dialogic
innovatie • interactie

Ministry of Economic Affairs

Universiteit Utrecht

Go with the dataflow!

Analysing the Internet as a data source (IaD)

Annexes

Go with the dataflow!

Analysing the Internet as a data source (IaD)

Annexes

Contents

Overview

The R&D project Internet as a data source began in May 2007 and data gathering took mainly place in the second half of 2007. In consultation with the principal and a steering committee overseeing the work, the following activities were performed, mostly in parallel:

- Conceptualisation and definition of methodological approach, including a paper outlining the usability of spiders/webcrawlers for gathering data and building statistical indicators.

- Analysis of international sources using IaD, both established indicators on the emerging digital economy and (mostly) examples of studies (scientific, market research firms, software firms) using IaD-methodologies. An overview of the sources identified – partly with the help of the international contact network of Statistics Netherlands through which various very useful suggestions were made – is given in table B below.

- Conducting 8 case studies in 8 different markets to assess the feasibility for gathering statistical data and developing indicators using IaD methods. In order to focus this quest for data, we used some very practical research questions (see figure 4.4 in the main report). Case study work involved desk research (including contacting providers of regular statistics, where relevant) i.e. looking for established and new statistical

Annex 1
Methodological approach

sources to describe the economic dynamics taking place within these markets dynamics, in-depth interviews with 3-5 key players in these markets (typically at the level of board of directors or strategy departments) and some experiments with spiders (see below). Case studies were performed in the following markets:

1. webstores and the leading electronic marketplace in the Netherlands (Marktplaats.nl)
2. product software market
3. recorded music market (mostly online)
4. Internet TV market
5. online gaming market
6. the market for social networking
7. housing market
8. pig market

- Each of the case studies was reported separately using a fixed format. All cases include a description of the structure and dynamics of the market at hand, an overview of concentration points (see figure 2.2 main report for an example), an overview of established statistical sources and possible candidates using IaD methods (beta-indicators) and the results of the spider experiment if applicable (see table A below). All cases are available in the Dutch language and two (music, gaming) are also available in English.

- Further considerable efforts were invested in organizing a network-centric measurement in the Netherlands using state of the art deep packet inspection software from the German firm, Ipoque. The various actors involved met and discussed in very practical terms the possibilities and limitations of performing such a measurement. This proved to be very informative when assessing what can be done using advanced network-centric measurements. Box 1 in Annex 2 summarizes some pros and cons of deep packet inspection.

- In the final stage of the project both discussions in the project group as well as with the steering committee proved instrumental in answering two questions: what did we learn from the case studies (1) and what are the pros and cons of using the various IaD-methods (2). We have in parallel analysed the material adopting a methodological, a statistical, a practical and a strategic policy perspective.

Below we reflect briefly on the approach adopted when using the spiders.

Spider approach

For this research several spiders and web scrapers were designed and developed. During this research, we used three different

Table A Spider subprojects within the IaD project, type of spider used and development time

Spider project number	Spider	Type
1	Vacancies	Distributed Webscraping
2	Housing market	Distributed Webscraping
3	Bol.com (major webstore)	Single site scraping
4	Hyves.nl (major social networking site)	Single site scraping
5	Product software	web spidering
6	Marktplaats.nl (major C2C electronic marketplace)	Single site scraping

approaches: distributed web scraping, single site scraping, and web spidering. The software was developed in Python, C Sharp and DotNet[1] A total of six spiders were developed.

Distributed Web Scraping

For two projects the question was whether an abstract indicator could be developed that could potentially compete with current statistics. To establish such an indicator, a number of leading sources were appointed. We scraped these sources in a distributed fashion to gather data from different sources.[2]

For project 1 a scraper was developed that could find the number of jobs on a site, simply by copying the number from a front page or from a search report. We included the top 9 job sites. These numbers were updated at real-time on all sites. The scraper would download the number of jobs posted on a daily basis.
We stored 120 reports from 9 sites over a period of 4 months. We were sure to include different types of job sites such as high-level jobs (intermediair, volkskrantbanen, etc) and low-level jobs (vacaturekrant, werk.nl, etc). These numbers were laid out in a graph to see whether trends could be discovered. For future work, sector specific information would make this a very sensitive instrument.

For project 2 a scraper was developed that could look on a real estate site and download different houses with different characteristics. These characteristics were price, postal code, and house type. Unfortunately, these characteristics were not similar on all sites. Some sites for instance, did not include the house category "parking place". Also, we do not know whether the price classes we picked were representative (€50.000-150.000, 150.001-250.000, 250.001-350.000, 350.001+). We downloaded a total of 2074 houses from two websites. These houses were checked on a weekly basis to see how quickly they would change status, over a period of 4 months.

Single Site Scraping

For three projects we scraped single sites. The main aim of these single site scrapes was to find specific data on a specific subject, such as the average price of CDs at bol.com, or the number of active Hyves users. The scraping libraries used in the distributed site scraping projects were reused extensively.

For project 3 a simple scraper could be used that automatically fills in data in form and that navigates through the site. For project 4 (Hyves) we defined a spider policy that describes how the spider should surf the website, since it would be exploring a social network. The spider would both download people's pages by doing sequential searches ("aaa", "aab", "aac",..) and by exploring the networks of the people whose pages had been downloaded. A total of 3.000 albums were downloaded from Bol.com. A total of 20328 people's pages were downloaded in project 4 from Hyves.nl.

Spidering

The final project consisted of downloading the content from 924 product software vendor's websites and doing linguistic engineering on these pages. The product software vendors were selected from a fixed list (Automatiseringsjaarboek 2006). The downloaded data was scanned for occurrences of keywords to establish the number of employees in a company, the number of countries in which the vendor is active and the type of products a software vendor builds. The final database comprised 10Gb. This project was

[1] The release of the hitherto geographically bound 'pig rights' is a strong driver for consolidation in the big breeding segment and for the establishment of enormous 'big flats'.

[2] The following libraries have been used: ClientForm (a library that automatically analyzes a form and posts data to that form), SQLite (a quick to deploy database platform), RE (Python regular expression matching module).

problematic. Sites often did not adhere to the W3C HTML standards, causing many sites to be downloaded but processed incorrectly. Furthermore, the use of Javascript restricted the pages that could be spidered. Finally, several times the results proved to be inconclusive and inconsistent, for instance if a site lists that it has both 3000 employees (working on project X) and 5000 employees (working in the company).

Research Problems

Spidering is not a trivial task. During this research many problems were encountered. Some problems were false positives for postal codes, the crashing of a spider, or significant changes to website layouts that were spidered regularly. Another example is when Volkskrantbanen cleaned up their database records and eliminated 20% of the entries. Even though nothing had happened in the real world, our spider detected a rapid decline in jobs.

Table B: International sources identified

Organization	Source/link	Description	User - Site - Network
European indicators, cyberspace and the science-technology-economy system (EICSTES)	http://www.eicstes.org/	EICSTES (research programme) intends to offer statistics and to obtain indicators on the European Science-Technology-Economy System in Internet in order to shed some light on the likely relationships between the R&D sector and key actors of the New Economy. These indicators will be disseminated in an open, user-friendly, graphical environment using new web visualisation techniques.	Combination of User - Site - Network
Internet World Stats	http://www.Internetworldstats.com/	Internet World Stats provides data concerning the use of Internet (by regions) and an analysis of the growth of Internet. The Internet usage information displayed comes from various sources: mainly from data published by Nielsen//NetRatings and by the International Telecommunications Union (ITU). Additional sources are Computer Industry Almanac, the CIA Fact Book, local NIC, local ISP, other public and private sources, and direct information from trustworthy and reliable research sources.	Combination of User - Site - Network
WISER (Web indicators S&T&I research)	http://www.webindicators.org/, http://www.virtualknowledgestudio.nl/	WISER is a research programme focussing on the increasing part of on-line scientific communication and research, which is not (or only incompletely) visible in traditional S&T indicators. WISER explores the possibilities and problems in developing a new generation of Web based S&T indicators. Web indicators should produce information about visibility and connectivity of research centres forming a common EU research area; innovations and new research fronts reached by e-science; about equal rights access and participation on e-science gender and regional.	Combination of User - Site - Network
Clickz	http://www.clickz.com/stats	Clickz collects and presents facts, figures, research, and data on every facet of the online industry, domestic and worldwide.	n.k.

Organization	Source/link	Description	User - Site - Network
Electronic commeRce Measurements through Intelligent agentS (ERMIS)	http://cordis.europa.eu/data/PROJ_FP5/ACTIONeqDndSESSIONeq112422005919ndDOCeq778ndTBLeqEN_PROJ.htm	ERMIS (research programme) is confined in the domain of new indicators for consumer oriented electronic commerce. The aim of the project is to design, develop and validate an integrated system for the efficient statistical measurement and monitoring of E-Commerce, that will be able to provide the final "information consumers", i.e. decision makers in the public and private sectors of the economy, but also the European Citizen the efficient means (i.e. a set of indicators and an appropriate conceptual framework for their interpretation) to assess developments and risks in this rapidly changing field. The target population quantification and the calculation of indicators will rely mostly on data captured from intelligent agent in the WEB.	n.k.
BigChampagne	http://www.bigchampagne.com/	BigChampagne is a research company specializing in data concerning peer-to-peer (P2P) networks and intelligence about media consumption (consumer behaviour). Information about media consumption is collected, aggregated and analysed. Data is is provided by web communities, retail partners, our strategic partners at Mediabase. Furthermore, P2P network data is used.	Network
Cooperative Association for Internet Data Analysis (CAIDA)	http://www.caida.org/	CAIDA gathers Internet data from and across a wide variety of Internet infrastructure, including commercial, educational, research, government, and exchange point links. Collected data is analysed in order to better understand current and future network topology, routing, security, DNS, workload, performance, and economic issues.	Network
Digital Era Statistical Indicators (DIASTASIS)	http://www.eurodyn.com/diastasis/	DIASTASIS (research programme) aims at defining, measuring and exploiting new socio-economic statistical indicators by combining: i) household research data and statistical data on SMEs; and ii) data on the use of the Web referring to the same base of households and SMEs. The new statistical methodology to correlate these two distinct data sets will be implemented on an information system, which will be demonstrated and assessed during its pilot operation. Statistical data on Web usage will be obtained from Internet Service Providers (ISPs) by using new technical means capable of gathering statistical data while ensuring protection of personal data (network-centric).	Network
Ellacoya	http://www.ellacoya.com/	Ellacoya is a leading provider of carrier-grade service control solutions that give broadband service providers full service control functionality for their networks. Its IP Service Control System identifies subscribers, classifies and controls applications on a per-subscriber basis, improves performance and customer satisfaction, and delivers revenue-generating IP services.	Network

Organization	Source/link	Description	User - Site - Network
Hitwise	http://www.hitwise.com/	The Hitwise online competitive intelligence service provides daily insights on how 25 million people interact with over 1 million websites in 160+ industries. Hitwise makes use of a network-centric model: data is sent to Hitwise from the ISPs including page requests, visits and average visit length.	Network
Narus	http://www.narus.com	Narus specialises in network traffic with full correlation between all the other elements on the network (routers, firewalls or IPS/IDS), across all of the links on the network as well as external storage facilities to access historical data.	Network
Netcraft	http://news.netcraft.com/archives/netcraft_services.html	Netcraft developed a toolbar for individual users, which provides information about the security and reliability of web sites. Furthermore, data is given on market share of web servers, operating systems, hosting providers, ISPs, encrypted transactions, electronic commerce, scripting languages and content technologies on the Internet.	Network
Packet Clearing House	http://www.pch.net	PCH developed a database of Internet topology measurements (operating since 1997). This archive of routing data from all major and many minor Internet provider networks is available to academic and commercial researchers and the operations community, to aid in the understanding of the dynamic nature and topology of the Internet. A network-centric approach is taken.	Network
Packeteer	http://www.packeteer.com/	Packeteer uses a unique intelligence about the application and the network to target the optimum technologies for the ultimate performance gain and extend the value across current and emerging enterprise initiatives.	Network
Awstats	awstats.sourceforge.net	AWStats is an open source Web analytics reporting tool, suitable for analyzing data from Internet services such as web, streaming media, mail and FTP servers. AWstats, just as Webalizer, analyses server log files and produces HTML reports.	Site
Chemconnect	http://www.chemconnect.com/	ChemConnect has established itself as a leading independent and neutral 3rd party commodity exchange, auctions provider, bulletin-board, back-end fulfilment service and market information source for NGL's, chemicals, feedstocks, polymers, fuel oil, and much more. Data provided on Chemconnect says something about (international) commodity trading.	Site
Elemica	http://www.elemica.com	Elemica offers total solutions focused on improving supply chain inefficiencies (Chemical industry). Offering a "one-stop" experience through browser-based and Enterprise Resource Planning (ERP) connectivity, Elemica represents an outstanding level of commitment and coordination.	Site

Organization	Source/link	Description	User - Site - Network
Google Analytics	http://www.google.com/analytics/	Google Analytics gives a site-centric analysis of Site usage (just as Webalizer and Awstats). Google Analytics has been mainly developed and designed to help you learn more about where your visitors come from and how they interact with your site.	Site
Google Trends	http://www.google.com/trends	Google Trends analyzes a portion of Google web searches to compute how many searches have been done for the terms you enter, relative to the total number of searches done on Google over time. We then show you a graph with the results -- our search-volume graph -- plotted on a linear scale.	Site
Innocentive	http://www.innocentive.com/	Innocentive provides a market place for R&D topics/ resources/ collaborations (Open Innovation model). Numbers of subscribers (Seekers and Solvers) could be taken into account as an indicator of R&D activity.	Site
NineSigma	http://www.ninesigma.com/	NineSigma enables clients to source innovative ideas, technologies, products and services from outside their organizations quickly and inexpensively by connecting them to the best innovators and solution providers from around the world (Open Innovation). Exchange and contacts presented on NineSigma could be used as an indicator of R&D activity.	Site
Webalizer	http://www.mrunix.net/webalizer/	Webalizer developed a web server log file analysis program, which produces Internet usage reports. Webalizer is meant for single server analyses and does not publicly present figures on an aggregated level.	Site
Yet2	http://www.yet2.com/app/about/home	yet2.com is focused on bringing buyers and sellers of technologies together so that all parties maximize the return on their investments. Yet2.com offers companies and individuals tools and expertise to acquire, sell, license, and leverage some of the world's most valuable intellectual assets. The concept of yet2.com looks at the concept of Innocentive. Exchange and contacts presented on Yet2 could be used as an indicator of R&D activity.	Site
StatMarket/ Hitbox	http://www.websidestory.com/products/web-analytics/datainsights/statmarket/overview.html/	StatMarket provides market share data on which browser versions, operating systems and screen resolutions web surfers are using worldwide. StatMarket information is based on a web browser interface.	Site - Network
Alexa	http://www.alexa.com/	Alexa has developed an installed based of millions of toolbars (user-centric), one of the largest Web crawls and an infrastructure to process and serve massive amounts of Internet usage data. Alexa's Toolbar provides a newly and innovative Web navigation and intelligence service for personal users. Collected data gives an overview of Internet usage behaviour (Marketing data).	User

Organization	Source/link	Description	User - Site - Network
Comscore	http://www.comscore.com/	ComScore provides real-time measurement of the myriad ways in which the Internet is used and the wide variety of activities that are occurring online. ComScore measures both offline and online activities in a user-centric manner and uses the Internet as a timely and powerful data collection medium (browsing behaviour).	User
Nielsen/ Netratings	http://www.nielsen-netratings.com/	Nielsen//NetRatings provides panel-based and site-centric Internet usage measurement services, online advertising intelligence, user lifestyle and demographic data, e-commerce and transaction metrics, and custom data, research and analysis.	User
Ranking.com	http://www.ranking.com/	Ranking.com calculates the online popularity of the most visited websites and provides these results free to the World Wide Web. Ranking.com makes use of statistics concerning unique visitors, page views and link popularity. A user-centric measure is used (installed software on a personal computer).	User
Statistical Indicators Benchmarking the Information Society (SIBIS)	http://www.sibis-eu.org	SIBIS (research programme) addressing innovative information society indicators to take account of the rapidly changing nature of modern societies and to enable the benchmarking of progress in EU Member States. These indicators have been tested and piloted in representative surveys in all EU member states, 10 Acceding and Candidate countries, Switzerland and the USA.	User - Site
Technorati	http://technorati.com/about/	Technorati is a leading company in the analysis of Live Web (dynamic and always-updating portion of the Web: mainly weblogs, social media, etc). Technorati searches and analyses blogs and the other forms of independent, user-generated content (photos, videos, voting, etc.) increasingly referred to as "citizen media." Technorati works with browser buttons, blog widgets, search plug-in and pinging.	User -Site

Annex 2
Detailed description of IaD methods

The IaD methods are described in detail in this annex. Section A gives a global overview of the seven IaD possible methods we identified: web surveys, benevolent spyware, traffic monitors, deep packet inspection, benevolent spiders and in-depth data analysis. Section B will present a taxonomy of IaD methods. Finally, Section C describes the four most interesting methods in detail.

