dialogic
innovatie • interactie

Ministry of Economic Affairs

Universiteit Utrecht

Go with the dataflow!

Analysing the Internet as a data source (IaD)

Main report

Go with the dataflow!

Analysing the Internet as a data source (IaD)

Main report

Foreword

For several years, Statistics Netherlands has issued a publication entitled "The Digital Economy". This national statistics bureau documents changes within the Dutch ICT-sector, examining the growing role that ICT plays right across the Dutch economy and society. If possible, the economic impacts are also assessed. For the last two years the figures for the Netherlands have also been presented in an international context. The Dutch Government is very pleased with this publication, as it helps everyone to better understand ongoing developments, identify problems and shape our policies.

Internet is changing the world, especially in the Netherlands since the widespread uptake of broadband. It is pervasive, weaving its way into the very fabric of our society and economy. Yet these rapid changes, in both business and behaviour, seem to escape the regular statistics reports. While some current trends may show up in traditional statistics a few years from now, we need to anticipate an increasing number of changes right now.

Take the classic dichotomy of producer and consumer and how that has been altered by the arrival of faster Internet connections. The public diversifies in its use of communication services. Some consumers become prosumers, creating their own media content (YouTube etc) or even become an entrepreneur (through electronic marketplaces for second hand products). Because shelf space limitations are non-existent in the virtual world, the supply of goods and services becomes more diversified. The "long tail" phenomenon offers us far more choice. The role of social networking sites (Linkedin, Facebook, Hyves, etc) broadens social and professional interaction. Knowledge is distributed much faster. On the other hand: in a globalised world people increasingly crave for local products, services and contacts.

We sensed these seemingly new phenomena were out there, but we couldn't see them reflected in our statistics. This gave rise to the question: is it possible to use the Internet itself as a second information source in addition to traditional statistics. Perhaps there are new methods, which can act as an early warning of new trends? What else can we do to improve our monitoring system? Could certain aspects of the Internet become a reliable substitute source for some existing statistics and contribute to a reduction in the administrative workload?

In an effort to answer these and other questions, the Netherlands Ministry of Economic Affairs started a research project in 2007. Entitled "Internet as a Data Source (IAD)", the first results from this study are now being published here. Based on eight case studies, the publication describes possibilities as well as limitations of the new data-collecting techniques.

This report doesn't pretend to provide definite answers to all the questions raised above, but it is a firm first step along new routes for data collecting. We believe these new methods can be applied well beyond the Dutch borders. Almost every data-collecting institute in the world strives to deepen its knowledge and improve its interpretation of changes. Please feel free to join us in that undertaking.

We would appreciate constructive feedback and recommendations to help us further with this fascinating data-project. A special website has been launched at http://www.ictbeleid.nl as part of a range of web activities from the Netherlands Ministry of Economic Affairs (http://www.ez.nl)

Mark Frequin
Director-General Energy and
Telecommunications

Contents

Politicians and policy-makers are increasingly recognising the significant economic and social impacts associated with the "Emerging Digital Economy", otherwise known as the "Information Society". At the same time, communities of statisticians, market research bureaus and scientists have started to map, measure and assess the impact of these economic and social transformations. Increasingly, they have come to realise that these transformations are not sufficiently reflected in the set of statistical indicators available today. Bearing all this in mind, the Netherlands Ministry of Economic Affairs defined a one year R&D project entitled "Internet as a Data source" (IaD). This was designed to test the feasibility of using Internet itself as an additional or alternative data source for developing statistical indicators.

The starting point of this project was the notion that organisations and individuals increasingly leave behind a so-called "digital footprint". This builds up during various economic and social activities mediated in whole or in part over the Internet. The challenge for this project was to assess the feasibility of mining these footprints to describe socio-economic phenomena. The project also looked at ways of developing new data and indicators for the emerging digital economy (which might replace existing indicators). The research questions were twofold: what methods can be used and what can be measured with them? We identified several methods of using the Internet as a data source spread over three inherently different types of

Management Summary

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

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

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

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

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

The usability of IaD methods is also high when:

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

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

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

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

The use of IaD methods clearly rides on the waves of the trend towards increased digitalization. The reach of the digital domain is still expanding and also covers an increasing part of the traditional economy. We compare most traditional statistical work with looking through the rear-view mirror of a car. By contrast, during our research, we gathered information in a more forward-looking way: what are the new developments online and how are they influencing the economy? We believe the function of Internet based measurement can be of use as an early warning mechanism. For example, if we had used network-centric measurements over the last three years, we could have predicted the enormous success of video sharing sites or specific P2P protocols in sharing music online. Timeliness is, therefore, a very important feature of IaD methods. Sometimes it is even possible to monitor certain markets real time (pigs, housing, etc.). IaD methods may lead to beta-indicators for the emerging digital economy. These indicators are a new category of socio-economic indicators for measuring the emerging digital economy more closely. They produce results with a clear early warning quality, but without the same statistical rigour or quality obtained from established indicators as published by statistical bureaus. Conversely, data generated by IaD methods might be used to calibrate (adjust) or even validate (confirm or reject) established traditional statistics.

Recommendations

IaD methods measure at the heart of the matter, namely changing patterns in data traffic. The fact that there is a true need for new types of statistics is witnessed by the rapid grow of various "Internet-based statistics" that are offered by commercial third parties. The quality of these commercial statistics is often rather dubious and unknown at best. We have made a plea for a clearinghouse function for these Internet-based statistics. Statistical agencies such as Statistics Netherlands could provide this function, but they will have to be stimulated or commissioned to do so. They are in the best position and well equipped to develop into a key player in this type of research.

Government agencies are first and foremost users of statistics, be it traditional or innovative statistics like IaD. The important question is not *whether* IaD methods should be used but rather *how* they should be used. Sometimes there are simply no alternatives to the use of IaD methods; existing statistics loose their relevance due to their lack of timeliness and/or because entire sections of the emerging digital economy are not covered at all.

