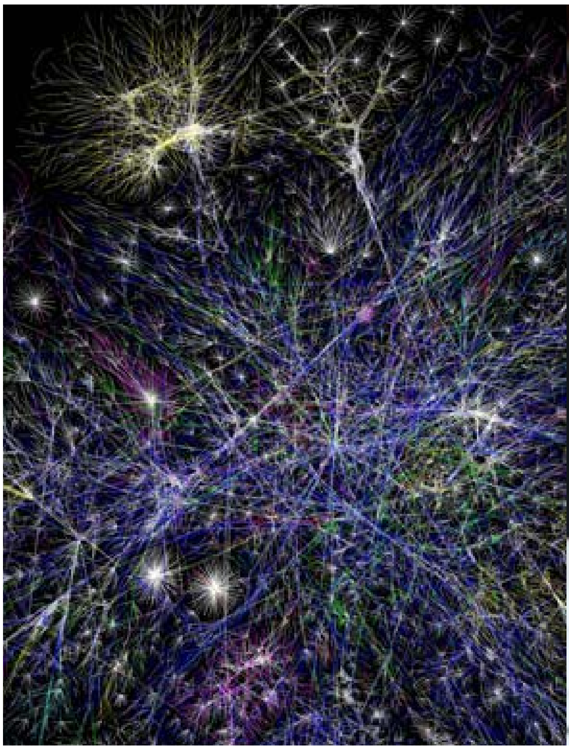
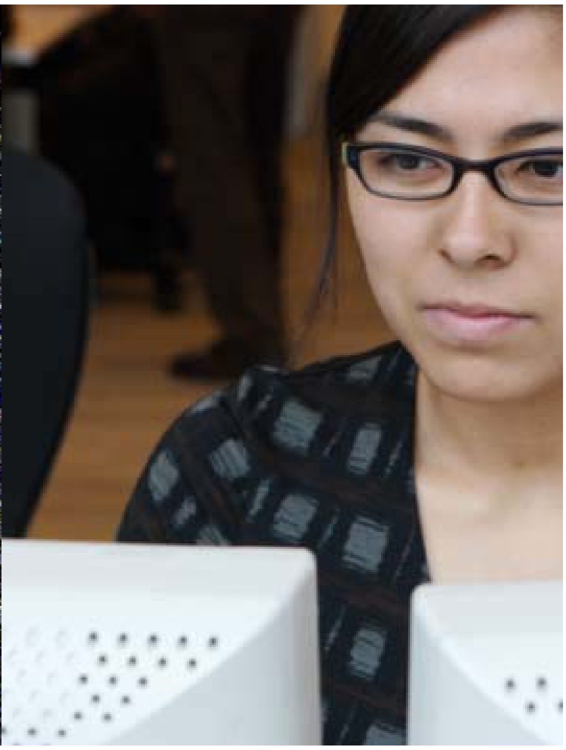
dialogic
innovatie • interactie

Ministry of Economic Affairs

Universiteit Utrecht





# Go with the dataflow!

## Analysing the Internet as a data source (IaD)

## Main report

# Go with the dataflow!

Analysing the Internet as
a data source (IaD)

Main report

measurements i.e. user-centric, network-centric and site-centric. In the course of eight case studies, we looked for data and indicators (new, extra and substitutes, if relevant) to characterize the markets concerned. We also assessed the usability of the IaD concept. This process enabled us to identify many advantages and disadvantages of the Internet as a data source.

We conclude that the Internet as a data source is a relevant method or information source for various markets (with specific characteristics). It is not only relevant for (public) policy makers, researchers and statisticians but also for market research companies, industrialists and trade organisations in the private sector.