Section A: Possible methods

1. (Web) surveys - user-centric

Asking questions to a representative set of Internet users in the form of a survey is a primitive way of gathering information regarding online behaviour. It can be conducted by the traditional (paper) questionnaires, but a web survey can also be used. The major advantage of using this method is the fact that the researchers know the user characteristics. Therefore, conclusions concerning subsets of the population can be drawn. Furthermore, the sample can be compared (and thus weighted) to the total population. A major disadvantage of this method is the relative high cost. This is due to the fact that the Internet user has to transform the analogue data (in his memory) into digital data.

There are many applications on the Internet offering web survey functionality. Some companies offer a hosted application and researchers only have to upload their (Word) questionnaire to obtain a fully web survey. There are also companies who structurally use web surveys and a panel of respondents to obtain longitudinal data, e.g. Peil.nl and TNS-NIPO.

2. Benevolent spyware - user-centric

Benevolent spyware can be used at the level of applications. The scope of the objectives of applications is very wide: There are applications allowing users to read and write e-mail, view websites (browsers), listen to web radio, view web-TV, use a Usenet server, share document with other users (P2P), play games, make phone calls, et cetera. In this context we are not discussing the regular malevolent spyware, but a more benevolent spyware. By this we mean that the use of the spyware is approved by the end-user and causes no harm to their system. It is a less invasive and more efficient way to obtain data than to ask users questions regarding their online behaviour. Of course, generalization of data is only possible if a p anel with well-described user characteristics is used.

When implementing spyware, user behaviour on a specific application can be monitored. In its original form, spyware is often designed to realize targeted advertisement. For example, if the spyware "sees" that a user lives in Tokyo and often visits web pages dealing with German cars, it can use this information to display advertisements for a Tokyo-based German car dealer. Spyware is often associated with browsers, especially Microsoft's earlier versions of Internet Explorer. But spyware can also be used in Instant Messaging plug-ins (e.g. Messenger Plus! Live), P2P applications (e.g. Kazaa), Media players (e.g. RealPlayer),etc. Basically, spyware can be developed for any application.

3. Traffic monitors - user-centric

Traffic monitoring can be used in the user-centric domain on PCs (hardware) and operating systems (software). When using traffic monitoring, all the communication between a user (or a private network of users) and the Internet is analyzed. This process can be conducted at the level of the operation system and a generic piece of software has to be installed on the computer. However, the condition has to be met that every user can be measured separately. The different user accounts in Windows XP, Linux and MacOS offer possibilities for this. It differs from the benevolent spyware by monitoring all the traffic from and to the Internet. On the other hand, the measurements are more generic and it is harder to obtain in-depth insight into the behaviour of the user. The measurement can also be conducted at the level of the personal computer. By placing a hardware device on the line between PC and the public Internet, traffic can be monitored. Making distinction between different users of the PC will be harder when using this method. Obviously, it is also possible to use this method to monitor the traffic between an office (or home) network and the public Internet. This is a hybrid of a network-centric and a user-centric measurement approach.

The traffic monitors we are presenting are very similar to firewalls. Firewalls are designed to regulate the traffic between a computer (network) and the Internet. To do this, they monitor traffic and block unwanted traffic. With the exception of blocking the traffic, this method roughly needs the same functionality. Furthermore, firewalls can be implemented on the OS (e.g. ZoneAlarm, Windows Firewall, et cetera) and by a hardware device (e.g. produced by Cisco, Juniper, NetAsq, et cetera).

4. Deep packet inspection – network-centric

Network-centric measurements focus on the traffic flow between many users and many suppliers of content. This connection between users and content is supplied by ISPs. Because of the limited number of Internet Service Providers (ISPs) a natural focal point occurs. Measurements can also be conducted in data centres. Almost every professional company active on the web houses its equipment in data centre. Several of these centres sell their "footprints" on the market and are not dedicated to one supplier.

The essential technology needed for measuring at network-centric level bears massive resemblance with firewalls. However, network-centric measurement equipment exists that is able to give a more in-depth analysis of the data flow. They are able to see which users (IP-addresses to be more specific) are communicating, how much data is exchanged, which protocol is utilized, which application is used and, sometimes, even properties of the content itself (e.g. video, audio, etc.). Moreover, in comparison to firewalls, the amount of data to be analyzed is much higher.

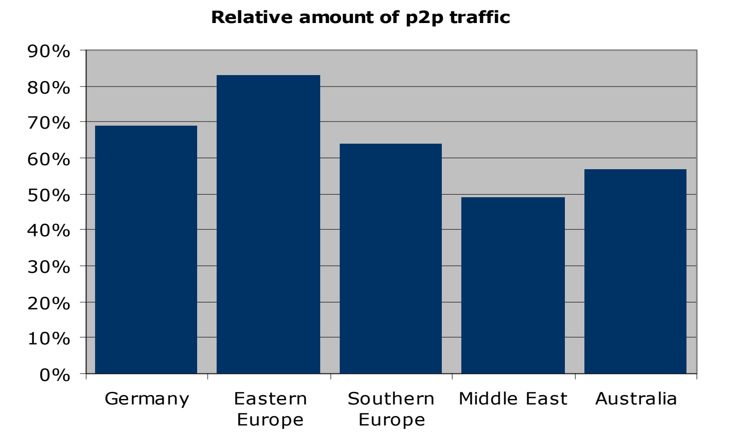
Box 1 presents some examples as well as some pros and cons of deep packet inspection.

Box 1: The possibilities of and barriers for deep packet inspection[3]

The Internet is constantly growing, both in terms of quantity (total amount of data traffic) as in term of quality (amount of different protocols used). This makes data monitoring a more important issue with every passing day for Internet service providers and network administrators. It enables them to optimize the flow of traffic over their networks or enhance the security of the network. A familiar example of traffic monitoring *(network-centric measurements)* is a firewall: a piece of (embedded) software able to inspect and block traffic. The first generation of firewalls analysed data traffic by looking at the packet headers. The content of the data remains unknown to them. Current applications are far more advanced and are able to actually look into the data *(deep packet inspection)*. And because most applications generate specific pattern in the data traffic, it is also possible to track applications that are not found by looking at metadata in the header of the traffic packages it generates. Moreover, several advanced methods are also able to examine specific behaviour within an application. By using these methods, it is possible to detect different protocols, even secured protocols. In some (P2P) protocols it is even possible to detect the file name and file size. This allows researchers to obtain in depth insight into the transferred data (e.g. the amount of video in P2P traffic or even the frequency of a certain film title in the traffic). The figure below is constructed by using data stemming from this method. It shows that a very large part of the Internet traffic in different regions consists of P2P traffic.

[3] A recent discussion on Internet topology measurements and their relevance for policy issues is included in Kaart et al. (2007)

Figure A: Relative amount of P2P traffic in the total amount of traffic (%) 2007[4]



Most applications using network-centric measurements are primarily focussed on network management and to resolve security issues. But as a side effect, they end up producing very interesting data that has potential use for statistical purposes. At its lowest level, the data could be used to examine the use of different kinds of protocols, as can be seen in Figure A. A shift in the use of protocols –which can be seen real time- often signs new trends, like the increase use of direct download links (e.g. RapidShare and Megaupload) instead of P2P applications. On a more detailed level, the rise and fall of different websites (e.g. YouTube) can be monitored. When focusing on P2P applications, even the number of downloads of a specific file (movie, music, software) can be monitored. Data shows that the top 10 BitTorrent Video in Germany for 2007 were the following: Next; Fantastic Four – Rise of the Silver Surfer (German edition); Spiderman 3 (German edition); Naruto Shippuuden – 020; 300; Private Mystery Island XXX; Die Simpsons – Der Film; White Noise 2 – The Light; Sterben für Anfänger; JTC – Mr Mrs Sexxx. [5]

Network-centric measurements have some practical challenges, aside from the obvious difficulties surrounding privacy. First, the decentralized structure of the Internet (the TCP/IP protocol) results in no single points of failure. Traffic always goes from one point to another by taking the best route at that time. However, when measuring data this decentralized design is a major obstacle. Therefore, network-centric measurements are measurements in a unstructured datacloud. But, when the point of measurement is chosen strategically, the method allows one to measure very small differences in the use of Internet in real time. These measurements are not representative for the Internet as a whole, but only for the subnetwork that is being measured.

A second practical obstacle when conducting network-centric measurements is related to network access. One always has to place some equipment directly on the network that is the subject of the research. However, the people responsible for network management are very reluctant to allow third parties access to their networks. Network performance, privacy and security could be important argument for this behaviour. But they also fear the outcomes of the research, showing that a very large of percentage of the traffic consists of P2P. They fear that this will provide governments and copyright organisations with arguments and evidence to force network administrators to shape the traffic.

[4] Ipoque (2007). Internet Study 2007.

[5] Ibid.

5. Traffic monitoring – site centric

Due to the fact that our model is mainly symmetrical, traffic monitoring on the site-centric site looks much like traffic monitoring on the user-centric site. In fact, for the P2P application, the methodology is exact the same. This is because by its nature, the P2P technology lacks traditional servers. The other part of the site-centric model deals with servers and operating systems that provide services to a large set of different applications. Here, measurements can be conducted by a hardware Firewall (at server level) or a software firewall (at OS level). Contrary to the user-centric domain, in the site-centric domain, servers (and thus also their operating systems) are predominantly used for only one application.

6. Benevolent spiders – site centric

Benevolent spiders are the opposite of benevolent spyware.

- Benevolent spyware installs an application at the user side (consumer) and uses the Internet to transfer collected data back to the server of the spyware "owner".
- Benevolent spiders are applications that are installed on the server of their owner and use the Internet to connect to (and harvest) online data sources.

For example, a benevolent spider could be used to download all the advertisements on several online marketplaces concerning green Alfa Romeo's built between 1960 and 1970. Basically, a spider runs a script that shows it the way to different data sources and it downloads specific information. In fact, search engines like Google can also be seen as spiders. They simply visit every page on the Internet and download its text.

The main feature of benevolent spiders is their ability to obtain very specific data. The massive amount of freedom in implementation allows researchers to built spiders capable of finding information on almost every topic present on the Internet. The use of benevolent spiders is hard in dynamic markets where there are a high number of suppliers, due to the very focussed nature of spiders. When designing a spider, we should pay attention to design a benevolent spider. For example, the spider should not weaken the performance of the server by sending too many requests for information. Furthermore, organizations that object being "spidered" should not be approached.

7. In-depth data analysis – site centric

An in-depth data analysis can only be conducted if the researcher has full and unrestricted access to a complete data set. Most Internet applications offer a gateway and an interface to obtain data, but they cannot be said to offer unrestricted access to the data. For example, it is usually possible to find the phone number of a certain person on the Internet, by using one of the many (national) online phone books. But it is usually not possible to do a reverse look-up and find out the name behind a certain phone number, or to give a full overview of all the registered phone numbers in a certain street, etc. Basically, the application layer limits the access to the database. In some cases a spider could use the application layer to obtain a full dataset, but in many cases this option is not possible due to security measures (usually in the application layer) or not feasible due to the gigantic size and dynamic nature of data.

To obtain the complete dataset, one usually has to contact the data owner. Analogous to conducting web surveys, this method is on the boundary between the Internet as a data source and conventional data collection. The advantage of this method is the depth of the analysis. Unrestricted access to a (dynamic) database offers almost endless possibilities in many sectors. Imagine having full access to all (or most of) the dynamic databases containing job offers. This could result in real time indicators for economic growth, the development of sectors, geographic concentration of specific economic activities, wage developments, etc. The disadvantage of this method lies in obtaining the data. It will be very hard to obtain full data access due to the fact that this data is usually an extremely important asset of the owner. Most companies on the web flourish because they have a unique set of data, e.g. Google, Facebook, YouTube, eBay, CNN, etc.

Section B: Taxonomy of IaD methods

In the remainder of this annex we mainly present the taxonomy of these different methods to use Internet as a data source. In this taxonomy both the web surveys and the in-depth data analysis are excluded. Both offer the possibility of many interesting analyses, but they are not an innovative way of obtaining data and have been used in practice for many years. Also, it is doubtful if these methods are genuine examples of Internet as a source of data.

To obtain a taxonomy, all the remaining methods that are both innovative and use the Internet as a source of data, have been rated on two dimensions. First, we can discriminate between the ability of a method to measure different applications. At one extreme we see methods capable of measuring all applications (e.g. deep packet inspection). At the other extreme we see methods that are only able to measure a single application (e.g. benevolent spyware).

Second, we can discriminate between the user numbers a method measures. A traffic monitor at OS level can measure the behaviour of one single person. On the other side of the spectrum, we see the benevolent spiders measuring the result of the actions of all Internet users on a certain application at once. If we put all the methods in a diagram displaying both dimensions, we obtain Figure B.

When analyzing Figure B, traffic monitoring in the site-centric domain directly attracts attention. This method covers a large surface of the

Figure B: Taxonomy of methods for using Internet as a source of data

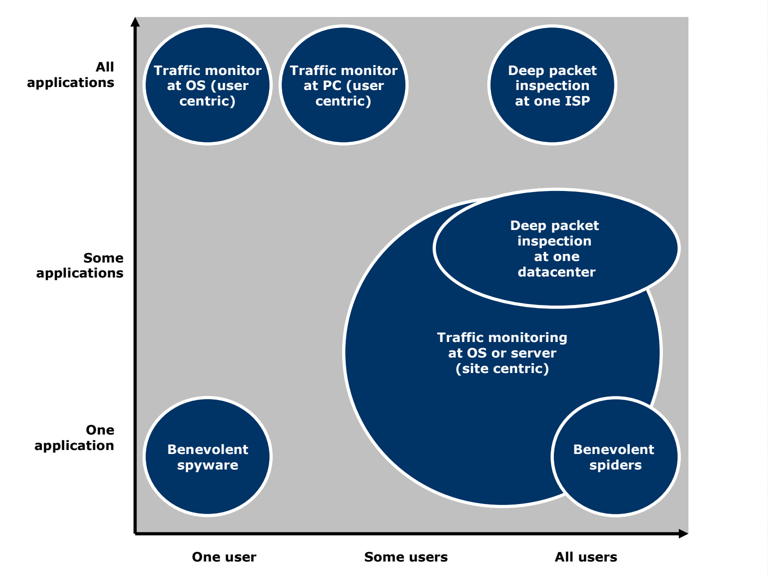

All applications
Some applications
One application
Traffic monitor at OS (user centric)
Traffic monitor at PC (user centric)
Deep packet inspection at one ISP
Deep packet inspection at one datacenter
Traffic monitoring at OS or server (site centric)
Benevolent spyware
Benevolent spiders
One user
Some users
All users

diagram. This is due to many possible variations in technical configuration at the site of the application(s) supplier. An application can run on one server (and thus OS), but it may also need a multitude of servers.[6] So when measuring a server, it is not certain if all or just some of the users are found.

But one server (and thus OS) can also host more than one application. If one applies traffic monitoring at server level, all applications running on the server are detected (which, of course, is not equal to all applications on the Internet). A similar problem occurs when using deep packet inspection at a data centre. To conduct a proper measurement we need to know *a priori* if the application is hosted in one or more data centres.

On the upper side of the taxonomy we see the methods using a broad but superficial measurement of all applications. On the lower side of the diagram we see the methods that apply a small but deep measurement. The most interesting methods in the taxonomy are displayed in the corners. If we focus on a single user, we are able to conduct analyses regarding user properties. Focussing on all users enables us to focus on the world in general. Everything in between (i.e. some users) poses a huge problem of generalization for us. If we look at the application dimension, the same logic applies. We need to know if we are measuring one specific application or all the applications. Everything in between gives us the same problem of generalization.

The taxonomy also makes clear that no single method can be defined as the best overall method.