An important advantage of IaD methods is that they generally offer a cost-effective solution and can significantly reduce administrative tasks. Privacy, however, is an issue of growing importance in the use of IaD methods. Under certain circumstances, privacy concerns might even block the use of IaD methods altogether. Therefore, the decision to use IaD methods for either substituting or complementing existing data that is collected by traditional methods should not be taken lightly. There is a trade-off between efficiency, objectivity, timeliness and cost-effectiveness on the one hand and validity and privacy on the other. Saying that a measurement is non-intrusive from a technical viewpoint may still mean that it can be highly intrusive from a privacy point of view. To boost the practical usability of IaD methods, the legal framework should be organised in such a way that the critical privacy and security issues can be resolved.

A very important role of government is to stimulate further research into the feasibility of IaD methods and to facilitate experiments. It is clear that using IaD is still in its infancy. We believe there is a strong need for further experimentation and research.

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

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

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

1
Introducing Internet as a data source

1.1 The emerging digital economy revisited

The Dutch economy is transforming gradually into a digital economy. At the same time, Dutch society as a whole is developing into an advanced Information Society. Dutch consumers buy billions of Euros worth of goods online, as well as an increasing number of services. They buy directly from online stores, producers with an electronic outlet or through rapidly growing electronic marketplaces. Social networking sites such as Hyves (a Dutch equivalent of Facebook), Schoolbank.nl or LinkedIn –– are being used by more and more companies and individuals for professional reasons (such as recruitment). In this respect, The Netherlands is like many other advanced economies.

There has been a swift rise in the use of applications such as YouTube, "uitzendinggemist.nl" (a site which allows the public to catch up on public TV or radio programmes they missed), as well as online gaming and music sites. All represent further examples of how the character of media consumption is changing. More companies are getting used to the idea of collaborating within virtual production networks. They are developing multi-channel strategies for interacting with clients. Information and Communication Technologies (ICTs) are facilitating new modes of production and opening up whole new niches of economic activity.

ICTs are not only empowering the typical representatives of the "new economy" such as the creative and ICT industries. They are also having an impact on typical "old economy" industries such as housing, agriculture, tourism and the hospitality industry. Here completely new digital (information) markets have emerged – even though the goods and services themselves are still traded physically.

Yochai Benkler in his book the Wealth of Networks (2006) very neatly describes how the most advanced economies have developed into what he coins the "networked information economy". He documents the rise of non-market and radically decentralized production and identifies two central shifts that led to this type of economy, namely (Benkler, 2006, p. 3):

1. "an economy centred around information (financial services, accounting, software, science), cultural (films, music) production, and the manipulation of symbols (from making sneakers to branding them and manufacturing the cultural significance of the Swoosh)";
2. a clear move "towards a communications environment built on cheap processors with high computation capabilities, interconnected in a pervasive network – the phenomenon that we associate with the Internet".

Not only scholars such as Benkler, but also politicians and policy-makers are increasingly recognising the significant economic and social impacts associated with the Emerging Digital Economy (EDE) or Information Society. Simultaneously, communities of statisticians, market research bureaus and scientists have started to map, measure and assess the impact of these economic and social transformations. They have all come to realise that these transformations demands new policy answers, new statistical measures and new research to keep track of the impact of these changes.

This has led to serious efforts by the OECD, Eurostat and several national statistical bureaus to measure the unfolding Information Society. This resulted in sets of indicators that provide some insight into the readiness, intensity and, to a partial extent, the impact in the use of ICT.[1] In fact we are witnessing a growing divide between the current statistics and indicators used to characterise the EDE and the complex, rapidly blurring of boundaries that can be observed in this EDE. This blurring of boundaries applies to

- industries that are increasingly interwoven and difficult to disentangle,
- individuals that, for example, switch between various roles (consumer, co-producer, citizen)
- new categories of producers and products that are hard to map using existing categorisations.

This implies that some phenomena linked to the emerging digital economy are not being covered by standard statistics. The development of new, well-defined and coherent statistical indicators usually takes several years, especially if international comparability is needed. The means that by the time the new indicators are ready, the phenomena may have changed beyond all recognition, or even evolved into something else. Examples of these so-called "blank spots" include new media consumption, the rise of social networking, various categories of electronic marketplaces, as well as entirely new professions linked to the emerging digital economy.

1.2 Central research questions and some points of departure

Against the background outlined in the preceding section, the Dutch Ministry of Economic Affairs defined a one-year R&D project entitled "Internet as a Data source (IaD)".

[1] It was recently acknowledged by OECD itself that "official statistics in the area of ICT impacts are generally not well developed" (OECD, 2007, p. 28).

The Ministry wanted to test the feasibility of using Internet as an alternative data source for developing statistical indicators. The starting point for the idea was that politicians and policy-makers could benefit from the wealth of data that can be derived directly from the Internet. It was felt that the same might also hold true for other sectors, and that statisticians, scientific and market researchers could profit as well.

The Internet now offers numerous new data sources, which can be effectively "mined" to describe both the "traditional" economy as well as the "new" economy (and other broader areas of society) more effectively and efficiently. It was decided to investigate whether IaD could be used for covering the blank spots that could not be filled using existing statistical indicators. Additionally IaD might also be used for producing an unofficial set of statistical (new, extra or substitute) indicators that could signal how the EDE is changing in a more real-time or even forward-looking fashion. In the course of the project this type of indicators were termed "beta-indicators". This also includes picking up weak signals on changes in data traffic and use patterns that might point to changes in economic and social processes associated with the EDE. For developing this type of indicators new methods need to be explored that really make use of the processes of digitalisation that are ubiquitous in the EDE. The aim of the IaD-project reported here is therefore twofold:

1. To identify new data and indicators derived directly from the Internet as well as to map and describe new phenomena associated with the EDE. This includes regular business processes being used in various industries and markets. For the most part, these new phenomena are not covered by established, regular statistics. In practice this means using Internet as a data source for developing new indicators and possibly substitute indicators for the EDE.
2. To explore and assess the usefulness of the various IaD-methods for deriving new and substitute data and indicators for the EDE. This means assessing whether by using IaD could use less invasive methods of data-collection, so that respondents (firms, households, individuals) do not have to answer survey questions. The project should describe the pros and cons of these methods in detail.