The main advantage is that IaD has been shown to provide insight into markets and phenomena in areas where the established statistical agencies have no information. IaD provides new or enhanced insight into relevant economic and social developments, in a quick and timely (near real time) fashion. In some cases, it can act as a substitute for existing indicators and data collection methods, leading to reduced administration and lower costs. Even if the statistical quality of the data collected with IaD methods is poor, it is better to have measured relevant developments partly or badly, than not to measure them at all. Relevant, in this case, means that these new developments can generate new economic activities (e.g. new services) with a considerable economic value. Currently, policy makers largely ignore these new developments and new economic activities when using established statistical indicators. In one of the case studies we were able to make a calculation of the total amount of transactions mediated by a leading Dutch C2C online marketplace (marktplaats.nl). On the basis of their web statistics, we have calculated that the total value of transactions in 2006 is €4,7 billion, which represents a very substantial part of online consumer spending in the Netherlands.

The usability of the Internet as a data source is dependent to a large extent upon basic market characteristics. The potential gain is highest when:

- value chains are highly digitalized;
- products are digital themselves, or information about the product is highly digitalized;
- markets are dominated by a few players;
- market players are very transparent;
- markets are highly regulated
- administrative tasks are labour intensive.

The usability of IaD methods is also high when:

- government registers can be accessed that are highly digitalized and contain good quality data;
- online activities are the subject of research;
- subjects of research are highly dynamic (and annual measurements are not sufficient)
- and/or real time information about the subject is required.

Also, IaD has more potential when various methods (user-centric, site-centric, network-centric) can be combined.

Statistical work is normally done on the basis of a clear demarcation between industrial sectors. In the course of our research, we encountered markets in the digital economy that are more difficult to delineate. In fact, through the development of new business models the barriers get even more fuzzy and diffuse.

An important lesson here is that many of the respondents interviewed during case studies, do not recognise themselves in the existing statistics. They ask for new indicators and revised definitions of products and markets. This point also signals an important problem for established statistical agencies. If they do not succeed in capturing dynamic, relevant developments in the emerging digital economy, they risk being overshadowed by those that do. In other words, statistical agencies will have to come up with new methods, such as IaD, in order not to run

methods and to facilitate experiments. It is clear that using IaD is still in its infancy. We believe there is a strong need for further experimentation and research.

Each product, service, specific economic activity in the value chain and each market has its own digital footprint. They have their own typical concentration points and, therefore, provide very specific opportunities for using IaD for indicator development or for the substitution of existing statistics.

A network of researchers, policy makers and statisticians could set up a innovative research program and new kinds of publications on this subject could be initiated. Also, governments should anticipate the use of digital (re) sources for statistical purposes when developing or implementing their own registers and ICT projects.

Finally, governments need to develop a roadmap for innovative methods and innovative statistics within the publicly funded statistical agencies, as well as through organizations like Eurostat and the OECD.

will increase further as the service functionality associated with a manufactured product becomes more important, sometimes to a point where it becomes more important than the actual product itself. Digital datastreams now command very large bandwidths, some of which flow through the public Internet as well through proprietary closed networks (some of which can be accesses for statistical purposes).

- **Measure digital footprints close to the actual users.** Essentially most statistics are the sum of changing behaviour by social or economic agents. In order to get a clearer understanding of the EDE, it is important to look at the changing behaviour of the individual users (who may be companies and/or individuals). By using the (sum of) changed behaviour as reflected in the digital footprints that users leave behind, behaviour of firms and individuals can be assessed.

  Datastreams over the Internet usually start by users making a (user) request for some kind of information or content. In a way, users signal their actual behaviour (i.e. their real behaviour as well as their perceived behaviour) through their digital footprints. If performed on an individual basis this might result in "Big Brother-like" concerns, but providing that privacy is guaranteed, analysing these digital footprints may result in accurate data and indicators.

- **Combine looking at the statistical rear-view mirror with developing more forward-looking beta-indicators.** Most established statistical indicators take years to develop. Even in the most favourable cases, they become available within a few years after they are first constructed and tested. The demands for validity, robustness, the possibility for constructing time series and international comparison are all time consuming and complex. In practice this means that new phenomena, such as a various aspects of the EDE, can only be measured a few years after they were first signalled. The Internet as a Data Source approach, however, allows us to use changing patterns in the current data for constructing new indicators and assessing how a new phenomenon is developing.