Which method we should apply to use for Internet as a source of data depends on which goals we want to fulfil.

Section C: The most promising methods

The merits of the methods in the corners of the diagram (Figure B) are discussed in the this section, i.e. (a) benevolent spyware, (b) a traffic monitor integrated in the operating system of an end-user, (c) deep packet inspection at the ISP-level and (d) benevolent spiders. In Section 4.3 we will reflect on practical usability of the various IaD methods.

One user – one application: Benevolent spyware

The use of benevolent spyware is interesting if the research focuses on a very specific behaviour of a (sub) population. The method needs a well-documented user panel and benevolent spyware focussing on one application, e.g. a browser like Internet Explorer or a media player such as RealPlayer. Typical questions that can be answered using this method are:

1. Which type of users visit financial services websites?
2. Do men spend more time looking at LinkedIn profiles than women?

The strength of this method is the possibility to generalize in-depth usage information to a (sub) population. It gives us a unique insight into the way Internet users act in practice. The data obtained by this method can be treated in much the same way, as most of other data used at statistic offices.

However, the method also has some downsides. First, the method will probably underestimate illegal and shameful user behaviour. This can be due to selective non-response, e.g. users very active in illegally using P2P-applications

[6] For example: While the exact size of the (probably) four main data centres of Google are unknown they were estimated in 2000 at 6000 processors (servers). Source: http://en.wikipedia.org/wiki/Google_platform.

will probably refuse to participate in the measurement. The underestimation can also be due to a change in user behaviour, induced by the knowledge of being watched. This method is also unsuitable for detecting very small effects, as a result of the limited panel size. Furthermore, benevolent spyware is also unable to notice if users start adopting new applications, which means some macro trends are hard to discover.

One user – all applications: Traffic monitor integrated in the OS of an end-user

By using a traffic monitor in the operating system and a panel of users, a broad profile of the behaviour of an individual Internet user can be obtained. Typical questions that can be answered using this method are:

- Do users who often use protocols associated with illegal multimedia content, like P2P and UseNet, also make more use of websites associated with legal multimedia content like iTunes, Jamba and YouTube?
- How many phone calls are made through Skype and who makes those calls?
- Do users living on the countryside make more use of Usenet than other users?

A traffic monitor at the OS level offers the opportunity to perform a broad measurement of user behaviour. This is a feature no other method can offer. Just like the benevolent spyware, this method suffers from the same problems of underestimating illegal and embarrassing user behaviour. The method is also less suitable for picking up small trends, due to the small panel size. On the other hand, this method does enable us to pick up trends in the applications usage. The traffic monitor is able to see all the traffic a user generates and therefore all the applications he or she used. Although this method can detect all applications at one time, the downside is the more limited insight into the actual use of application.

All users – all applications: Deep packet inspection at an ISP

In our model in figure 3.1 Deep packet inspection at an ISP is placed between the user and the market. This method enables us to build a real time link between user behaviour and the applications use. This method can answer question like:

- Which website experienced a massive growth over the last month?
- What is the level of P2P-traffic and how does this change?
- How much streaming multimedia content is sent over the Internet and how does this change during a 24-hour period?

The biggest advantage here is that it is the only method that measures a very large level of Internet traffic that is unaffected by social desirable bias. Therefore, shameful, embarrassing and illegal behaviour can be measured easily. Also, the huge amounts of data that can be collected enable us to track down very small changes and spot trends at an early stage. The huge amounts of data also enable true real-time measurements. The disadvantages stem from the limited insight into the properties of the user and the content. The only property that is always known from the user is the IP-number of their connection.

Moreover, since it is impossible to apply this method at the networks of all the ISPs, generalization of the data may develop. This is because ISPs often focus on certain clients in specific geographic areas. Furthermore, in practice it proves to be very hard to find a single ISP that is willing to cooperate. The harmful effect this method can have on users privacy can also attribute to this problem. From the point of view of the content, the major limitation is the incremental perspective on datasets. This method is only able to see users changing and reading content, but not the content (e.g. a database of an online marketplace) as a whole.

All users – one application: Benevolent spiders

The method of applying benevolent spiders to online content enables us to measure economic effects of markets by applications known to the researcher. Its profound view of data enables us to answer questions like:

- What is the average value of a house in Amsterdam that is for sale (on several web portals) and how does this develop over time?
- What is the current level of job offers (on several web pages) and where are these located?
- What is the size of the money flow that is generated by online marketplaces?

The main characteristic and advantage of this method is its ability to give in-depth insight into content. Furthermore, it proves to be relatively cheap compared to the other methods – at least, when it comes to the use of simple spiders. A disadvantage is its inability to find (macro) trends in Internet use. The method only analyses the applications that it is programmed for. In markets with a low concentration level, and therefore many (supplier) websites, this method is hard to implement. Another disadvantage is the fact that a spider mainly finds information on the phenomena known to the spider's programmer. In other words: To find the number of videos on YouTube implies you know beforehand that YouTube is a major player in its market segment. The programmer of the spider has to feed the spider with YouTube's URL to conduct a search on this website.

Annex 3
Stylized results 8 case studies

1. Case webstores/marktplaats.nl

In this case study, we focussed on C2C marketplaces in the broader context of (B2C webshops and) online shopping. Therefore, the range of products is very broad and these are mainly physical products. Recently, services are also being offered by C2C marketplaces, where second-hand goods dominate. C2C marketplaces are predominantly nationally (and more specific: regionally and even locally) oriented, with the exception of very specific categories of advertisements, such as holidays and holiday homes. The online shopping (webshops) segment is B2C and the sites for advertisements are mainly C2C. There is a big grey area, however, filled by the so-called "webtraders" who are active on a more or less professional basis. The market for C2C marketplaces is highly concentrated. The biggest player (Marktplaats.nl) was the main object of research in this case study.

There are no existing indicators for the C2C segment. In the B2C segment there are figures on the number of webshops from the Netherlands Chamber of Commerce and the Statistics Netherlands. Due to differences in definition these figures do not match. Also, the branch organisation representing webshops (Thuiswinkel.org) has its own home shopping market monitor which indicates that consumer spending online in the B2C segment has grown by 30% per year over the last few years.

Trends and developments

First of all, a relevant trend is the enormous growth in the reach of Marktplaats.nl, both in terms of the number of advertisements (a rise from 1 million new advertisements per month in the first quarter of 2004 to 6 million in the third quarter of 2007) and in visitors (20% reach among active Internet users in the Netherlands in 2002 rising to 60% reach in 2007).

Trust is the main driver in this market and this corresponds with all sorts of developments in the area of safer payments, trusted third party solutions, etc. Trust in e-commerce is definitely going up when we look at increased online spending, the number of transactions and the average amount spent per transaction.

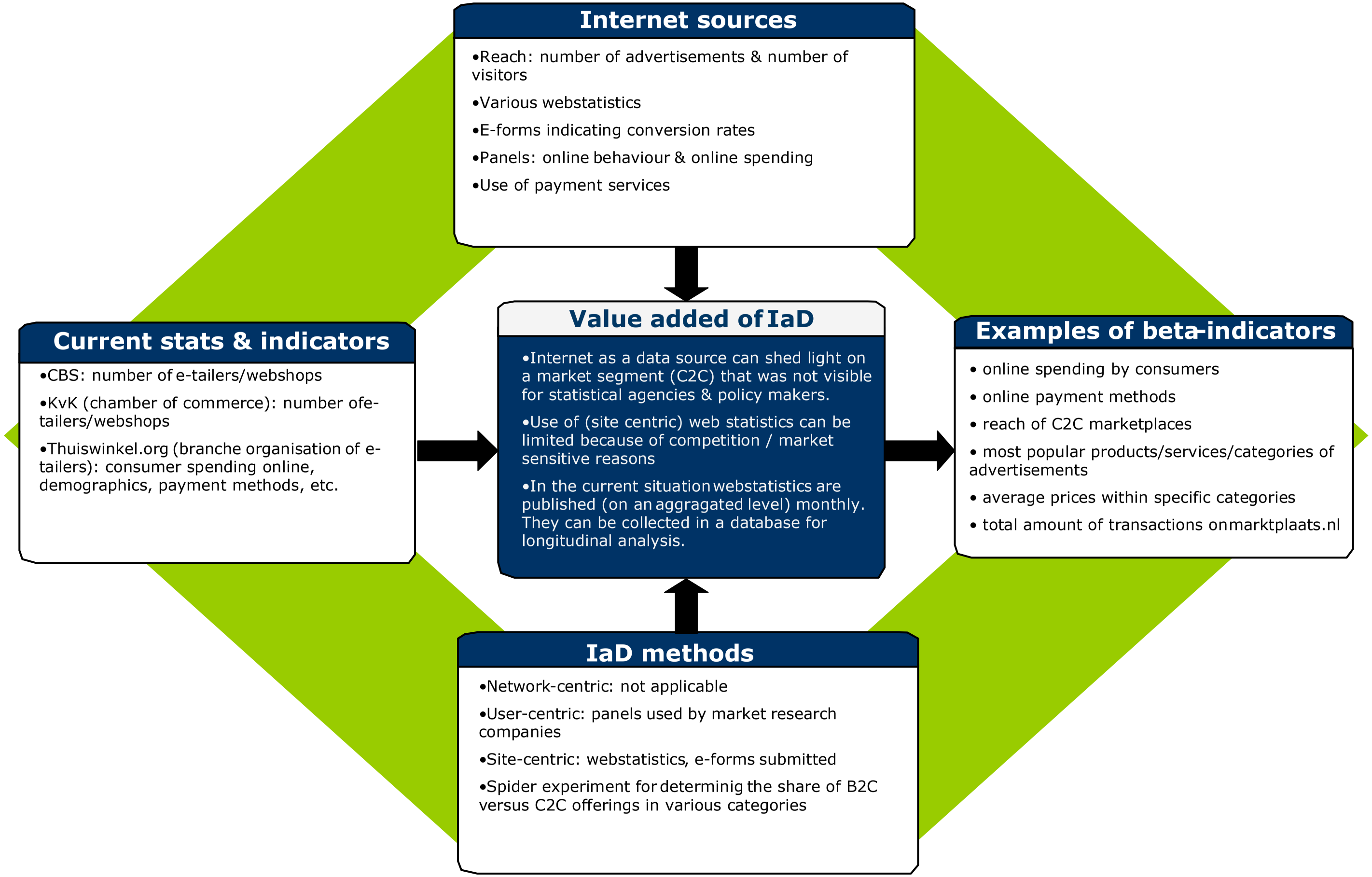
Added Value of IAD

When we determine the added value of using Internet as a data source as compared to traditional methods of data collection and existing sources of information, the following conclusions can be drawn (Figure C.).

Research conducted

Specific research questions addressed	1. What is the reach of marktplaats.nl? 2. What is the share of B2C offerings versus C2C offerings? 3. What is the total amount/value of transactions?
Sources	Sources of information or digital footprints can be found in: 1. Web statistics of marktplaats.nl 2. E-forms used by marktplaats to measure conversion rates 3. Use of payment services 4. use of logistics/fulfilment services
Methods & experiment	We used data both from user-centric measurements (dedicated panels from market research companies) and data from site-centric measurements (web statistics). In addition, we built a spider to collect advertisements and to distinguish between B2C offerings and C2C offerings. In addition, this spider collected information on average prices. This information is used in determining the total amount of transactions.

Figure C: Overview of value added of using IaD-methods when measuring webstores/marktplaats.nl



Internet sources

- •Reach: number of advertisements & number of visitors
- •Various webstatistics
- •E-forms indicating conversion rates
- •Panels: online behaviour & online spending
- •Use of payment services

Current stats & indicators

- •CBS: number of e-tailers/webshops
- •KvK (chamber of commerce): number ofe-tailers/webshops
- •Thuiswinkel.org (branche organisation of e-tailers): consumer spending online, demographics, payment methods, etc.

Value added of IaD

- •Internet as a data source can shed light on a market segment (C2C) that was not visible for statistical agencies & policy makers.
- •Use of (site centric) web statistics can be limited because of competition / market sensitive reasons
- •In the current situation webstatistics are published (on an aggragated level) monthly. They can be collected in a database for longitudinal analysis.

Examples of beta-indicators

- • online spending by consumers
- • online payment methods
- • reach of C2C marketplaces
- • most popular products/services/categories of advertisements
- • average prices within specific categories
- • total amount of transactions onmarktplaats.nl

IaD methods

- •Network-centric: not applicable
- •User-centric: panels used by market research companies
- •Site-centric: webstatistics, e-forms submitted
- •Spider experiment for determinig the share of B2C versus C2C offerings in various categories

We were able to use site-centric measurements (web statistics: reach, e-forms: conversion rates) and a spider experiment for determining average prices – in order to make a calculation of the total amount of transactions on marktplaats.nl. This calculation is based on (conservative) assumptions that were used earlier by Marktplaats in 2004 (before the company became part of eBay and therefore could make public this type of market-sensitive information). Our calculation shows that in 2006 the total value (of the total number of transactions on marktplaats.nl) was approximately €4,7 billion. This information on the size of this specific market segment was not known by statisticians, policy-makers and – most important – the tax authorities. The tax authorities in the Netherlands have, however, developed their own program called Xenon that is used to determine how much value added tax (VAT) has been evaded by the more professional sellers on C2C marketplaces.

Wider economic impact

We have learned several important lessons from this case, which relate this specific phenomenon (growth of C2C marketplaces) to the "real world".

- We noted the fact that second-hand goods can be sold very easily and this will influence substitution demand for specific products and markets.
- A second important lesson is that when selling goods online the low barriers to entry can be a stepping-stone for start-up entrepreneurs (a gradual shift from the C2C segment to the B2C segment).
- A third relevant impact on the real world is the fact that spatial patterns are emerging in C2C marketplaces.

Especially in the rural areas in the Netherlands there is a lot of activity in C2C transactions. These are all reasons to further invest in IaD methods so as to derive a more comprehensive understanding of the phenomenon, use and economic impact of C2C marketplaces.

2. Case – the market for recorded music

In this case study we focus on the market for fixed music fragments. This contains the sales of music associated with a physical carrier, like a CD, and music not associated to carriers, like MP3 files. Broadcasting, such as radio stations, is excluded from this case study. At this moment, the market predominantly consists of (carrier and non-carrier based) digital music.

Traditionally, the music industry has an international focus. A relatively high level of concentration typifies the market: only four record labels account for approximately 75% of the total turnover. Regulation in this market is relatively low. However, due to increasing illegal activities, more focus is being put on the protection of intellectual property rights. The B2C-segment is dominant, but B2B and C2C play a role in this market as well.

There are many existing indicators around. Statistics Netherlands has figures on the development of the (music) retail sector as well as Internet user behaviour concerning music. The OECD uses figures on the size of the retail sector, as well as the worldwide record sales. Representatives of the recording industry (NVPI, IFPI) have detailed information of their sector, including illegal behaviour, that they say is harming their sector. Some companies actually use the Internet as a source of data, like Big Champagne, Nielsen Soundscan and Ipoque. They have data regarding the use of P2P applications.

Trends and developments

Over the last 20 years the market transformed from an analogue carrier (e.g. vinyl records) by way of the digital carrier (e.g. CD's) and seems to be moving towards a situation where a substantial amount of digital music is being sold without a carrier (e.g. MP3). Digital music, especially music not associated to a carrier, opened a major window of opportunity for illegal distribution. However, in the recent years we see major companies using digital music without a carrier for regular business models, like iTunes, Zune, etc.

The rise of (legal and illegal) MP3's has (had) major effects on the traditional market. The traditional record stores and the major record labels experience major negative effects, while many consumers and small artists experience major positive effects. Legal and illegal music can be distributed on the Internet through a multitude of channels, like P2P, one-click-hosting, Usenet, etc.

Added Value of IAD

When we determine the added value of using Internet as a data source as compared to traditional methods of data collection and existing sources of information, the following conclusions can be drawn (Figure D).

Research conducted

Specific research questions addressed

1. Which music is made available for (illegal) file sharing?
2. What is the popularity of (illegal) P2P transfer for recorded music?
3. What is the current average price of a CD or music-DVD?