These two aims can be combined as will be shown in the set of 8 experimental case studies. The main lessons learned from these case studies are reported in chapter 4 and described in somewhat more detail in Annex 3.

Let us now consider four points of departure that, to a large degree, define the approach we have chosen, namely:

- **The ongoing digitalisation and the resulting digital data explosion are at the heart of measuring the Emerging Digital Economy.** As will be explained in Chapter 2 in greater detail, the essential (and revolutionary) characteristic of the Internet is that it allows a very efficient, direct two-way communication between two nodes in a network irrespective of the technical infrastructure that is used for transporting the data. Nodes are defined here as computers themselves or products in which computers are embedded and which generate data. As the number of nodes continues to explode, as well as the speed of connectivity, the frequency at which they exchange data and the volume of the datastreams they generate also increases rapidly. This is facilitated through rapidly falling prices for processing power and bandwidth. All these factors mean the digital domain is constantly expanding while the analogue domain is either static or declining.

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

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

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

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

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

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

1.3 Approach adopted and outline

The Internet as a Data-source project is an R&D project. Researchers in cooperation with statisticians, policy-makers and other stakeholders (represented in a steering committee) assessed the usefulness of Internet as a data source for characterising the EDE. One of the key dilemmas at the start was either to focus on covering "blank spots" in current statistics on the EDE or to focus mainly on assessing the possibilities of IaD methods. Eventually it was decided that the emphasis of the study was to be testing the core IaD methods by undertaking eight case studies.

The following activities were performed, mostly concurrently (see Annex 1 for more details on the research approach adopted):

- Conceptualisation of an IaD-model and methodological approach, including a paper outlining the usability of spiders for gathering data and building statistical indicators;
- Analysis of international sources, both established indicators on the EDE and (mostly) examples of studies, using Internet as a data source;
- Conducting 8 case studies using a fixed format in webstores - a leading electronic marketplace in the Netherlands (Marktplaats.nl), product software market, music market (mostly online), Internet TV market, online gaming market, the market for social networking and more traditional markets such as the ones for houses and pigs;
- During these case studies various small experiments were performed using spiders. These are reported separately in the individual case study reports and are synthesized in Chapter 3. Considerable efforts were made to organize a network-centric measurement in the Netherlands using state of the art deep packet inspection software from Ipoque, a German company. The various partners involved have met and discussed in very practical terms both the possibilities and limitations of performing such measurements. This proved very useful in assessing what can be done using advanced network-centric measurements.
- Overall analysis focusing on 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).

In this final report, Chapter 2 first discusses the potential of digital footprints and some conceptual considerations when using Internet as a data source. Chapter 3 looks at the various IaD methods and present their most important pros and cons. Chapter 4 summarizes in a stylized fashion the results of the 8 case studies and lessons learned from these case studies. In Chapter 5 the strategic policy implications and relevance of using Internet as a data source are given and suggestions made as how to proceed. These include some suggestions for further research and experimentation. In chapter 6 we present the overall conclusions. The annexes are presented in a separate document.

2 The potential of "digital footprints"

2.1 Introduction

In 1999 we developed a monitor for measuring E-commerce. Back then we observed that we see "products, actors and business processes become more digital in character, or are given shape and form by digital means".[2] This process of digitalization has progressed much faster and further since then, to a point where it may provide more direct opportunities for measuring economic phenomena associated with the EDE.

The project starts with the assumption that increased digitalization means that companies, organizations, as well as individual consumers and citizens leave a traceable digital footprint when they engage in economic and social activities that somehow involve the Internet. We do so in our electronic quest for information, by ordering and paying for goods and services or arranging the transport of these goods. These digital information streams are often generated through "transactions"[3] between users and producers of this information.

[2] See, den Hertog, Holland and Bouwman (1999, p. 19). We then used the well-known model of Choi, Stahl and Whinston (1997, p. 18) as one of the building blocks for measuring E-commerce. This is still relevant as an element of a conceptual model for measuring the emerging digital economy.

[3] The notion of 'transactions' need to be interpreted broadly as this is not necessarily a traditional economic transaction. These transactions could equally be transactions between users among themselves (in all sort of peer-to-peer type of applications) or between users, companies and authorities (as part of a registration process for example).

Eventually, participants operating in the EDE are carrying an "electronic cloud" around with them, which increasingly defines and signals their activities. We have labelled this as a "digital footprint", rather like environmental footprint marking the sort of use and impact individuals, firms and society make on the environment. We believe that digital footprints are not the addition or end result of all economic activity, but a by-product that can be mined to assess the development of these economic activities. The challenge of this project is to assess the feasibility of mining these growing electronic footprints for describing socio-economic phenomena. From this we want to develop new, extra or substitute data and indicators for describing the EDE.

In this chapter we will first examine relevant broader technological, economic and societal trends that provide the context for the ongoing process of digitalization (section 2.2). We then briefly discuss the basics of this process of digitalization and how this translates into a data explosion, which may be mined for constructing statistical indicators (section 2.3). Subsequently we introduce two simple matrices. The first one combines the dimension type of market and elements in a value chain in order to identify and map typical concentration points of digital footsteps in markets (section 2.4). The second matrix once again points at the fact that Internet as a data source may also be relevant for mapping and describing established industries and markets (section 2.5).