  Five years ago, if it had been possible to measure more accurately the various sorts of data streams flowing over the Internet, we would have been able to predict that the video usage was about to explode. Currently, we can only conclude this (long) afterwards. There is clearly a trade off between stable, well defined, well tested and relatively less relevant indicators on the one hand and less stable, less well defined and tested but highly relevant indicators on the other. The latter could be seen as "beta-indicators" which help us to identify and measure new phenomena fast. Only a fraction of these will eventually be developed into regular high quality indicators, but in the mean time the other beta-indicators might have been very useful in pointing towards relevant trends and issues.

- **Test the applicability of IaD in both "New Economy" and "Old Economy" markets.** In early discussions on the EDE there was often an implicit comparison with the "Old" Economy. This assumed that there is a new, different set of economic rules for the new economy industries. But as digitalisation has evolved over the last few years, IaD increasingly applies to traditional economy industries and markets as well. The information streams concerning physical goods have become increasingly important. Some have developed into markets of their own such as the market for estate agents, logistics providers or information sources that track and trace the quality and origin of agricultural products.

the data packets that pass by (*data packet inspection*) and thus to look at a much more detailed level.

The final destination of the user request for a specific piece of content is always another computer. The data flow is either directed to an application that is hosted on a server (such as a web page) or directly to the computer of another user (P2P). Similar to user-centric measurements, site-centric measurements can be at the level of individual applications (spiders) or at the level of the server as a whole (*traffic monitoring*).

The usability of a particular IaD method is largely dependant on the specific research question at hand.[8] User-centric measurements are the only methods that generate detailed data at the level of individual users. Network-based measurements, on the other hand, are particularly suitable for gathering aggregated data over large user populations. Site-centric measurements are somewhere in-between. They give information about the behaviour of all users of a particular application on a particular site. Furthermore, each method has its specific pros and cons in terms of practical and statistical usability and privacy.[9] A first overview is given in figure 3.2 and further elaborated upon in the following paragraphs.

[8] See again Appendix 2 and Chapter 4.

[9] A detailed description of the statistical usability of the various IaD methods has been included in Appendix 4.

Figure 3.2: Some pros and cons of the various IaD methods

| IaD method[10] | Advantages | Disadvantages |
|---|---|---|
| User-centric (spyware & traffic monitoring at OS level) | Provides detailed insight into user behaviour on a specific application.<br>Data allows to construct user profiles<br>Low capex and opex (traffic monitoring: software firewalls provide a cheap tool for measurement)<br>High scalability due to complete control over composition of panel, same set of applications used across countries.<br>High internal validity (but underestimates shameful and/or illegal behaviour)<br>High external validity (depends on size and composition of panel) | A (costly) panel is needed. Due to privacy concern, probably difficult to find panel members<br>Due to inherently limited panel size hard to find small effects<br>Abuse by malevolent third parties is a severe security risk<br>Spyware needs to be custom-made for every individual application |
| Network-centric (deep packet inspection at ISP) | Users are not aware that they are being monitored therefore illegal and/or shameful behaviour can also be covered.<br>Highly efficient measurement method. Many users and many types of content can be measured at one place at the same time.<br>All applications are being covered (but not automatically detected)<br>Real-time measurement and size of acquired data enable detection of minor changes and trends at a very early stage<br>High scalability in technical terms – repository of re-engineered footprints can be deployed anywhere<br>High internal validity | Data does not allow to construct user profiles<br>High capex for the development of (sophisticated) equipment and high opex for constant updating the footprint repository. However since re-use of data is often a by-product of traffic optimization actual purchase costs of data might be relatively low.<br>Very hard to find ISP's who are willing to cooperate (very reluctant to place equipment at the core of their network and even more reluctant to inform their subscribers.)<br>Low external validity |
| Site-centric (Spiders) | Widely applicable. Any online data source that is accessible to a regular user is also accessible to a spider (but heavy use might cause problems)<br>Provides detailed insight into content<br>Relatively little privacy concerns.<br>Development costs of simple spiders are relatively low (but possible trade-off with higher operational costs for filtering and interpreting data). | Low internal validity due to difficulty to interpret richer content<br>Usually difficult to retrieve origin of visitors to a website<br>Development costs and operational costs of more sophisticated (e.g., adaptive and/or semantic) spiders are relatively high<br>Scalability is low because (targeted) spiders are tailor-made for a specific setting (particular site in a particular setting)<br>Rising costs due to possible technological "arms race" between site administrators and spider developers. Number of sites that are not or only partly accessible is rising (thus external validity goes down)<br>Copying large chunks of data is in conflict with Database rights (only an issue in EU) |