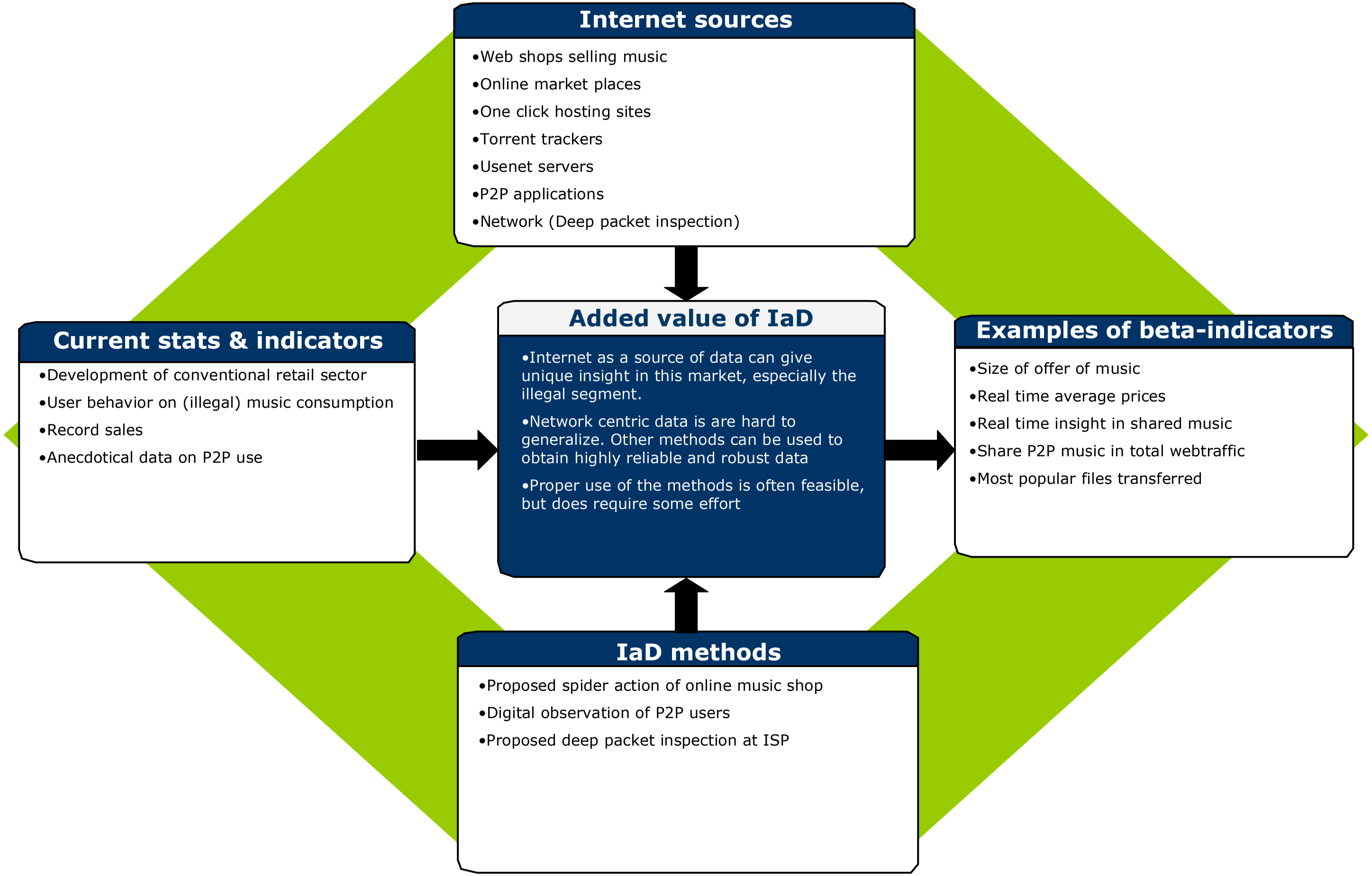
Sources

When applying site-centric measurements, interesting data can be obtained by online suppliers of music: Web shops selling CD's, ring tones, MP3's, etc. The C2C-segment can be examined by using online marketplaces (see other case). A part of the illegal distribution issue can be covered by focussing on the one-click-hosting sites, Usenet servers and legal P2P-sites, like torrent trackers. User-centric measurements can be used on P2P-applications. These applications almost always show the properties of the (content of the) users. Benevolent spyware (or traffic monitoring) can also be used to monitor P2P-applications, but will probably underestimate its use. A network-centric measurement using deep packet inspection at an Internet service provider can help us to obtain insight into the true use of music distribution by P2P.

Methods & experiment

The case presents a proposal to perform a spider action at a major online music webshop (site-centric). Furthermore, a small experiment on a P2P application was conducted (user-centric). Deep packet inspection is, of course, also valuable in this case study (network-centric).

Figure D: Overview of value added of using IaD-methods when measuring recorded music



Internet sources

- Web shops selling music
- Online market places
- One click hosting sites
- Torrent trackers
- Usenet servers
- P2P applications
- Network (Deep packet inspection)

Current stats & indicators

- Development of conventional retail sector
- User behavior on (illegal) music consumption
- Record sales
- Anecdotical data on P2P use

Added value of IaD

- Internet as a source of data can give unique insight in this market, especially the illegal segment.
- Network centric data is are hard to generalize. Other methods can be used to obtain highly reliable and robust data
- Proper use of the methods is often feasible, but does require some effort

Examples of beta-indicators

- Size of offer of music
- Real time average prices
- Real time insight in shared music
- Share P2P music in total webtraffic
- Most popular files transferred

IaD methods

- Proposed spider action of online music shop
- Digital observation of P2P users
- Proposed deep packet inspection at ISP

From this case study we can conclude that Internet as a source of data can give us a unique insight into this market (especially the illegal segment). We conducted an experiment in which we were able to see which types of files are shared by users in a P2P-application. This, and methods focussing on torrent trackers, Usenet, one-click hosting providers, all gave us a real time insight into the type of music that is being shared. Moreover, existing research on deep packet inspection showed that P2P-traffic is the largest type of Internet traffic. A substantial amount of this traffic consists of music. Finally, spidering sites containing price information on CD, like the web shop of C2C marketplaces, can be used to construct a real-time indicator of the CD price.

Wider economic impact

Illegal music distribution via the Internet has (had) a major negative influence on the traditional sellers of music. Several artists and record companies tried to incriminate users active in file sharing. But, these developments also lower the entry barriers to this market. This results in new artists being able to publicize themselves to the world at almost zero costs. Moreover, the illegal file sharing is, undoubtedly, a major driver of the market for MP3-players, like the iPod. Paradoxically, this opened up new opportunities for employing new business models for exploiting legal digital music.

3. Case market- Internet TV

This case study discusses Internet-TV. Internet-TV is defined as video that can be watched using a PC and requires regular Internet access; IP-TV is therefore excluded. The video can be delivered to the end-user by a stream or a downloadable file. Another distinction is made between video data on a server and another computer (P2P). Digital video can be copied easily. However, in many cases there is no need for copying since the (legal) content can be found on the Internet. Of course, illegal content also plays a big role in this market.

Internet service providers play a significant role in this market by physically connecting supply and demand. ISPs usually have a rather national focus and limited market power. They experience a relative high amount of regulation and suggestions for additional regulation are done on an almost daily basis (e.g. even to stop online bullying). Suppliers of a broad spectrum of content –like YouTube- usually have an international character and experience a medium amount of market concentration.

Suppliers of specific content –like online football matches- often operate in a national playing field. Regulation of both is usually focussed on the protection of intellectual property right and compliance with national regulation with respect to the content. Traditionally the TV market is predominantly B2C, but the Internet added the C2C component to this market.

The Dutch audience research foundation (SKO) uses a survey to collect data regarding TV viewing habits on a PC. Furthermore, they also use (site-centric) data of the Internet-TV website of the Netherlands public broadcasting organization to correct their viewing figures. There is also some anecdotal evidence from different sites offering video content. Network-centric data has been gathered by different organizations but the reliability and opportunities for generalization are low.

Trends and developments

The use of Internet-TV has exploded since 2005. Before 2005 technological barriers hindered the successful spread of this technology. Internet-TV changed the TV market from a market with extremely high entry barriers to one where they are now extremely low.

Research conducted

Specific research questions addressed	In this case study we addressed the following questions: 1. Who uses Internet-TV? 2. What are properties of the offer made by Internet-TV? 3. What is the share of Internet-TV against the total Internet traffic level?
Sources	Structural use of site-centric measurement can give us more insight into the supply side of this market. A significant part of this market can be covered by analyzing (spidering) just a few initiatives, such as YouTube.com and video.[Yahoo/msn/Google].com. Site-centric measurement can also be used to track developments in the illegal segment by monitoring NZB sites and torrent trackers. Log data of media players offer a unique opportunity in user-centric monitoring. User-centric measurements can also be applied to monitor the properties and use of P2P and P2PTV applications. The gigantic size of video data files, with respect to all other data on the Internet, strengthens the usefulness of network-centric measurements.
Methods & experiment	To obtain data relatively easily in this case study, we proposed three methods. The first one uses a site-centric measurement in a sub-domain of this market: video reports of meetings at a city council. The high concentration and low dynamics in thus sub-market make site-centric measurement very suitable. The second method proposed a deep packet inspection at a (Dutch) Internet service provider. Although the generalization of this data is hard, it provided us with a unique view of (especially illegal) activities. The third focused on the use of benevolent spyware to conduct user-centric measurements on this topic.

Added Value of Internet as a data source

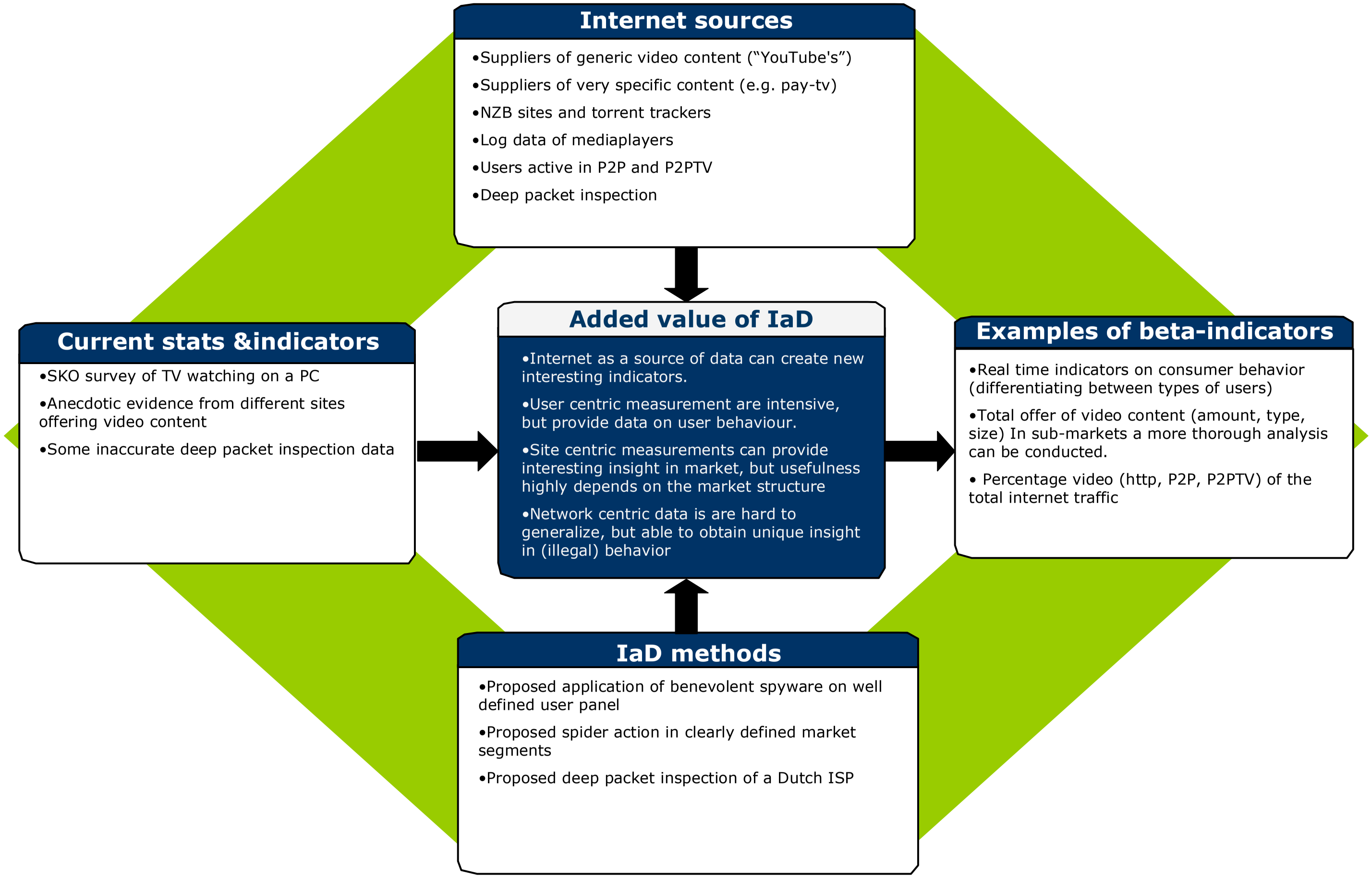
The added value of Internet as a source of data can be seen in Figure E.

Using Internet as a data source can result in several interesting indicators.

First, by using some kind of benevolent spyware, possibly integrated into current media players, interesting data can be obtained on user behaviour. It enables us to see what types of users use Internet-TV and at what times.
Second, site-centric measurements give us the opportunity to obtain insight into markets. The illegal segment can be covered by spidering NZB-sites, torrent trackers or other gateways to illegal content. The legal segment can be analyzed by spidering regular website like YouTube and all its competitors. Especially very clearly defined market niches offer great opportunities for obtaining in-depth market info by using spiders. We examined the case for webcasting meetings of from a municipal council.

Third, deep packet inspection is very suitable for Internet-TV. Internet-TV is one of the most data-intensive types of content. Although very hard to generalize, DPI could give us unique insight into illegal behaviour.

Figure E: Overview of value added of using IaD-methods when measuring Internet-TV



Internet sources

- •Suppliers of generic video content (“YouTube's”)
- •Suppliers of very specific content (e.g. pay-tv)
- •NZB sites and torrent trackers
- •Log data of mediaplayers
- •Users active in P2P and P2PTV
- •Deep packet inspection

Current stats &indicators

- •SKO survey of TV watching on a PC
- •Anecdotic evidence from different sites offering video content
- •Some inaccurate deep packet inspection data

Added value of IaD

- •Internet as a source of data can create new interesting indicators.
- •User centric measurement are intensive, but provide data on user behaviour.
- •Site centric measurements can provide interesting insight in market, but usefulness highly depends on the market structure
- •Network centric data is are hard to generalize, but able to obtain unique insight in (illegal) behavior

Examples of beta-indicators

- •Real time indicators on consumer behavior (differentiating between types of users)
- •Total offer of video content (amount, type, size) In sub-markets a more thorough analysis can be conducted.
- • Percentage video (http, P2P, P2PTV) of the total internet traffic

IaD methods

- •Proposed application of benevolent spyware on well defined user panel
- •Proposed spider action in clearly defined market segments
- •Proposed deep packet inspection of a Dutch ISP

Wider economic impact

Given the vast size of video content, the rise of Internet-TV has had an important influence on the demand for broadband capacity. This has stimulated the rollout of new broadband networks and the upgrade of current networks. Furthermore, it also allows new artists to gain a strong reputation within a short time. This development has a direct negative influence on the number of consumers watching conventional TV and thus this value chain as a whole. Finally, it has had a negative impact on all the organizations using business models that rely on exclusive video coverage, e.g. the movie industry, pay-TV, the video rental segment, etc.

4. Case - online gaming

This case focused on three segments of the (online) gaming market: traditional console games and PC games (which are increasingly moving online), massive multiplayer online role-playing games (MMORPGs) and casual games. The three segments are very different and cannot be compared (thus essentially our research comprises three individual cases).

The market for console games has an international focus and is dominated by just a few big players, both in the hardware (consoles) and software (games) layer. With regard to the latter, the market is highly dynamic at this moment, with a high degree of consolidation.

The market for MMORPG's has many players but is also highly concentrated. For a long time the market has been largely neglected by the big players from the traditional (console) games industry. Much of the action was in specific niche markets, e.g., South Korea and/or teens. But due to the exponential growth of players worldwide (and in particular the phenomenal commercial success of World of Warcraft), it is now also subject to similar dynamics as witnessed in the console software market.