2.2 Some relevant trends in Technology, Market and Society

In their book, Information Rules, authors Varian and Shapiro (1999) state that technology changes, but economic laws do not. If we are to believe some of the opinion leaders in the current Web 2.0 or participative web debate, this statement by Varian and Shapiro is now up for discussion. Examining the most popular websites at this moment (YouTube, MySpace, Wikipedia, Facebook, Hyves, etc.) we see that users are adding value. These active users (Alvin Toffler coined them pro-sumers in the early 1980's) are willing to spend time, contribute their knowledge and skills, make available their computing power etc for a cause that is not commercial, non-proprietary and more commons based.

Brynjolfsson (2005, p. 51) has used the term *Gift Economy* to address this new phenomenon. Web 2.0 is all about connecting people and providing these people with the tools to create, share, tag, comment, review and recommend the so-called *user generated content*. The OECD (2007) states that most user-generated content activity is undertaken with no expectation of remuneration or profit. Benkler (2006) introduced the phrase peer production and argued that sharing or collaborating on a production can be considered a new modality of economic production.

Does this mean that economic laws might be changing after all or is this just change in some very specific markets and in consumer behaviour? Have many Internet users suddenly become creative content producers instead of passive couch potatoes? Are we all digital volunteers now fighting for all the right causes instead of just being self-absorbed consumers? In a paper by Pouwelse et al. (2007) the authors state that there is a lot of emphasis on this *(Good) Samaritan Side* of peer production, but there is also of course a less positive aspect that they call the *Pirate Side*.

This pirate side of peer production is something we can observe in Peer-to-Peer (P2P) networks. According to the authors, P2P file sharing can also be seen as a form of peer production because it creates vast libraries of content. The pirate side confronts us with matters such as illegal downloading, privacy issues and spam that accompany the spectacular growth in this aspect of the Internet.

First of all, we intend to elaborate on the various trends that, in our vision, underpin the digital economy. These trends are a combination of new developments in technology, markets and demographics as well as user behaviour.

2.2.1 Trends in technology

First and foremost, there is a trend that "everything goes digital" and "everything goes IP". This means that all kinds of information, media and business processes (information gathering, ordering, payment, logistics, service etc.) will be increasingly transmitted through the Internet. The resulting giant "digital cloud" engulfs more and more of our daily personal life, but also business activities and the work of public sector organisations. The digital cloud or data-explosion is something we can measure, for it contains not only data in the form of text, images and sound (e.g. music, MP3) but also radio, TV and moving images. Also, more and more devices are equipped with an IP address and all sorts of intelligent web services enable communication and information exchange between these devices (application2application).

A second trend that serves as a technological driver is the increased broadband availability. The Netherlands is considered to be one of the leading countries in terms of broadband infrastructure. The availability of this infrastructure triggers new services that are dependant upon fast and reliable connections and high quality of service. When the infrastructure is no longer an issue, development concentrates on content and content based services.

A third trend in technology is that the network (the Internet itself) is becoming the platform that provides all sorts of functionality (software tools, scripts, applications, webservices) and data. Increasingly so, applications are available that allow users to create and distribute content in a highly decentralised way. After the mainframe and PC era, we are witnessing now the concept of distributed functionality where connected users are offered nearly unlimited access to data and services.
A fourth trend in technology has to do with standards for exchanging and structuring data on the web. The World Wide Web Consortium (W3C) has proven to be very successful in developing interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential. XML is the best-known example. It is a simple, very flexible text format in which form and content are separated and it plays an increasingly important role in the exchange of a wide variety of data on the Web. Another relevant phenomenon is OpenAPI (the term API stands for Application programming interface) which refers to sets of technologies that enable websites to interact with each other by using SOAP, Javascript any other web technology. While its possibilities aren't limited to web-based applications, it's becoming an increasing trend in so-called Web 2.0 applications.

The last trend in technology is the possibility to use the Internet with a variety of mobile devices thus enabling Mobile Internet applications, location based services etc. Interesting possibilities are e.g. Google mashup, by which maps and geographical information are combined with data from other sources into a single integrated tool; an example is the use of cartographic data from Google Maps to add location information to housing and real-estate data, thereby creating a new and distinct web service that was not originally provided by either source.

2.2.2 Trends in economy/markets

An important economic driver is the low entry barrier for all sorts of service providers who are doing business on the web. The lower costs of hardware and software tools, of connectivity etc. serve as a big stimulus for new business. The low costs and increased availability of software tools and connectivity also stimulate users (prosumers) to become more active (also in terms of user generated content) and spend more time online.

The second trend is that increasingly users are adding value. User generated content is becoming an important economic phenomenon, despite the fact that most ventures start as non-commercial. The impact on traditional media and content suppliers is significant. Digital content is increasingly important across all media and publishing industries and is becoming pervasive in sectors not previously considered to be content producers or users (e.g. business services) and in the public sector (public sector information such as weather information, public sector content such as archives, and cultural content), education and health. This adding value by users in the form of peer production has increased commercial interest and investment by traditional media and publishing industries, mobile operators and telecommunication providers.

The third economic trend is long tail economics, which refers to the possibility to produce and distribute small volumes of products in a very cost-efficient way. The long tail (Anderson, 2006) is in fact a demand curve that flattens, because the sum of many niche goods can rival with the best sellers in markets like books, music, movies and other media. Because of the low cost of production and distribution, combined with very powerful tools to connect supply and demand (e.g. search engines, profiling technology, recommendation systems etc.) new business opportunities arise where they previously did not exist.

A fourth trend is the emergence of new business models. There is great flexibility in ways to get paid for a specific value proposition on the web. Income can be derived from providing a marketplace (brokering), generating traffic (affiliate model, advertising model), selling products or services (merchant model, subscription), generating leads, demand aggregation, data mining (e.g. in consumer profiles) etc. A firm may combine several different models as part of its overall Internet business strategy. Business models have taken on greater importance recently as a form of intellectual property that can be protected with a patent.

The last trend in the realm of economy and markets is the fact that active users are taking over functions and activities in various value chains. Self-service by users can mean that in certain value chains disintermediation of traditional players (e.g. travel agencies as a result of online booking) will be inevitable.