[10] Traffic monitoring at the operating system level and benevolent spyware are merged because there are very similar in terms of advantages and disadvantages. Traffic monitoring at the server side is dropped altogether because a server administrator who will allow this kind of measurement will probably also allow direct insight in the data.

in general also accessible to spiders. However, problems could arise due to the sometimes large amounts of data that are requested by spiders. Copying large chunks of one particular dataset could imply an infringement of database protection law.[20] More importantly, in practical terms, the inherently limited upstream capacity at the server side constitutes a bottleneck.

If a spider sends many requests for information, the speed of the server might significantly slow down –affecting all the other users who are simultaneously trying to access it. It is therefore considered to be bad practice to crawl large parts of a website on a regular basis.[21] Obviously this problem is especially relevant to targeted spiders.

[20] The legal protection of databases – next to the copyright law – is a specific trait of European Law (Directive 96/9/EC). It creates a sui generis right for the creators of databases which do not qualify for copyrights (for instance, because they are not the owners of the original data but have merely assembled the data). The database protection law is not applicable outside the EU.

[21] The problem is partly solved by using a so-called 'courtesy policy' (see Eichman, 1994). Most open source spiders have a courtesy policy built in by default. In addition to this, most web servers have explicitly declared which part of the site is accessible to spiders and which parts are not (in robots.txt). There is a growing tendency among webmasters to block ever greater parts of their website for spiders.

In the first case (webstores), for instance, the unique visitor numbers to Marktplaats.nl are directly based on the web statistics collected by the webmaster. Visitor numbers for an individual site are usually not relevant for official statistics, yet what makes the figures interesting is the dominant market share by Marktplaats.nl and the sheer volume of the transactions involved. Thus Marktplaats alone already generated €4.7 million in transactions on an annual basis. Although this number only represents the minimum market size for web stores it is the best we have got so far. It is already much larger than is often assumed.

Examples of substitutes can especially be found in the last case (pigs). The relevance of using the market price set by the German online pig exchange Teleporc is not so much in the information itself – (it can also be derived from other conventional sources – but in its timeliness. It simply provides the first price known in the market (Germany is the most important market for pigs in Europe). Other pig markets (for instance, the domestic Dutch one) often follow that price some hours or days later.

In the same case, the relevance of figures generated by the "benevolent spider" at Agrovision lies instead in the level of detail (and again partly in the timeliness). Thus they could be used as extensions to statistics that already exist. In fact, it is exactly their level of detail that represents their (commercial) value: pig farmers use them to benchmark their own production processes against others and banks use them to assess the vitality of farms that ask them for loans.

Figure 4.4 links the type of research questions addressed in the individual case studies to the resulting beta-indicators for the 8 case studies. The figure shows that we started our search for examples of indicators using IaD-methods with fairly simple research questions. At the same time, the equally simple beta-indicators that we ended up with in our small scale case studies already provide a more detailed insight into most of the markets analysed. This is mostly due to the fact that most of these markets are not well covered in regular statistics. As already noted in Figure 4.3 most beta-indicators are completely new or extra. Only a few are alternatives to already existing statistical indicators.