Finally, casual games are a completely different market – much more nationally oriented than the console and MMORPG market and with another dominated business model (advertisement). In terms of players, casual games by far surpass any of the other two markets, partly because it has opened up new user groups. In terms of revenue, however, it remains to be seen how viable the market is.

There are hardly any reliable statistics available on (online) gaming. The only exceptions are figures on the console and PC game market but these are solely based on retail figures and because of the switch towards online distribution (both legal and illegal) cover an ever shrinking part of the market. For the other two markets, even for the key statistic of the total number of active users, only rough and widely varying estimates are available. In general, the total number of active users is grossly overrated (by factors of between 5-10).

Trends and developments

Due to rampant software piracy and competition from online games, the position of PC games vis-à-vis console games is deteriorating. The big players in the console market have retained and even strengthened their position due to the control over the proprietary technical standards. These big players have never been able to operate successfully on the MMORPG market – currently they are just buying up specialised MMORPG developers.

The development of the total number of MMORPG players worldwide shows a neat exponential growth (doubling about every year) and stands currently at a respectable number of 40 million – of which 10 million are for World of Warcraft. The spillover from the consolidation in the MMORPG market has also created a lot of dynamics in the business models. The dominant traditional model of monthly subscription fees (still highly successfully used by World of Warcraft) is now competing with new business models (one-time purchasing price or free to play with revenues from paid added functionalities (primary market for virtual objects or Real Money Trade, RMT). A well-known example is the sale of plots of virtual land in Second Life. There is also a growing segment of companies that specialise in the "harvesting" and re-sale of such virtual objects (secondary RMT). The total value of global RMT has already grown to €2 billion. Most of the trade occurs in/from Asia (Korea, China, and, to a lesser extent, Japan).

In contrast to console/PC games and MMORPGs, casual games have a very short learning curve. This has opened the market for entirely new groups of players (e.g., middle-aged women). Worldwide several hundreds of million of people are already playing casual games. Generic portals such as Yahoo and MSM still have a dominant market position (esp. in their home market the US) but similar to the developments in the MMORPG market it seems that specialised developers are more successful (including several Dutch firms). The established bigger players are also buying up these niche players. Although there are substantial amounts of money involved in the mergers and acquisitions, the margins in the business remain very low. It remains to be seen how the enormous number of players can be turned into commercial success.

Added Value of IAD

Each of the three markets distinguished above (console/PC, MMORPG, casual games) has distinctive characteristics, hence the added value of IaD methods differs greatly.
In Figure F, conclusions for the most relevant market (at least from the IaD perspective) – the market for MMORPGs – have been summarized.

Research conducted

Specific research questions addressed

1. How has the number of players of online games developed over time and what is the current reach?
2. What are the three largest online games (MMORPGs, casual games) – in the Netherlands and worldwide?
3. How do these large online games perform in terms of market share and turnover?

Sources

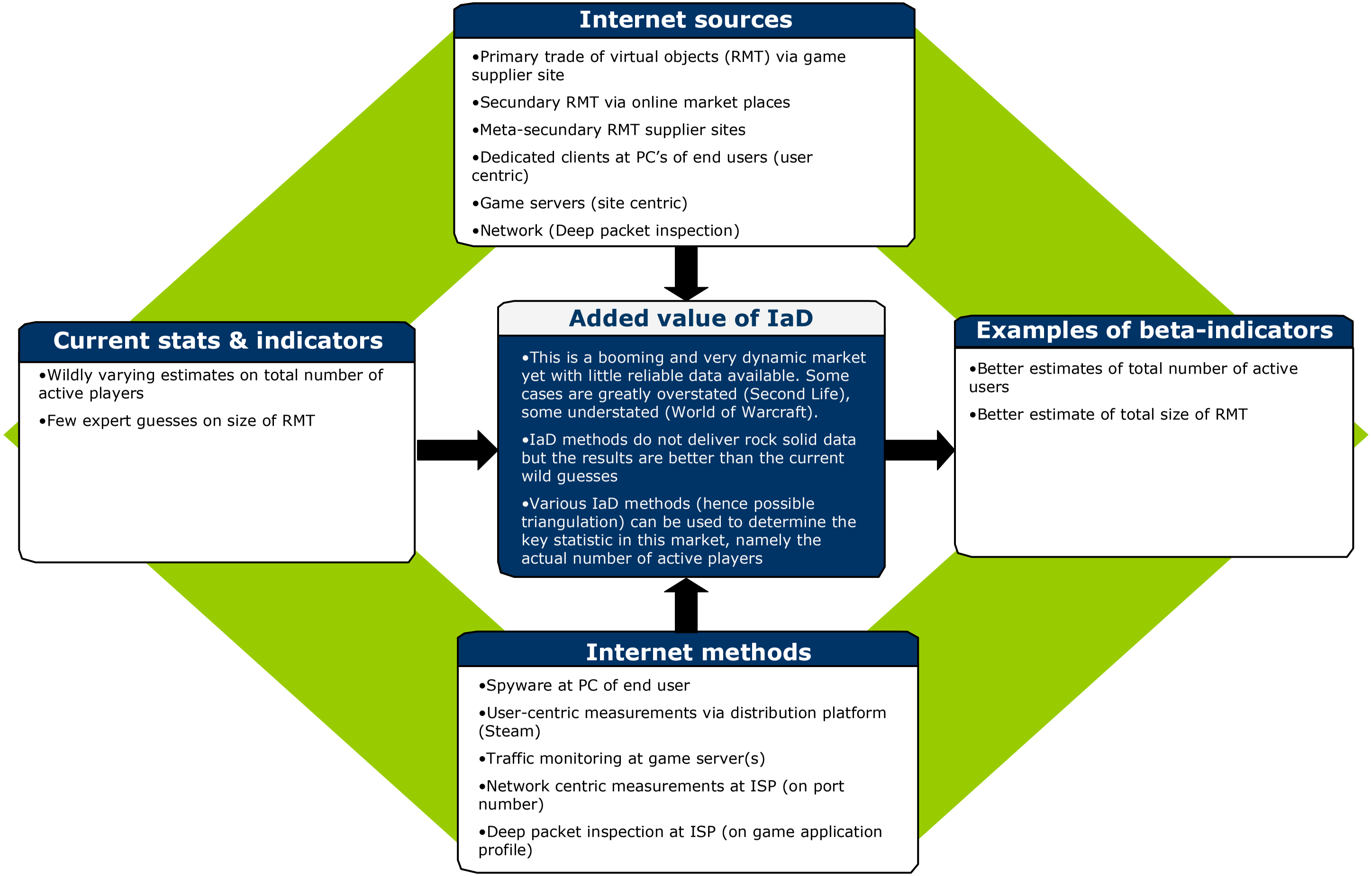
A great number of sources of information has been identified, such as:

1. Web statistics from dominant market players
2. Market data from (non)commercial marketing research bureaus (e.g,. Stichting Internetreclame in the Netherlands, Comscore internationally).
3. Network-centric measurements (downloaded games in P2P traffic, online games with fixed ports)
4. User-centric measurements conducted by third parties (Steam PC software distribution platform)
5. Data on number of users published on specialised data portals (MMOGData for MMORPG market, VZChartz for console and PC game market)

Methods & experiment

In this case study, only secondary data has been used. This data is predominantly based on IaD methods applied by third parties (either suppliers themselves or research organisations). The full range of methods has been used (from user-centric and network-centric to site-centric) albeit with a focus on site-centric measurements.

Figure F: Overview of value added of using IaD-methods when measuring market for Massive Multiplayer Online Games (MMO's)



Internet sources

- Primary trade of virtual objects (RMT) via game supplier site
- Secundary RMT via online market places
- Meta-secundary RMT supplier sites
- Dedicated clients at PC's of end users (user centric)
- Game servers (site centric)
- Network (Deep packet inspection)

Current stats & indicators

- Wildly varying estimates on total number of active players
- Few expert guesses on size of RMT

Added value of IaD

- This is a booming and very dynamic market yet with little reliable data available. Some cases are greatly overstated (Second Life), some understated (World of Warcraft).
- IaD methods do not deliver rock solid data but the results are better than the current wild guesses
- Various IaD methods (hence possible triangulation) can be used to determine the key statistic in this market, namely the actual number of active players

Examples of beta-indicators

- Better estimates of total number of active users
- Better estimate of total size of RMT

Internet methods

- Spyware at PC of end user
- User-centric measurements via distribution platform (Steam)
- Traffic monitoring at game server(s)
- Network centric measurements at ISP (on port number)
- Deep packet inspection at ISP (on game application profile)

In principle, for the measurement of the scope of the MMORPG market all IaD methods can be applied. Benevolent spyware (such as already being used by Valve) could be deployed to monitor the online behaviour of gamers in detail (this could, for instance, give new insights into the growing concerns surrounding gaming addiction). Worldwide traffic patterns of specific MMORPG applications can be measured right in the middle of the network or preferably at the hinge between the network and the servers on which the online games are hosted. In contrast to most other cases, site-centric measurements seem to be the least applicable (although they might be very relevant in the specific case of casual games; that is, "traditional" web statistics such as the number of unique visitors, average page view, click conversion rates etc.). The combination of various methods (possibly by triangulation) could generate more robust and valid data on the crucial statistic of the actual number of active users.

Wider economic impact

The gaming market is essentially built on the principle of making money on from people's spare time. Contrary to leisure activities, such as tourism and sports, there are seemingly little or no broader societal benefits involved. Thus the money made by individual firms (micro level) could, to a certain extent, be regarded as a waste on the macro level (opportunity costs in terms of time spent on non-productive purposes, e.g., casual gaming). In the particular case of MMORPGs, explicit costs occur as a result of gaming addiction.

Online games are probably the most extreme case of the wider trend of the blurring of boundaries between the analogue and the virtual world. As such, they make up a valuable experimental space (also in a negative sense, see earlier). A highly interesting phenomenon in this respect is the legal controversies that surrounded the rise of real money trade – literally making (real) money out of sheer "virtuality". The objects that are being sold have no legal status whatsoever (the administrator of the online game could delete them at anytime without further consequences) but nevertheless they are already worth €2 billion. Also, the first virtual theft charges have already been brought to court (in the Netherlands).

5. Case - social networking

In this case study we focussed on Social Networking Sites (SNS) and their impact on the economy. In fact, there are two types of effects:

- direct (different ways in which SNS make money: business models)
- indirect (effect on existing markets, hardware/ software, ISP & telecom, advertising etc. and the effect of user generated content on traditional media, marketing and brands, etc.).

There are also two types of SNS. First of all, there are generic profiling sites in which network-externalities play a predominant role. The number of members is an indication of the value of the network. Secondly, there are niche-players which bring people together on very specific topics, hobbies, etc. Within the niche segment, the economic value of the networks is mainly determined by the opportunities for advertisers to address very specific target groups.
The biggest players in the Netherlands are Hyves (5.6 million users), Schoolbank and MSN spaces. The three dominant players worldwide are MySpace, Facebook and Hi5. At present the statistical agencies do not collect data on this specific type of online activity.

Trends and developments

An important trend is the discussion on privacy issues. It appears that especially young people are very open in sharing profile information and this can easily be abused (spam, stalking, cybercrime etc.). A second development is the effect of SNS on social behaviour. Research shows that contacts online do not substitute but complement real life friendship. Also, social networking sites are becoming very important in the area of labour market communication. The same goes for using SNS for political communication and mobilisation. A relevant trend in terms of markets and consumer behaviour is that people are using SNS before buying a product or a service. Information from (trusted) peers is becoming more important than other sources of information. A relevant technological trend is to combine profile information within location based services, e.g. to locate friends from your network in your vicinity with the use of a mobile device.

Added Value of IAD

When we determine the added value of using Internet as a data source, as compared to traditional methods of data collection and existing sources of information, the following conclusions can be drawn (Figure G).

The most important conclusion is that Internet sources and methods are (at present) the only way to collect and analyse information on this specific market segment. Another interesting conclusion from this case study is that the self reported number of members from the dominant player in the Netherlands (Hyves) is 30% higher than the information from our own spider experiment indicates. As these numbers play an important role in determining the value of a SNS, this is something to examine in more detail.

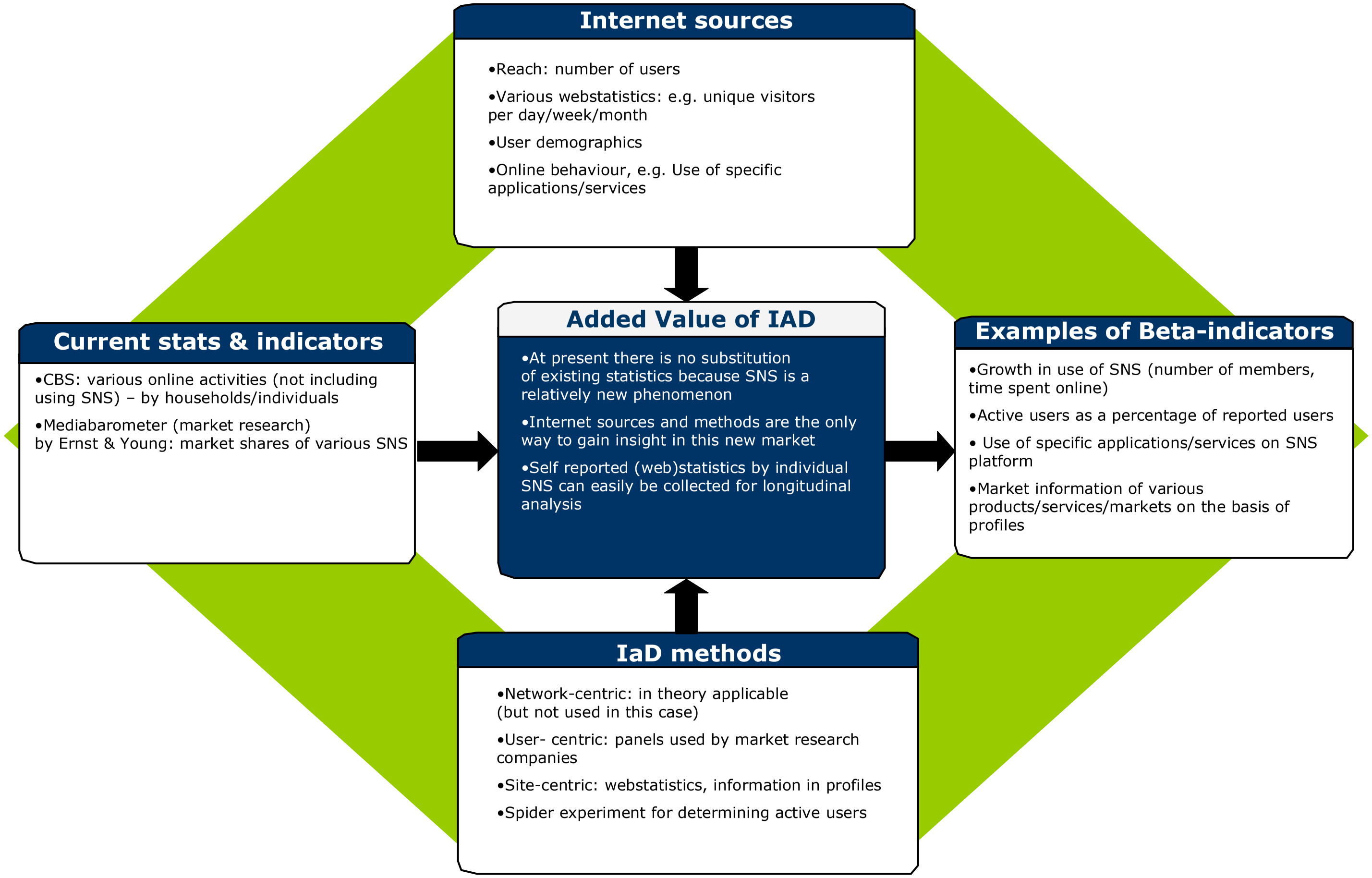
Wider economic impact

In an OECD publication on participative web and user created content (2007), an overview is presented of economic incentives and benefits for various market segments and value chains. What is important to stress is that User generated content starts with (voluntary) peer production in a social context, and can easily be translated into commercially interesting activities in an economic context.