2.2.3 Trends in demographics and user behaviour

The third category of trends deals with demographics, Internet usage and consumer behaviour. First of all, there is a generation aspect to active use of the Internet. The OECD (2007) refers to these active users as digital natives. These young users have many years of online experience and therefore have the skills to use all sorts of tools that are available on the web.

Secondly, there is an apparent need among these young users to express themselves and to connect and communicate with others via the web. In the Netherlands, three quarters of all Internet users are a member of one or more social networking sites (Ernst & Young, 2007). Also, the need for entertainment is very well addressed by websites such as Youtube. The OECD mentions that almost three quarters of people who publish amateur video content online are under 25 (OECD, 2007, p. 29).

A third trend that constitutes a social driver for the EDE is Internet usage. In the Netherlands, people spend, on average, 8 hours a week online (STIR, 2007). The growing popularity of the Internet is affecting the use of other media. The time spent with traditional media (newspapers, radio, television) is dropping.

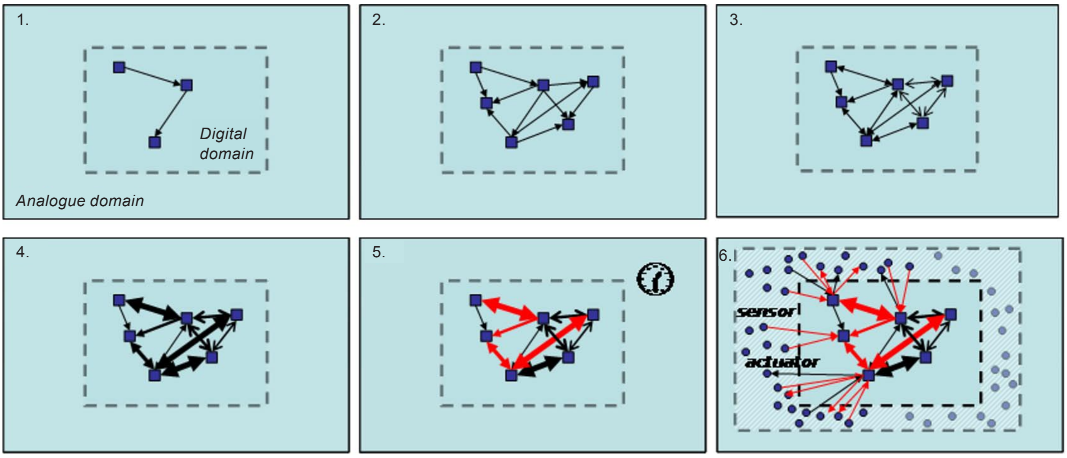
In our opinion, it is the *convergence* of these trends in technology, markets and user behaviour that constitutes a disruptive force. This may have a profound impact on the digitally enabled economy, especially in markets and sectors that are *highly digitized*. By this we mean that both the products and services are highly digitized or the business processes are highly digitized; sometimes even the economic participants can be highly digitized). Carlota Perez (2004) refers to these developments as a new techno-economic paradigm: in her opinion, new all purpose technologies, new organisational principles, different business models and low cost facilitating infrastructure may all contribute to a quantum leap in productivity for all economic and social activities.

2.3 Digitalisation

The notion of digitalisation has already been introduced in Chapter 1. Digitalization is not limited to a few selected product markets or business processes but is a pervasive trend, which is omnipresent in current society. It affects ever more domains, processes applications and uses. Digitalisation is also closely interwoven with another mega-trend in this part of the world, namely the shift from production to services. In the latter perspective, physical products are merely regarded as the basic layer of economic activity – increasingly, added value is generated in and by the information streams that surround these physical products. Thus even when a product itself cannot be digitized (e.g., a house or a pig) the production, trade, delivery and consumption of these goods generates massive flows of information that can be digitized. Even if a house as such is not being sold over the Internet, such a transaction often leaves a digital footprint somewhere on the Internet.

The overall increase in digitalisation is itself

Figure 2.1: Digitalisation as a mega-trend



driven by a number of underlying fundamental trends. In a logical order (and summarized in Figure 2.1 below):

1. The *number* of digital nodes is increasing. The total number of computers and Internet users continues to rise across the world. In addition to this, software is increasingly embedded into a great number of products. All these nodes generate data flows.
2. The *connectivity* between the nodes increases. Information processing nodes are embedded in a dense global mega computer network (a.k.a. the Internet) that is increasingly meshed.
3. There is a shift from one-way (push/pull) to *two-way* (dialogue) relations. In computer networks data and information is increasingly exchanged in both directions. Interactive television is just one of the many manifestations of this trend.
4. The *volume* of the data flows increases. This is mainly due to the shift from text and static graphics towards audio and video. Video-based applications have become very popular lately. The phenomenal growth of YouTube is perhaps the most striking example – this single video site alone is said to generate 10% of all Internet traffic worldwide.[4]
5. The *frequency* of data exchanges increases. There is a shift from static to streaming and event-driven to real-time exchanges. Examples range from weather to agriculture and finance.
6. The traditional physical analogue domain and the virtual digital domain are getting increasingly interwoven. In a growing number of domains the analogue domain is observed by sensors and interacted upon by actuators. These devices often communicate with each other or with other nodes (see 1) on the Internet (see 2). Greenhouses, for instance, often automatically self-regulate the inside temperature and climate based on input from external weather sensors; computers in cars are diagnosing problems themselves and offer real-time reporting to mechanics and so on. In a sense, then, the digital domain is invading the analogue domain.

Taken together, all these trends have already caused an enormous explosion of data traffic. A major part of these traffic flows are through the Internet and are, therefore, relatively easy to trace. The remaining flow occurs within all kinds of closed and/or local networks. In principle, from a technical viewpoint at least, these flows are still traceable or could at least be made accessible for statistical data collection. These increased data flows could be used for more accurate, timelier data that can be made available at a lower cost, without interfering with the work of individuals and companies. The methods that could be used for this are discussed in chapter 3.