Using Internet as a data source also allows for a more detailed understanding of the dynamic changes taking place in more mature markets. Housing sites have rapidly become a key marketplace where supply meets the demand for property and, increasingly, relevant additional services. With regard to the pig market, we are better able to understand the growth in electronic trading and a very detailed insight into the administrative processes surrounding both pig farming and trading. In both cases very detailed, timely statistics and indicators can be built using Internet as a data source.

It is clear from performing the 8 case studies that there is no "one size fits all" approach when it comes to measuring the various parts of the EDE. We also learned that the added value of using IaD methods differs considerably per industry and market. We found that the latter were mainly caused by differences in key market characteristics, such as the level of digitalization in the value chain, level of market concentration, level of regulation and geographical scale. Additionally, the availability of good quality digital concentration points also impacts on the usefulness of using IaD methods and eventually the value added by IaD in a particular market. These are points where a datastream can be tapped in order to generate data that can be used for developing relevant specific market indicators. However, in this case, the availability of a few good concentration points is to be preferred above numerous less suitable or partial concentration points.

Figure 4.5 gives an overview of the nature of the product (is it digital or not), some of the key market characteristics just mentioned and a separate score for the overall usefulness of the IaD methods, regardless of the market characteristics. The latter score is based on the experiences gained performing the case studies.

When considering the nature of the product (first row), we only differentiate between digital products (black square) and analogue products (white square). However, we observed that as analogue products information is sometimes abundantly available, the opportunities for using IaD are not necessarily small. On the contrary, the markets for pigs and houses can be assessed quite well using IaD-methods. We scored the market characteristics qualitatively, on a five-point scale and the availability of concentration points. A higher score means that IaD methods are more suitable in that particular case. The idea is that a higher level of digitalization leads directly to a higher probability of using IaD methods successfully (i.e. more potential places where IaD methods can be used).

Markets with higher levels of regulation are in a similar vein. They coincide with more accountability and hence registrations as well as higher levels of market concentration. In concentrated markets a market can be measured relatively more easily by following the main producers. A higher availability of concentration is regarded to be more suitable to be measured using IaD methods. For example, both markets for houses and social networking are highly concentrated (with dominant market shares for Funda and Hyves respectively) but in the latter case there are many more places where information can be found on social networking (that is, online social networking sites are just one of the many places where people meet). In the housing case Funda is really the prime site to find statistics on the property market in the Netherlands.

In a similar vein, the geographical dimension is inversely related to the feasibility of using IaD methods as it is difficult to tie statistics to the Dutch situation when online demand and supply are dispersed across the world (as is the case with online gaming and recorded music). Geographical borders hardly matter anymore. If they still matter, this is usually due to the relevance of language (e.g., in the case of some social networking sites – the language of communication on Hyves is Dutch thus the vast majority of users are Dutch (which does not necessarily mean they are resident in the Netherlands). Some gaming sites, especially casual gaming sites, are also relatively strongly tied to national regions.[29]

In Figure 4.5, the eight cases are ranked (from top to bottom) on the degree to which they are suitable IaD measurement methods. The ranking is based on the simplest heuristic, that is, adding the scores of all rows without attaching weights to the individual rows.

[29] Which probably explains why the Dutch Spill Group – one of the major operators of casual gaming sites in the world – was willing to pay considerable prices for 'local' domain names such as jeux.fr (France), juegos.com (Spain) and games.co.uk (UK).

phenomena - are fuzzy. They are hard to link to a single set of established industries, markets or actors. Is Internet TV still part of the broadcasting market?

Is the electronic trade in music still part of the music industry? Are social networking sites developing into a specific industry? Is the online gaming industry a submarket within the market for game consoles, the product software market or the market for electronic creative content? Does marktplaats.nl carve out a complete new type of industry or does it simply compete with regular retailing outlets and flea markets? Statistical work is normally done on the basis of a clear demarcation of industry sectors. What we encountered is that EDE markets, through the development of new business models, are more difficult to delineate. In fact, webstores are no longer limited to the retailing industry. Music is sold and spread through a wide number of channels by an increasing number of actors and C2C informal "markets" coexist alongside B2C markets. An important lesson here is that many of the respondents in the various cases, do not recognise themselves in the existing statistics. They ask for new indicators and new definitions of products and markets.