Research conducted

Specific research questions addressed	1. What is the reach (number of users) + the development in time of SNS? 2. What are the three largest SNS – in the Netherlands and worldwide? 3. What are the effects of SNS on the economy?
Sources	Sources of information or digital footprints can be found in: 1. Web statistics 2. Ordering/payment functionality (only in cases when specific services or applications are not free of charge) 3. ISP (using deep packet inspection)
Methods & experiment	We used both data from user-centric measurements (dedicated panels from market research companies) and data from site-centric measurements (web statistics). In addition, we built a spider to determine the percentage of active users within the reported total number of members of the biggest SNS in the Netherlands (Hyves).

Figure G: Overview of value added of using IaD-methods when measuring SNS



Internet sources

- Reach: number of users
- Various webstatistics: e.g. unique visitors per day/week/month
- User demographics
- Online behaviour, e.g. Use of specific applications/services

Current stats & indicators

- CBS: various online activities (not including using SNS) – by households/individuals
- Mediabarometer (market research) by Ernst & Young: market shares of various SNS

Added Value of IAD

- At present there is no substitution of existing statistics because SNS is a relatively new phenomenon
- Internet sources and methods are the only way to gain insight in this new market
- Self reported (web)statistics by individual SNS can easily be collected for longitudinal analysis

Examples of Beta-indicators

- Growth in use of SNS (number of members, time spent online)
- Active users as a percentage of reported users
- Use of specific applications/services on SNS platform
- Market information of various products/services/markets on the basis of profiles

IaD methods

- Network-centric: in theory applicable (but not used in this case)
- User- centric: panels used by market research companies
- Site-centric: webstatistics, information in profiles
- Spider experiment for determining active users

6. Case product software market

Product software is defined here as a packaged configuration of software components or a software-based service, with auxiliary materials, which is released for and traded in a specific market[7]. These can be enterprise-wide packages or systems or modules of software components. In this case we define the Dutch market for product software as the market for standard packages and applications that has a solid developed base, and deliveries to several businesses and consumers.

In this case study, we focus on the Dutch product software industry that is mainly B2B-oriented. The size of the sector is difficult to estimate. Statistics Netherlands has recently adapted its industry classification and now specifies more categories of ICT enterprises and product software companies.[8] Still, other sources need to be consulted to obtain recent estimates about the structure and developments of the Dutch product software market.

Statistics Netherlands has counted around 17,000 computer service companies in the Netherlands. Product software companies are a (unknown) part of this. From alternative sources - vendor comparison portals as ict-base.nl, softwaregids.nl and softwarepakketten.nl - we estimate that there are at least between 930 and 2000 product software companies in the Netherlands. From these sources, however, not much can be said about the market's economic size, composition and characteristics.

During our desk research we also found that the (Dutch) open source software service industry remains unknown when it comes to size and composition. This is inherent on the more or less "hidden" economic and labour activity of this sector. Still, we believe the measurement of the open source software industry (or market) is one of the challenges of the measuring the product software sector.

Trends and developments

In general, competition and consolidation is changing the Dutch software market. While Microsoft, Oracle and SAP dominate large parts of the market, many smaller product software companies exist too - particularly those serving niche markets or particular sectors of the Dutch SMEs.

The product software market is sensitive to economic change, as well as technological hypes and bubbles. From 2003 onwards, however, the market has grown rapidly, driven mainly by new Internet-investments and the growing need to interconnect and integrate applications. Consumers and companies demand mobile and web-based applications, business processes need to be connected and automated (SOA, webservices, Software as a Service, SaaS). Also the Netherlands government and SMEs are catching-up in their IT-maturity using product software.

Because of the nature of their product, product software companies are frontrunners in using the Internet for commercial activities. These include on-line sales, product delivery, product updating and product support. Product software companies therefore have a 100% web-presence that makes them quite suitable for investigation within this Internet as a data source project.

[7] Xu, Lai & Brinkkemper, Sjaak (2007), Concepts of product software, European Journal of Information Systems, Volume 16, Number 5, pp. 531-541.

[8] An overview of changing industry classifications at Statistics Netherlands can be found at http://www.cbs.nl/nl-NL/menu/methoden/classificaties/overzicht/sbi/sbi-2008/default.htm

Research conducted

Specific research questions addressed	1. What is the size and composition of the Dutch product software sector in the Netherlands according to industry statistics and the existing product software portals? 2. How many websites from product software companies can be spidered in order to retrieve on company features such as number of employees, year of foundation and the presence of business software terms?
Sources	As an alternative source, the portal softwaregids.nl and softwarepakketten.nl in the Netherlands can be used to collect the URLs of product software companies and hence build a "bottom-up" database to derive industry indicators. In addition, the websites of the larger product software companies can be monitored on their software delivery, back-up, update and download activities. From this, the economic value of product software companies can be partially estimated.
Methods and experiment	We experimented with a spider build to crawl the websites of Dutch product software companies. The spidering was based on semantic analysis of website content. From the spidered websites the popularity of IT-terms can be indicated, to some extent, as well as the number of employees and company age. The success rate of the spider is limited, however, and its bias for results need to be further investigated.

We also expect that product software company websites have certain uniformity in structure and content, as most of these specialized companies use their home pages to present their basic data such as year of foundation, number of employees, type of products and customers.

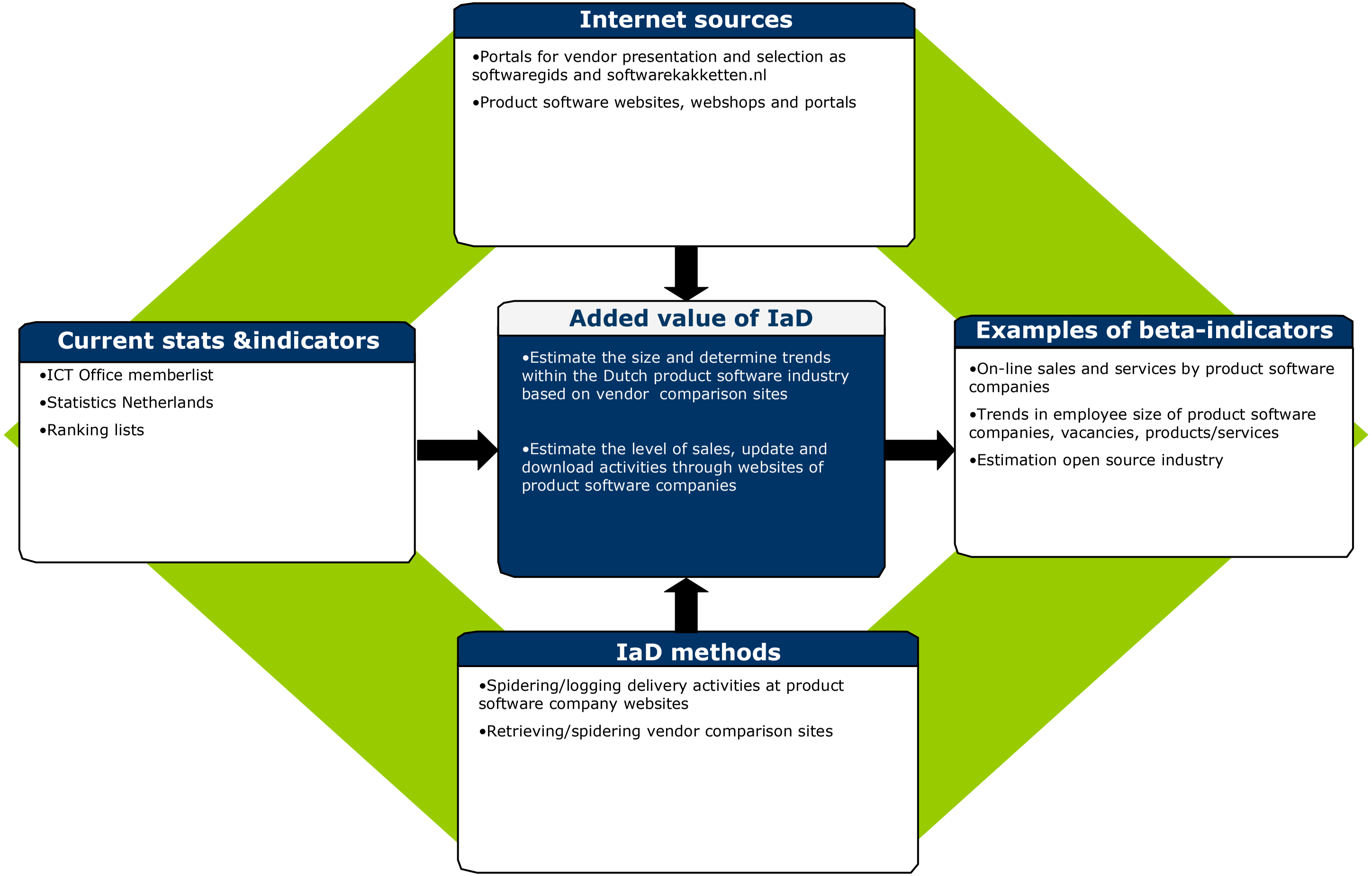
Added Value of IAD

When we determine the added value of using Internet as a data source as compared to traditional methods of data collection and existing sources of information, the following conclusions can be drawn (Figure H).

Wider economic impact

If spiders can be developed that are more successful for searching, collecting and storing information from software comparison portals, this can contribute to the improved estimation of the wider economic impact of the product software industry in the Netherlands. In addition, the larger Dutch product software companies can contribute to this by providing logging data about their on-line sales and delivery activities. As of now, however, both traditional and new methods for data collection are needed to estimate the actual size, composition and developments of the Dutch product software sector.

Figure H: Overview of value added of using IaD-methods when measuring product software company URLs from vendor selection portals



Internet sources

- Portals for vendor presentation and selection as softwaregids and softwarekakketten.nl
- Product software websites, webshops and portals

Current stats &indicators

- ICT Office memberlist
- Statistics Netherlands
- Ranking lists

Added value of IaD

- Estimate the size and determine trends within the Dutch product software industry based on vendor comparison sites
- Estimate the level of sales, update and download activities through websites of product software companies

Examples of beta-indicators

- On-line sales and services by product software companies
- Trends in employee size of product software companies, vacancies, products/services
- Estimation open source industry

IaD methods

- Spidering/logging delivery activities at product software company websites
- Retrieving/spidering vendor comparison sites

7. Case – the housing market

In this case study we have analysed the market for private property in the Netherlands. There are almost 7 million houses in the Netherlands that are highly heterogeneous in terms of size, quality, location and price. Their quality - and hence price - is dependent on the object and the quality of the region and neighbourhood where it is located.

The housing market is a market where the stock is relatively fixed as only a modest share of new houses is added (and demolished) each year. The housing market is embedded or linked to large markets, such as markets for construction and project development, brokerage services, financial services and advice, legal services etcetera. Markets can be segmented by region, rented versus owned houses, housing characteristics. Information about housing and the house market is a key asset. Increasingly, more detailed and timelier information on the houses for sale (and to a degree also for houses for rent) is becoming available through the Internet. This considerably increases market transparency.

The housing market is highly regulated. Government involvement is high in various capacities ranging from spatial planner, financer, designer of housing related tax provisions, guardian of affordable housing and so on. Although most houses are sold using brokerage services (B2C), the number of individual buyers and sellers – empowered by detailed information – operating more independently (partly C2C) is on the rise.

There is a wealth of statistics available, produced by various parties including Statistics Netherlands (Housing, housing stock, permits for building houses, WOZ and mortgage statistics), housing preferences & living conditions (VROM), Buildings & addresses (Dataland), statistics based on purchasing and mortgage notes (Cadastre), houses for sale & transaction process (NVM) and housing prices index (WOX) & integrated housing market information and forecasting systems (ABF).

There are various broad based databases on housing in which a large number of sources and indicators based thereon (collected using established methods) have been brought together. Most of these can be accessed electronically, e.g. the VOIS database published by VROM and produced by a private research and consultancy firm ABF.

Trends

Housing markets are stagnating in the Netherlands due to the low production volumes of new houses and low price elasticity (supply does not follow changing demand) resulting in decreasing affordability of houses among especially starters on the housing market and stagnating "housing careers". A new "information on housing" market has emerged based on new business models in less than a decade. Housing sites have developed into the main tool used by house hunters (and related services!). Housing sites compete in providing the complete overview of houses on offer and the detail of that information is on the rise. These housing sites have empowered individual buyers and sellers on the market and direct transactions between consumers are increasing too. Other private and public players (e.g. Kadaster) have invested heavily to ensure that information on housing and the housing market is available in electronic form as well.

Research conducted

Specific research questions addressed

Zoomed in into two key player questions included are:

1. Can the use of Kadaster Online be used as proxy for development of the housing market (prices, market demand, number of transactions, transaction speed)?
2. To what extend can Funda.nl be used for assessing the developments on the housing market (prices, demand, transactions), housing preferences, potential buyers
3. To what extend is the trend towards selling houses without a broker real?

Sources

There are already a fairly large number of data sources in the housing market. Some of these could potentially produce more housing statistics using IaD (e.g. Cadastre Online). Most important new sources are housing sites (Funda, self service sites) and electronic marketplaces, which are increasingly being used in both the B2C and C2C housing market (including marktplaats.nl).

Methods & experiment

Funda (dominant housing site) participates in two projects in which user-centric measurements are used for assessing the popularity and use of its site (Visiscan by Multiscope and STIR by Intomart). Most promising are site-centric measurements of dominant sites in the housing market (Funda, Kadaster Online), although there is a thin line with mining existing available databases that are more readily available in those organizations. Already simple web statistics may provide surprising insights. Network-centric measurements are not feasible as there are no protocols that typically refer to the housing market. We performed a small spider experiment on the Funda and marktplaats.nl websites to assess housing prices over time, transaction speed and most courant housing types.

Added Value of IAD

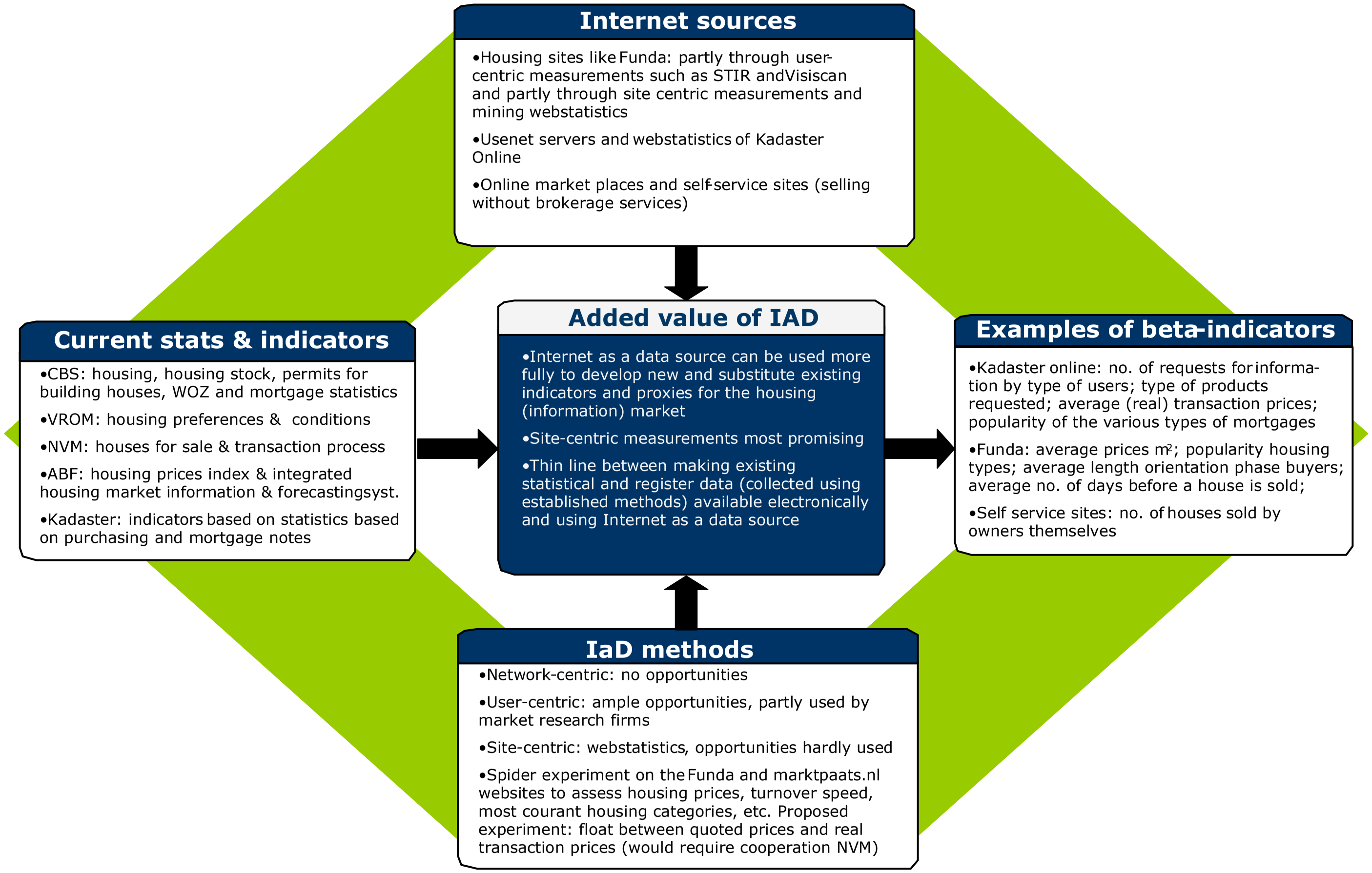
When assessing the added value of using Internet as a data source as compared to traditional methods of data collection and existing sources of information, the following conclusions can be drawn.