2.4 Digital data concentration points in markets

If digital footprints can be found just about everywhere, where do you start collecting information? In order to locate digital footprints, or possible data sources in the various case studies, we constructed a matrix. This combines market-types on one axis, limiting our research to three types: Business-to-Business (B2B), Business-to-Consumer (B2C) and Consumer-to-Consumer (C2C) and on the other axis a simple generic value chain or ordering model which could apply to almost every market or economic activity. We are aware that other variations are possible, if we consider government as a separate player.

[4] See Ellacoya (2007). According to this study, based on several network-centric measurements (see Ch. 3), 46% of all data traffic worldwide comprises of HTTP, of which YouTube takes 20%.

We have limited ourselves here to four stages or activities, starting with information on the product or service and ending with fulfilment. In between a distinction is made between the actual ordering and payment activities. When we combine these two dimensions, it is possible to generate a picture in which different digital data sources or concentration points can be located per market. The purpose of this matrix is to assist in finding possible (alternative) data sources or potential concentration points that can, in principle, be used for constructing new, extra or substitute indicators. By way of example, we have described the market for recorded music in figure 2.2.

2.5 Digital data sources for the emerging “new economy” and established “old economy”

Using digital footprints to measure or describe markets or economic activities is by no means limited to the emerging digital economy. The Internet as a data source equally applies in old economy or “established” markets. In several case studies (housing, pigs, see Chapter 4) we came across a wide variety of digital data sources because of the high level of digitization of relevant economic participants and processes.

Figure 2.2: Potential IaD concentration points, the example of the market for recorded music

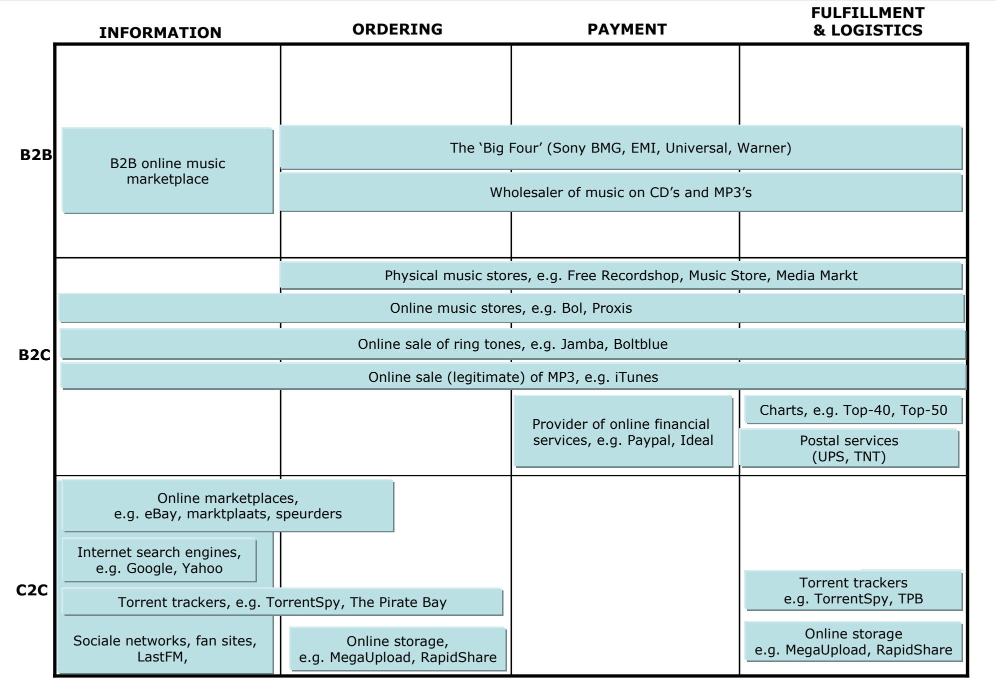

INFORMATION
ORDERING
PAYMENT
FULFILLMENT & LOGISTICS
B2B
B2B online music marketplace
The ‘Big Four’ (Sony BMG, EMI, Universal, Warner)
Wholesaler of music on CD’s and MP3’s
B2C
Physical music stores, e.g. Free Recordshop, Music Store, Media Markt
Online music stores, e.g. Bol, Proxis
Online sale of ring tones, e.g. Jamba, Boltblue
Online sale (legitimate) of MP3, e.g. iTunes
Provider of online financial services, e.g. Paypal, Ideal
Charts, e.g. Top-40, Top-50
Postal services (UPS, TNT)
C2C
Online marketplaces, e.g. eBay, marktplaats, speurders
Internet search engines, e.g. Google, Yahoo
Torrent trackers, e.g. TorrentSpy, The Pirate Bay
Sociale networks, fan sites, LastFM,
Online storage, e.g. MegaUpload, RapidShare
Torrent trackers e.g. TorrentSpy, TPB
Online storage e.g. MegaUpload, RapidShare

Figure 2.3: What we can measure and how we measure it: four examples

	"Old economy" (established) markets & phenomena	"New economy" (emerging) market & phenomena
Established (analogue, mostly invasive) data collection methods	(1) ICT investments in industries measured	(2) New media use by final users through a survey among a panel of households
Internet-based (digital, mostly non-invasive) data collection methods	(3) Price of pigs traded over electronic markets through a network-centric measurement	(4) Share of illegal content in P2P traffic as measured through a network-centric measurement

While the product or service can be physical, the information flows within the market are highly digitized and thus can be a source for measurement activities and statistical analysis. The matrix in figure 2.3 combines two dimensions that are central in the research we conducted.

The first dimension on the horizontal axis is what we wanted to measure. Here we see established markets and "old economy" on the one hand, and emerging markets and phenomena ("new" economy; digital economy) on the other hand. The second dimension deals with how we measure: established ways of data collection versus new, Internet-based methods. Of course there is a thin line to distinguish between the two methods because established ways of data collection are increasingly based on using registers or digital tools to collect the input.