**Use of IaD concentration points (see 4.2) have been useful in some, but not all case studies.** A seventh lesson is that, in terms of our selection of market types, the C2C-market is more and more important. This is evident in the markets for online music, Internet TV, marktplaats.nl and social networking, but also present in markets such as the housing markets. In some cases B2C, B2B and C2C are hard to disentangle such as in online music and marktplaats.nl. In addition, we noticed that most data sources could be traced at the beginning and end of the value chain (information gathering and fulfilment). The latter is mainly for products that are delivered electronically, as some products (houses, pigs) cannot be delivered over the Internet. It proved much harder to detect electronic data on ordering and payment. The data is not absent, but in practice it is more difficult to get access to, since it is often covered by confidentiality agreements.[32]

**Technical and practical availability of digital sources for third parties may differ considerably.** An eighth lesson is that the availability of digital footprints does not imply these are readily available for statistical purposes. On the contrary, getting access to these types of data, which are mostly collected for other purposes, can be very difficult in practice. Most actors involved have no direct need for nor a stake in the production of statistical indicators. A lot of the data or information we wanted to use in the various case studies was simply not accessible. This could be a matter of privacy, market-sensitive information, or because relevant actors (public or private) had sealed off (or exploited themselves) possible data sources. For example, we found it much harder to measure typical B2B-markets using the Internet as these may use proprietary networks and are generally harder to trace.

**Added value of using IaD may be higher in newly developing markets.** A final lesson is that by starting from fairly simple research questions, and by using Internet as a data source, we were already able to generate new or extra indicators, especially in those markets that are not properly covered by existing statistics ("blank spots"). The advantages of E-data, as we call it, are especially relevance, richness, timeliness/speed, limited

[32] It would be a useful experiment to see what type of statistical indicators could be based on the payment or the various payment systems. The growth of electronic payment systems typically used on the Internet – preferably split between the various markets and industries – would already provide interesting data on the further growth of the digital economy. Similarly, it would be good to know to what degree electronic information searches also result in electronic or analogue ordering.

places and what it will mean to the development of electronic payment systems, the new types of logistics that emerge (micro logistics), the impact on regular retailing and so on and so forth. Some of the cases also point to new forms of illegal or at least "grey" economic activities such as the use of P2P networks for distributing audio and video.

We also considered the wider impact digitalization has on innovation. In most of the case studies analyzed we came across new types of services or business methods that can often be applied to other domains as well. Consider, for example, what the availability of high quality digital maps has done to property sites and how it will affect various other markets ranging from the hospitality industry to social networking.

Finally, digitization impacts considerably on the ways in which consumers spend their time. The effects of all kinds of electronic markets and communities are not limited to the economic realm but are firmly based in the social realm (SNS). Especially the young "digital natives" have very different communication patterns and consequently different ways in which they spend their time. There are also effects noted in the political realm (e.g. the use of Facebook in US primaries, or the speed at which protests against the Burmese military were organized). In several cases the traditional barriers between the social and economic realm are blurring. There are also numerous examples where developments that have started in the social domain turn out to have a substantial economic value.

## Colophon

Published by
the Ministry of Economic Affairs.

*Authors:*
Reg Brennenraedts (Dialogic)
Christiaan Holland (Dialogic)
Ronald Batenburg (Dialogic, Utrecht University)
Pim den Hertog (Dialogic)
Robbin te Velde (Dialogic)
Slinger Jansen (Utrecht University)
Sjaak Brinkkemper (Utrecht University)

The Hague, April 2008

More copies can be ordered via
www.ez.nl/publicaties
or +31-(0)70-3081986
or 0800-6463951 (within the Netherlands only).

## Information

Directorate-General Energy and
Telecommunications

P.O box 20101
2500 EC The Hague
The Netherlands
Internet: www.ez.nl

Publication number: 08ET12