The most important conclusion is that although the housing market is well served by statistical indicators, using Internet as a data source could be helpful in developing new and valuable indicators. IaD could also be used to build proxies for monitoring the development of a traditional , mature market such as housing. It was also concluded that Internet as a data source offers some possibilities for substituting existing indicators. Site-centric measurements are most promising as there are a clearly a limited number of concentration points in the housing market (Housing sites, Cadastre). However, as existing datasets and registers are used so intensively in the housing market (by for example VROM) it is a thin line between making these available electronically and using Internet as an alternative data source. When developing Internet as a data source possibly the main challenge is how to convince third parties that it is beneficial for them to contribute to producing reliable and up to date statistics on the housing market.

Wider economic impact

The continued digitalization of housing market information has led to new intermediaries like Funda and other housing sites. These may be linked to banking groups that may offer additional services and represent an important economic value in their own right. Apart from creating new economic activity and introducing completely new business models, their main contribution is probably to improve market transparency. Eventually, end users stand to benefit the most. On the other hand, the availability of digital housing information has triggered new search behaviour amongst consumers

Figure I: Overview of value added of using IaD-methods when measuring the housing market



Internet sources

- Housing sites like Funda: partly through user-centric measurements such as STIR and Visiscan and partly through site centric measurements and mining webstatistics
- Usenet servers and webstatistics of Kadaster Online
- Online market places and self-service sites (selling without brokerage services)

Current stats & indicators

- CBS: housing, housing stock, permits for building houses, WOZ and mortgage statistics
- VROM: housing preferences & conditions
- NVM: houses for sale & transaction process
- ABF: housing prices index & integrated housing market information & forecastingsyst.
- Kadaster: indicators based on statistics based on purchasing and mortgage notes

Added value of IAD

- Internet as a data source can be used more fully to develop new and substitute existing indicators and proxies for the housing (information) market
- Site-centric measurements most promising
- Thin line between making existing statistical and register data (collected using established methods) available electronically and using Internet as a data source

Examples of beta-indicators

- Kadaster online: no. of requests for information by type of users; type of products requested; average (real) transaction prices; popularity of the various types of mortgages
- Funda: average prices m²; popularity housing types; average length orientation phase buyers; average no. of days before a house is sold;
- Self service sites: no. of houses sold by owners themselves

IaD methods

- Network-centric: no opportunities
- User-centric: ample opportunities, partly used by market research firms
- Site-centric: webstatistics, opportunities hardly used
- Spider experiment on the Funda and marktpaats.nl websites to assess housing prices, turnover speed, most courant housing categories, etc. Proposed experiment: float between quoted prices and real transaction prices (would require cooperation NVM)

–surfing property sites has developed into a form of leisure activity for some - and fuels a trend towards more self service.

8. Case - pig market

Pigs are an archetypical conventional product. The use of pigs is strictly singular – they are solely bred for human consumption. This makes it relatively easy to describe the market for pigs and pork.

The Dutch pig market is highly professionalized. The production of pork meat has been optimized for efficiency –yield some of the lowest production costs per kilogram of pork. There is a fierce competition on price and the margins are small, sometimes even negative. This has lead to a continuous increase in scale of the sector and a large domestic overproduction (thus a heavy reliance on exports, especially to Germany). Some Dutch companies such as TOPIGS (upgrading), Nutreco (animal food) and VION (meat processing) have grown into genuine multinationals.

Dutch pork is (still) a commodity product, thus competition is predominantly on price, not on quality. However, the general quality level across the value chain is very high, mainly because of the recent concerns about food safety. This concern has also been translated into a high degree of market regulation. All pig breeders are obliged to keep extensive records of all their animals. Supply chain responsibility has been carried through to a considerably extent.

The pig market is well covered by existing (traditional) statistics. The Netherlands Bureau of Statistics (manure statistics), the Agricultural Economics Institute LEI (general agriculture statistics), and the Public Boards for Lifestock, Meat and Eggs PVE (actual market prices, import/export) all have very detailed and relatively up to date data.

Trends and developments

The most important current trends are the continued pressure to increase the scale (culminating in the concrete plans for the building of so-called "pig flats" at industrial areas, and, at the same time, the increasing popularity of biological breeding). These seem to be two opposite trends but closer inspection reveals there is really no contradiction – the broader adoption of biological breeding inevitably leads to the professionalization of the segment. As such it slowly but surely starts to emulate its original counterpart, the global industrial agricultural industry.

As a reaction to the growing professionalization of biological breeding, several parties involved in biological breeding from the early days have turned their backs on the movement and gone back to their organic roots. Consequently the Dutch pig market now seems split three ways: the conventional mass production based on fierce price competition, the professionalized biological production based on the best price/quality ratio, and the ecological/organic production that is aimed at the original niche markets for idealistic consumers.

The second important trend is the current focus on reducing the administrative burden within the pig sector. The administrative obligations with regard to the transport of pigs ("Regeling Varkensleveringen") have already been lightened.[9] Several agencies have recently been merged into one central organisation ("Dienst Regelingen" at the Ministry of Agriculture). This organisation administers several basic pig registries. Pig farmers have direct online access to these registries and can register new and/or change existing entries.

[9] However recent public indignation about the cruelty of (international) pig transports has forced the Dutch government to reverse some measures.

Research conducted

Specific research questions addressed

In this case study the particular focus was on the substitution and/or improvement of existing statistical data. Research questions were:

1. How will the asking price for one kilo of pork on the Dutch market develop in the very short term?
2. What are trends in the geographical distribution of the consumer's market for Dutch pig farmers (esp. is there a tendency to produce pigs for the German instead of the Dutch market, requiring leaner meat ?
3. What are the trends in the geographical location of new pig farms (esp. where are the multiple-storey mega pigsties established?)

Sources

Sources of information of digital footprints can be found in:

1. Basic registers from the Ministry of Agriculture (I&R registration of farm animals, data from the automated system for import and expert certification system CLIENT, data from the geographic information system GeoBOER)
2. Aggregated data from users of specific administrative applications for pig farmers (user-centric measurements conducted by the supplier of the application, Agrovision)
3. Price development on online auction sites for pigs (e.g., Teleporc) and "pig rights" (e.g., varkensrechten.nu)[10]
4. Closed intra/extranets of several big players in the pig value chain (e.g., Pigbase database TOPIGS, Farmingnet VION, Nutrace system Nutreco).

Methods & experiment

Because this case study has been one of the first conducted, no experiments have been conducted. Instead, the focus was on the detection of alternative IaD-based data sources from third parties. Given the physical nature of the product, network-centric measurements are not relevant. Site-centric measurement is possible at the online auction sites and at the public (basic registers) and closed (company) databases.[11] Agrovision is an odd but highly interesting case, where aggregated results of user-centric measurements are available off-the-shelf, together with a very good coverage of the (SME) big breeder market.

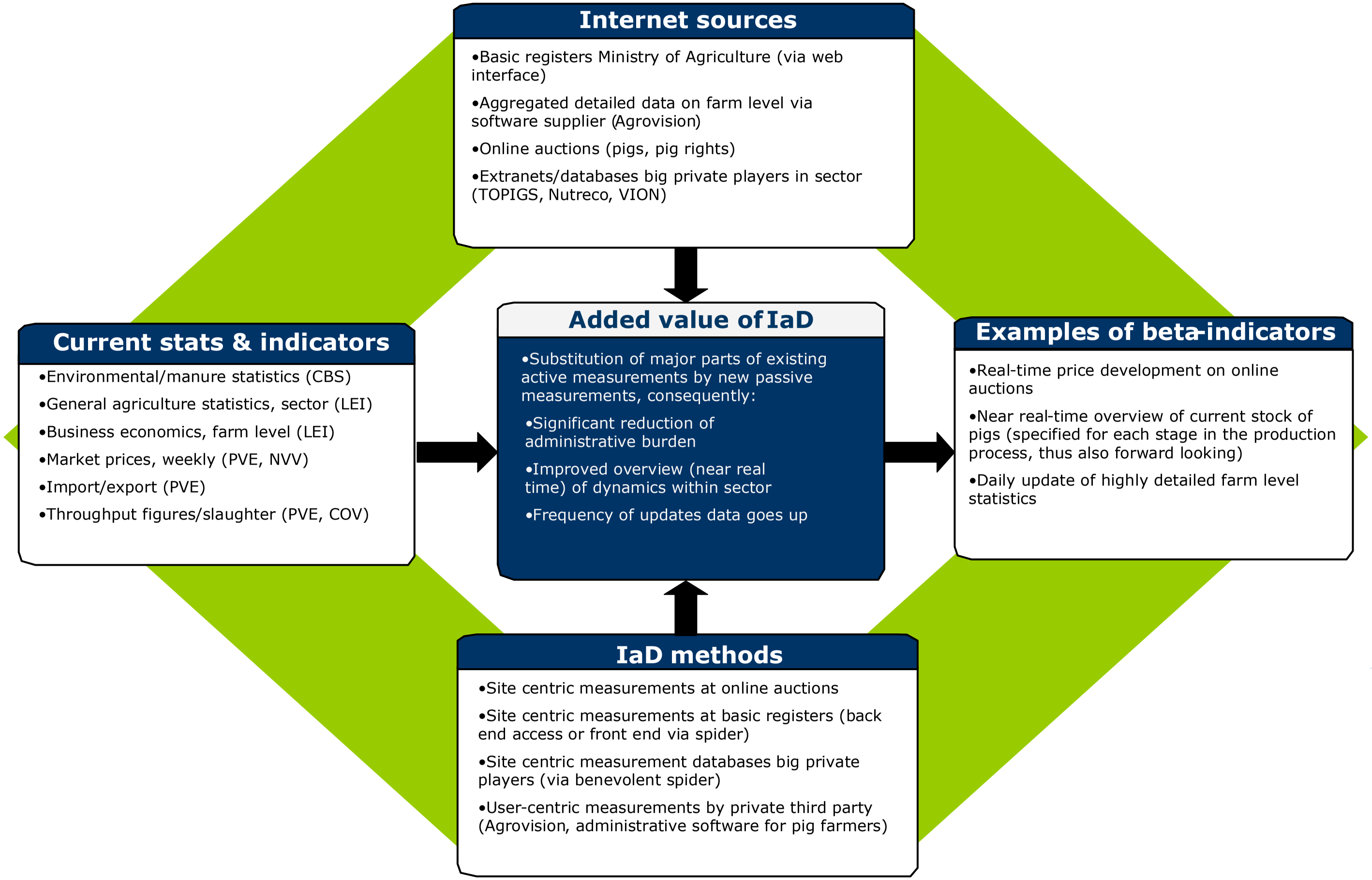
Added Value of IAD

The following conclusions with regard to the pig breeders' market case can be drawn (Figure J). The Dutch agricultural sector (including the pig market) is already well covered by traditional statistics. The very high degree of supply chain integration has resulted in a similar high degree of digitalization. Combined with the relatively straightforward layout of the market (singular use, commodity product), the market dominance of a few big players in subsequent phases of the value chain and the high degree of regulation makes this market particularly suitable for Internet-based measurements. In theory, a major part of the existing traditional active measurements (e.g., the century-old "landbouwtellingen") could be substituted by passive, automated measurements (e.g., keeping track of mutations at basic registers). In theory, the entire pig life cycle (sic!) could be followed online in real-time.

[10] The release of the htiherto georgraphically bound 'pig rights' is a strong driver for consolidation in the big breeding segment and for the establishment of enormous 'big flats'.

[11] All things considered this is not really site-centric measurement but rather direct access to content (see figure 3.1)

Figure J: Overview of value added of using IaD-methods when measuring the pig market



Internet sources

- Basic registers Ministry of Agriculture (via web interface)
- Aggregated detailed data on farm level via software supplier (Agrovision)
- Online auctions (pigs, pig rights)
- Extranets/databases big private players in sector (TOPIGS, Nutreco, VION)

Current stats & indicators

- Environmental/manure statistics (CBS)
- General agriculture statistics, sector (LEI)
- Business economics, farm level (LEI)
- Market prices, weekly (PVE, NVV)
- Import/export (PVE)
- Throughput figures/slaughter (PVE, COV)

Added value of IaD

- Substitution of major parts of existing active measurements by new passive measurements, consequently:
  - Significant reduction of administrative burden
  - Improved overview (near real time) of dynamics within sector
  - Frequency of updates data goes up

Examples of beta-indicators

- Real-time price development on online auctions
- Near real-time overview of current stock of pigs (specified for each stage in the production process, thus also forward looking)
- Daily update of highly detailed farm level statistics

IaD methods

- Site centric measurements at online auctions
- Site centric measurements at basic registers (back end access or front end via spider)
- Site centric measurement databases big private players (via benevolent spider)
- User-centric measurements by private third party (Agrovision, administrative software for pig farmers)

Wider economic impact

The physical transport of the animals is one of the most suboptimal stages in the overall production process of pork meat. It results in direct economic losses in terms of wounded, stressed or even dead animals. It is also a major liability in terms of hygiene and food safety. During the last decade there has therefore been constant pressure to minimize the physical transport of animals, for instance by concentrating all the stages of the breeding process at one location (closed system, or in Dutch: "gesloten bedrijf"). Another way to minimise transport is to conduct transactions online, as far as possible. Thus, the further digitalization of the sector (e.g., embodied in online auctions and ever tighter chain integration) fits this image perfectly.

This trend also coincides with the general policy to further reduce the administrative burden in this heavily regulated sector. Digitalization and subsequent automation of the administrative flows could significantly lower the operational costs – which could have major impacts in an industry that is characterized by very low margins. A near-real time updated administration could lead to even higher cost savings during exceptional circumstances, such as the outbreak of infections and/or contagious diseases or cases where public healthcare is somehow jeopardized.

1. Introduction

The added value of using measurement instruments rather than direct observation depends on a number of variables of which the most important are efficiency, objectivity, reliability, and validity (see Swanborn, 1987; Segers, 1999; Van der Zee, 2004). Each particular measuring instruments has specific pros and cons, hence has different scores on the variables mentioned. However in comparison to traditional methods (usually surveys), all IaD methods have the distinctive trait that they are based on non-reactive and spontaneous behaviour.[12] The fact that IaD methods rely on revealed preferences of respondents, rather than stated preferences, result in several advantages and disadvantages when compared to traditional methods. This will be discussed in the next paragraph.

On a more detailed level, each of the IaD methods has specific traits which makes it more or less suitable for use in particular circumstances. These specific issues are discussed in the last paragraph.

[12] Cf. deducing the popularity of a painting by measuring the relative wear of the carpet in front of the painting (spontaneous behavior) rather than asking visitors of the museum for their opinion (provoked behavior).

Annex 4 Statistical usability of IaD methods

2. General differences between traditional and IaD-methods

Efficiency

A measurement is at its most efficient when it is performed at the right place (where the process under study is actually occurring) and at the right moment. When the research specifically aims at the use of the Internet (e.g., the use of a specific application or protocol) IaD-methods measure right at the spot (where they derive statistical data directly from the Internet). A traditional survey, on the other hand, is always based on indirect evidence (namely the statement of the respondent). For example, when determining the actual (sic!) time spent on the Internet, it is more efficient to use IaD-methods than to use surveys.