It is important to stress that established ways of data collection almost always include a survey and hence a respondent who is asked to fill out a questionnaire or provide data in some other form. Here we see a fundamental difference when compared with the idea of using digital footprints, since these are collected or generated without the user being surveyed in one way or the other. This is one of the biggest advantages of Internet based methods: they are very efficient and non-invasive. It is the non-invasive aspect that is especially important, especially in the current climate of looking for a reduction in administrative burdens and costs.

The last row in the matrix (examples 3 and 4) points out the two segments on which we focussed our research. Internet based methods for data collection is our point of departure, but we looked at both established and emerging markets (or phenomena).

3 Methods using the Internet as a data source - their pros and cons

3.1 Background

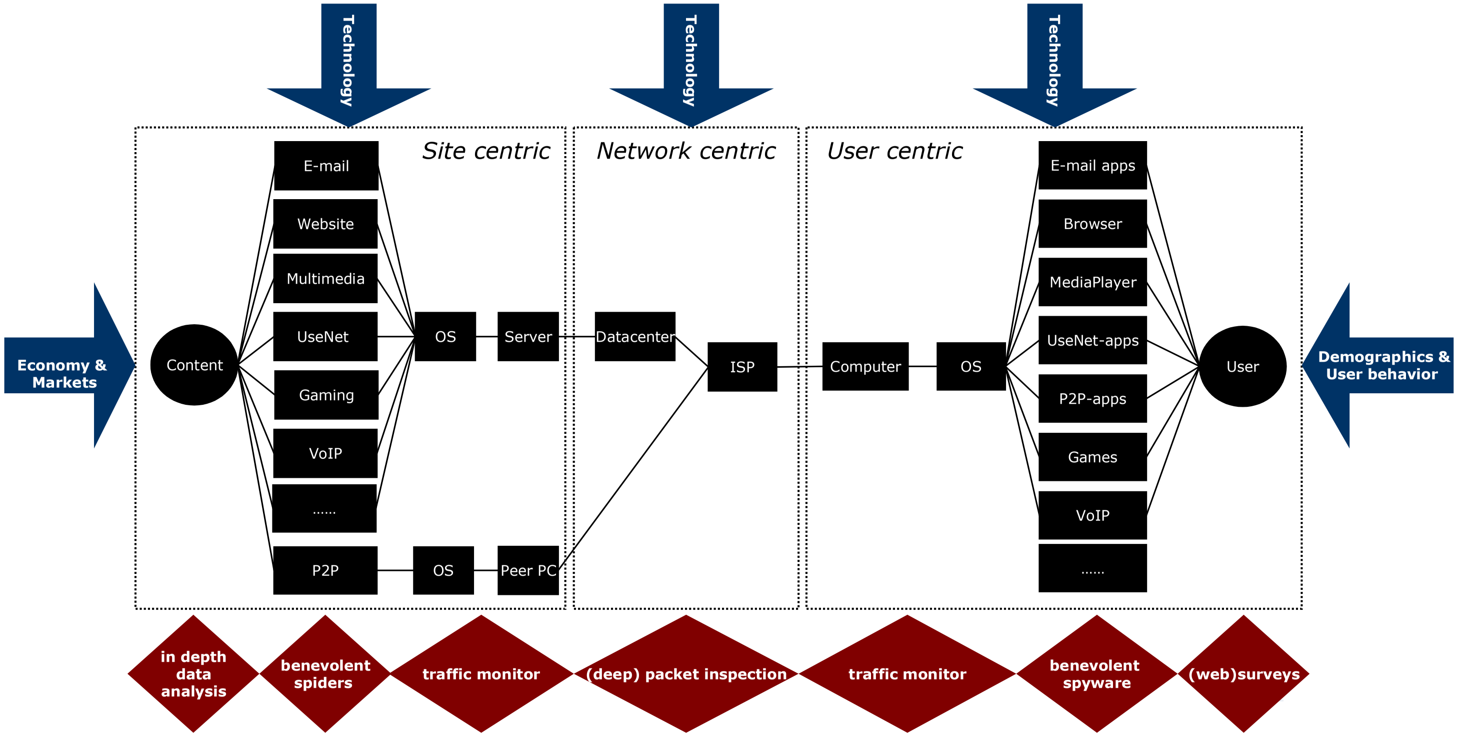
As described in the previous chapter, the emerging digital economy is strongly intertwined with the growth of the Internet. Thus, in mapping the growth of the digital economy, the development of the Internet itself is one of the central topics. Because ever greater parts of the economic and social interactions are dealt with over the Internet – and inevitably leave their digital traces – the Internet has become an important site for data discovery in itself: this Internet as a Datasource (IaD).

Using the Internet to directly gather data has three major advantages.

- First, the automated collection of data has relatively low start-up and operational costs compared to traditional methods.[5]
- Secondly, data collection does not require the collaboration of respondents. This could significantly reduce the administrative burden.

[5] But note that savings from labour/capital substitution are partly offset by higher initial costs (esp. development costs).

Figure 3.1: Operational implications of using Internet as a data source



- Thirdly, since changes in the behaviour of households and firms almost immediately result in corresponding changes to data flows, Internet-based measurements make it possible to track shifts much faster than traditional measurements.

The use of IaD measurements is therefore particularly promising for highly dynamic areas – such as the emerging digital economy.

3.2 Introducing the various IaD methods

One of the most important challenges of using the Internet as a data source is to handle the enormous amounts of data that are involved (see paragraph 2.3 on digitalisation).[6] In order to find our way through this flood of data we have used the following conceptual model :

The model is based on the way data flows from a particular user to a particular piece of content. Demographic and/or social trends cause changes in online user behaviour, in that it changes the demand for certain types of content. Shifts in demand lead to shifts in supply (or the other way around), thus generating new economic possibilities and eventually entirely new market structures. The process in which online demand and online supply are linked is obviously strongly influenced by technological changes.