With regard to the second point, in general, methods that rely on the detection of spontaneous behaviour are less efficient than methods that are based on evoked ("elicited") behaviour (see Swanborn, 1987). This is especially true when highly infrequent events are being studied (e.g. an application that is rarely used, or a site that is rarely visited). On the other hand, due to the low operational costs, the lack of targeted observations is offset by the fact that the measurements are usually "always on". This means that measurements are done continuously and the results can be stored for filtering and analysis (data mining) at a later stage. A major advantage is that all events are covered – also the ones that were initially not part of the experiment – and that during the post analysis new patterns may be found that may have otherwise been neglected (in a traditional research design).

Objectivity

When it comes to objectivity, IaD methods (e.g., use of spyware) obviously perform better than traditional surveys .This is due to the fact that there is no human agent involved in the collection of the data, hence no interviewer bias). This means the data is based on non-reactive and spontaneous behaviour.

With regard to the first issue it should be noted that the direct interaction between an interviewer and an interviewee is also a blessing in disguise. Because of the richness of the communication and the possibility of direct feedback the interpretation of the data is generally better than with automated data collection. For instance, the poor interpretative skills and lack of direct feedback are two of the major disadvantages when using spiders (see below) and pose a significant threat to the validity of the results.

As for the second issue, from a technical viewpoint of it is, of course, possible to use IaD-methods without informing the respondent in advance that data is being collected. This would avoid potential biases due to the "sympathy effect" that are generally regarded as a threat to the reliability of survey results, especially when sensitive issues (such as., pornography, dating, illegal downloads) are being researched.[13] If the respondent is informed about the data collection – for example because of legal obligations – the effect may still ebb over time because the respondent may forget (or get used to) to the fact that he or she is being monitored.[14]

Objectivity can also be improved by explicitly stating how the measurement has been done. This enables third parties to replicate the measurement. But this (scientific) claim is often at odds with commercial interests. There are several firms that publish statistical data based on IaD-methods (e.g., NielsenNetratings,

[13] Schmitt and Oswald (2006) have recently argued that the importance of the 'sympathy effect' (Dutch: 'sociaal wenselijke antwoorden') is highly overstated, and that ex post corrections (using 'sympathy effect scales' or specific 'profiles') have little or no added value.

[14] It is a privacy issue whether the respondent should be informed at all, only be informed on a single occasion, or be actively reminded of a monitoring presence every time s/he goes online.

ComScore, Hitwise, Ellacoya, BigChampagne) but it is often rather unclear how they actually conducted the measurements. At the very least, they should be more specific and open about the external validity of the measurements (when and where have the measurements being done, and under what circumstances). With regard to internal validity, the code of the scripts that are being used for the measurements should be (made) open (source).

Reliability

In general, the automated collection of data is more reliable than traditional methods such as surveys. In the absence of human agents, there is no variance due to subjectivity. In this respect, reliability and objectivity are closely related (see earlier). Furthermore, automated measurements are generally easier to standardise. Under similar circumstances, automated measurements will consequently return the same results.
Since it is relatively easy (and cheap) to repeat measurements, the reliability of the instrument can also be relatively easily checked. Aberrations can also automatically be detected and/or filtered.

Furthermore, in contrast to traditional surveys, which are always based on subjective opinions of respondents, IaD-methods are always based on objective quantitative data. This means that the scales of the measurement instruments are inherently more precise and easier to normalise. Although there are various statistical techniques available to improve the reliability of scales that are being used in surveys, this does not improve the quality of the underlying data. On the other hand, over the years the use of traditional surveys has yielded various highly standardized and vigorously tested scales to measure specific concepts. In the field of Internet measurement, such scales are still missing. Even worse, common definitions are missing for basic concepts such as "visitor" or "active user" (see the cases on social networking and online gaming). These are very serious issues that need to be investigated further.[15]

Validity

Last, but by no means least, we touch upon the issue of validity – the extent to which a test measures what it purports to measure (Cronbach, 1949). Note that validity is not a property of the instrument itself, but of the results of the measurement and the interpretation. This distinction is highly relevant in the context of Internet measurements, as the results are often valid but the interpretation is not. The key issue with internal validity is that the non-reactive measurement of spontaneous behaviour, on which most IaD-methods are based, gives little clues for interpretation. As long as one stays close to the original data (which here means: Internet traffic) statements are still very much valid. However when the scope of the statements is broadened (e.g., to firms) the number of alternative interpretations of the data explodes. Consequently the "semantic exclusivity"[16] can no longer be guaranteed and the validity of the statements becomes questionable.[17] Thus one should be very wary of "hineininterpretieren". The statements should only be related to the objects that are actually measured, even within

[15] A particular reference can be made to the establishment of a clearing house. This is a meeting and/or market place for the community of producers and users of statistical data (statisticians, market researchers, scientists, policy makers). One core task of the clearing house would be to agree upon common methods and definitions. See also chapter 5.3 (#3) and chapter 6 ('Way forward and the role of CBS').

[16] Refers to Swanborn's original Dutch notion of 'betekenisexclusiviteit' – hard to translate.

[17] Consider the example of the painting in the museum again (see ft.1). The fact that the carpet in front of a painting is relatively worn could also be explained by the fact that it is located near to a frequently visited object (e.g., toilets). The validity of the causal link between carpet wear and popularity of a painting can only be made if all other possible explanations have been excluded. It is, in terms of validity, much easier to deduce the most popular walking routes in the museum from the wear of the carpet (in this case, the relation is much more direct).

Internet traffic. For instance, P2P is (despite its dominance on the market) not synonymous to BitTorrent and the use of BitTorrent cannot be equated with illegal downloads, e.g. P2P can also be used to transfer files which are not protected by copyright laws.

More or less similar points can be made with regard to external validity – the extent to which the statements can be generalized to apply to other populations and/or settings than the original ones. If statements based on Internet measurements are generalized to households in general, there are obviously problems with the model, as not all persons have Internet access. Given the current high rates of Internet penetration the effects of under coverage are negligible. However the problem might reappear for specific advanced uses of the Internet.[18]

All Internet measurements are based on the digital footprint that people leave on a computer or on the Internet, not on the person themselves. The link between the digital trace and the person can never be completely established. With the exception of user-centric measurements, the most detailed level of identity is the IP-number. This number refers to a physical device (a piece of hardware), not to a person. Furthermore, most ISP's allocate their IP-numbers dynamically thus the most detailed number of identity is then the block of IP-numbers that is being assigned to that particular ISP. Once again, if the statements are only made at the level of traffic flows and not on an individual or household level the problem does not occur. An exception is network-centric measurement where it is not even known whether the particular traffic flow that is being measured is representative of the average Internet traffic (see later comments).

In general, there seems to be a trade-off between objectivity and reliability on the one hand and validity on the other hand. IaD-methods (based on non-reactive, spontaneous behaviour) have the highest score on the first two variables and traditional methods (based on reactive, evoked behaviour) on the third variable. The threats to validity are less severe when the statements only refer to data traffic. This is precisely the part of the new economy that is particularly hard to cover by traditional methods. Statistical data should not only be accurate, but also relevant, timely and actually accessible (Eurostat 2000a, 2000b, Blackstone, 1999). In fact, data quality only becomes an issue after the latter three criteria have been met (Blackstone, 2001). In the realm of "hard" data (that is, statistics on the actual use of the Internet) IaD-methods are probably more relevant than data gathered by traditional methods, definitely more current (which is a major issue in the highly dynamic emerging digital economy), and technically easier accessible.

3. Specific issues for each IaD method

User-centric measurements

The representativeness of user-centric measurements is relatively high. If the software that is being used (spyware) is installed with the prior consent of the user, the same quality levels can be achieved than with traditional panel surveys.[19] In both cases, the panel size is ultimately decisive for the data quality.

[18] The profile of early adopters might differ significantly from the average profile.

[19] When spyware is distributed without prior consent – as in the case of malevolent spyware – the spread is not random but is partly determined by historical lock-in (the point in the population where the spread has started) and by the technical profile of the user (advanced users have better security settings thus are less affected – or the other way around: spyware makes use of specific security holes in certain advanced applications). The latter case is an example of self-selection which is for instance also a major threat to the validity of anonymous web surveys (see Bethlehem, 2006).

Likewise, similar problems with "panel behaviour" arise in the case of longitudinal use.[20]

Both spyware and traffic monitoring at the level of the operating system can be linked to the user accounts of the computer that is being observed.[21] This means that the finest level of detail is a user profile. Note that this still does not always refer to the person itself. Problems might arise if several users use one general account on the same computer, or log in under different accounts. Also, collective use of applications (e.g., watching a video together) cannot be registered. Both problems could be circumvented when users have to actively disclose their identity every time the measuring software is activated. This bears much resemblance with the system currently used in the Netherlands by SKO ("Stichting Kijkonderzoek") to figure out TV viewing profiles. This would also solve potential problems with privacy (see earlier) but inevitably reinforce the "sympathy effect".

Network-centric measurement

In terms of efficiency, network-centric measurements seem to be very promising. At a central point in the Internet, all traffic that passes is being measured.[22] The problem with the Internet is that it is largely designed in a non-hierarchical way, and so such central points are missing. This means that within the network all the points have to be measured before the results can be aggregated to the network as a whole. The matter is complicated by the fact that traffic on the Internet is constantly rerouted in a highly dynamic way (low reliability).

Furthermore at individual points there might be structural errors due to the fact that ISPs often have particular peering agreements with other ISP's or with their big clients. Thus, from a technical point of view, it is difficult to determine to what extent the results are representative for the network of one particular ISP, let alone for the Internet as a whole. The problems with external validity are somewhat less severe for two-way measurements but these are notably harder to implement in a network than the less reliable one-way measurements. The reliability of the data can be improved by repeating the measurement over longer periods of time and a set of ISPs. In this way, potential structural biases can also be observed from recurrent patterns in the data.

The internal validity of the results of network-centric measurements – whether the protocols within the data stream are correctly identified – depends on the state of the technology being used. Earlier generation network-centric measurements only measure at the packet level; hence valid statements could only be made on that level (e.g., total aggregate size of the data). By – partially – opening up the packets (deep packet inspection) later generations of network-centric measurement can now measure at the protocol level. In this way, for instance, applications can be detected which use non-standard ports (most peer to peer applications), or which mask themselves (such as Skype). Deep packet inspection cannot only determine which protocols are being used, but also how these are being used. The validity of these measurements is generally very high because they are often applied in commercial settings in which the tolerance for type I and type II errors is very low.

Despite all the difficulties mentioned earlier, the use of network-centric measurements is a major advantage compared to (the much more targeted) user-centric measurements in that massive amount of data and users are involved. This means that the distribution tails are much

[20] However it might be more difficult to find participants for Internet-based panels than for traditional panels.

[21] Traffic monitoring at the side of the network (that is, on hardware) can only be done at the level of IP-numbers.

[22] In practice this is often a sample – albeit a highly representative one (e.g., 25% of a huge amount of passing traffic is being inspected).

longer and that – at least in theory – also very rare events and/or minor changes can be detected. Due to their inherently limited panel size, such events or changes cannot be detected by user-centric measurements. Consequently network-centric measurements are exceptionally suited for tracing new trends in the use of Internet at a very early stage. Since the aim is not to do statements at the overall network level, this ability is hardly affected by the low external validity of network measurements.[23] The "predictive validity" of the measurements rather depends on the availability of robust and measurable criteria (Cronbach & Meehl, 1955). The relative growth of a certain application or protocol could be such a criterion, although it remains to be seen how durable trends can be distinguished from (local) fads.

Another important property of this method is the fact that it is able to obtain data that is unaffected by a social desirability bias. This can be very helpfully in the case of illegal or shameful content.

Site-centric measurement

Site-centric measurements rely heavily on the use of spiders. Spiders are being used to automatically retrieve information from many sites (as in the case of search engine crawlers) or information from few sites. External validity is a major problem in the first case because the dataset of the Internet as such is just too big to handle, even by giants such as Google. This is more problematic, since the actual coverage rate of spiders is often unknown.[24] In either case, massive amounts of data have to be processed and filtered in a meaningful way.[25] The latter is the Achilles' heel of the method because spiders are notoriously bad in interpreting especially richer kind of data. The validity of the direct results of spiders is often low. Thus when it comes to semantic interpretation, the help of a human agent is almost inevitable.[26] Based on the ex post evaluation of the retrieved data the spider then has to be reprogrammed. The fine-tuning of spiders involves considerably efforts. The problem is less prominent when the data is less rich (e.g., only requires binary assessments such as the presence or absence of a certain object).[27]

An overview scheme of the statistical usability of the various IaD-methods is provided below (Figure K).

[23] This is, the trends that are detected might be genuine and valid but we do not know how many other potential trends (that occur in other data flows) there are. Thus the selection of the trends that are being found is always rather arbitrary and far from complete.

[24] Google is said to cover (index) between 10% and 70% of all websites on the Internet. The very wide margin between these estimates clearly illustrates how difficult it is to assess the external validity of the results of a spider.

[25] Just for the sake of illustration: the spiders of Google have so far collected about 1000 Terabytes (1000 x 10^{12} bytes) of information from websites. Google alone is said to represent 50% of all spider activities on the web.

[26] Recently much progress has been made in the field of the so-called Semantic Web. W3C has for instance introduced certain technologies such as OWL (McGuiness & Van Harmelen, 2004) and RDF (Brickley, 2003; Biddulph, 2004; Manola & Miller, 2004) that are specifically designed to make web pages easier to understand for software agents (such as spiders, RtV) and web services. However semantic web crawlers (or 'scutters') cannot interpret rich data tabula rasa – they rely on hints that (ex ante) provided in special kind of meta data tags (such as the RDF seeAlso relationship, see for instance Dodds, 2006). If these tags are missing, the intelligence of the agents is of little use.

[27] In this case we have done various experiments with the use of spiders. The validity of the results differed greatly. Best results were achieved in the marktplaats case which only involved one specific site that was crawled for one specific trait (presence or absence of a hyperlink in an advertisement). The results were less satisfactory in the product software case that involved many websites and a more complex trait (number of employees).

Figure K: Overview scheme of the statistical usability of IaD-methods

IaD-method	Data provider	Robustness (internal validity)	Representativeness (external validity)	Transparency	Longitudinal use
Benevolent spyware	Individuals (user profiles)	High. Like regular surveys, underestimates shameful or illegal behaviour. Possibly, not all activities can be monitored.	High. Like regular surveys, depends on the limited size and composition of the panel.	Very High. Looks like conventional surveys. However, spyware has to be designed transparent (i.e. open source)	High. Real-time measurements can be conducted. To avoid "panel behaviour", some changes in panel composition have to be applied. Changes in applications could require changes in spyware.
Traffic monitor at OS (user-centric)	Individuals (user profiles)	High. Depends mainly on composition of panel. Underestimates shameful or illegal content. New or uncommon protocols could be hard to distinguish.	High. Like regular surveys, depends on the limited size and composition of the panel.	Very high. Looks like conventional surveys. However, traffic monitor has to be designed transparent (i.e. open source)	High. Real-time measurements can be conducted. To avoid "panel behaviour", some changes in panel composition have to be applied.

Figure K: Overview scheme of the statistical usability of IaD-methods

IaD-method	Data provider	Robustness (internal validity)	Representativeness (external validity)	Transparency	Longitudinal use
Deep packet inspection at ISP	ISP's, indirectly every Internet user and service provider	High. All the traffic of all users in a network can be measured. But some structural bias since advanced users can hinder deep packet inspection –allowing only a shallow version.	Low. Generalizations on the total amount of traffic are impossible. Qualitative aspect are hard to generalize. ISPs usually focus on different market segments. and user characteristics are usually unknown. Measuring all the traffic of an ISP is usually impossible.	Very low. Developers of DPI usually use non-disclosure agreement. Method is not like other methods applied in statistics.	Medium. Real-time measurements can be conducted. Small changes in the infrastructure can have major implications.
Benevolent spiders	Any online data source (website, database with Internet front end)	Low-medium. Differences between initiatives, e.g. websites, hinder measurement. Structural bias since advanced sources (sites) are hard to spider. Highly dependable on the quality of the spider. Often underestimates illegal content	Varies. In highly concentrated markets relatively high –or even irrelevant since the total population is measured. In highly fragmented market usually (very) low.	Medium. Method has similarities with regular data mining.	Low. Continuous changes in the relevant set and layout of data sources make this hard.

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: 08ET11