Social, economic and technological changes interact. The rise of visual culture, for instance, is the basis for the increased demand for video images. The continuous increase in affordable bandwidth has enabled the online exchange of such images. This has generated new initiatives in the market (such as the establishment of specialised video sites like YouTube) and these initiatives have in turn influenced general user behaviour (e.g., boosting the re-use of digital video images).

Based on the actual location of the data, a split can be made into three general types of IaD measurements. Going from the user (right) to the content (left) these are user-centric, network-centric and site-centric measurements. Within each type, specific measurement methods are available.[7]

User-centric measurements trace changes in behaviour at the computer of the individual user. At the most detailed level, *benevolent spyware* can be used to track each specific application for what types of content are being used. At a more general (*operating system*) level, it is possible to determine which types of users use which type of applications (traffic monitoring as being used in software firewalls).

As soon as there is communication with other computers – which is by definition the case for Internet applications – changes in patterns will also occur at the network level. The recent shift back from P2P to HTTP-traffic can for instance largely be attributed to the phenomenal success of YouTube. Such shifts can be traced by network-centric measurements. The most basic types can only generate aggregate data (e.g., trends in total volume of data). However, recently more sophisticated methods have been developed which are able to actually look within

[6] At the Amsterdam Internet Exchange (AMS-IX), one of the central regional nodes in the Internet, during peak hours every second 400 Gigabits of data passes by.

[7] These are represented are red diamonds at the bottom of the figure. A detailed description of the methods can be found in Appendix 2. Note that two methods on the outside (web surveys and data analysis) are, strictly speaking, not IaD-methods. Although both methods involve the use of digital data, neither of the cases produce digital footprints that are left behind on the Internet. The distinctive difference is that IaD measurements are based on non-reactive and spontaneous behavior (that is, make use of digital footprints) whereas traditional methods such as surveys are not (see also Appendix 4).

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

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

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

[8] See again Appendix 2 and Chapter 4.

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

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

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

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

3.3 User-centric measurements

User-centric measurements provide detailed insight into user behaviour. The most specific method (spyware) enables measurements at the level of individual user accounts. Typical research questions that could be addressed by user-centric measurement concern the use of specific applications by specific user types. Examples in the case studies include the use of benevolent spyware to gather detailed data from an administration application used by pig farmers or to monitor the online behaviour of gamers.

But this high degree of detail comes at a price. User-centric measurements make use of panels, which are costly and thus often limited in size. This makes it difficult to detect small changes and thus to detect early trends. In terms of privacy, a similar trade-off occurs. User-centric measurements are the only type of IaD-methods that allow for targeted messages to respondents that their behaviour is being monitored. On the other hand, the use of (benevolent) spyware at the personal computer of respondents implies a serious security risk as third parties might abuse the access for malevolent uses.

The compilation and the support of the panel also comprise the main costs in the deployment of user-centric measurements. Although traditional surveys also use panels, the costs are higher because privacy concerns will probably complicate the search for panel members. The unavoidable establishment of a helpdesk will further affect the support costs.[11] Although each individual application requires a specific type of spyware, the development costs of software for user-centric measurements are relatively low. This is because the measurement generally concerns standard applications and/or uses. In the latter case, readily available firewall software might even provide a cost-effective tool for measurement.

The (largely fixed) support costs and, especially, the development costs become less of an issue when user-centric measurements are being used at a larger scale. The scalability of user-centric measurements is actually very high because the measurements are hardly affected by local circumstances (e.g., language) and concern standard applications that are generally the same across national boundaries.[12] Furthermore, given the full control over the composition of the sample (the user panel), standard panels can be used for each country.

3.4 Network-centric measurements

Network-centric measurements concern the flows between users (user-centric) and content (site-centric). At one or more central points on the Internet, all traffic that passes these points is being measured. Because the measurements are being done at the network itself, it is possible to cover a great number of users and types of content simultaneously. It is therefore a highly efficient method. Furthermore, because of the massive amount of data involved it is possible to detect even minor changes in the dataflow. Thus, in contrast to user-centric measurements, network-centric measurements are especially useful to trace and track new trends in the use of Internet from a very early stage.

[11] However part of the support can be automated, such as the central distribution of software updates. In the case of benevolent spyware, this is an important issue because the frequency of updates is rather high (for traffic monitoring software the frequency is much lower). Every change in an application requires a new version of the particular piece of spyware that has been installed to track that specific application.

[12] For instance, a central repository (or at least a core set) of scripts can be developed than can be deployed anywhere. On top of that, country-specific subsets can be used.

A current example would be the use of direct download links instead of P2P-applications for the (illegal) distribution of digital content.

A major drawback of network-centric measurements is that the external validity of the results is relatively low. This is mainly due to the particular (non-hierarchical) design of the Internet and the particular market structure of Internet access.[13] For instance, small changes in the infrastructure can strongly influence the measured results. As a result, although the latest technology (*deep packet inspection*) makes it possible to perform highly reliable and rather detailed measurements at a high throughput speed, it is difficult to generalize the results of a specific measurement to an aggregate level. In other words, the changes that are detected are real, but there is uncertainty to what extent they are representative of changes in Internet usage as a whole.

Because network-based measurements intercept aggregated data flows between users and content, they are particularly suitable for identifying illegal and/or shameful behaviour. However the very fact that users are – or cannot even be – aware that they are being monitored raises important privacy issues. Inspecting passing-by traffic goes against the basic principles of the (free) Internet (see, for instance, Sugden et al., 2003). Even the mere discrimination between various types of data traffic is already in conflict with the principle of net neutrality.[14] It is therefore very difficult to find ISPs that want to cooperate – i.e., willing to place measurement equipment directly on their network – because they are extremely wary about informing their subscribers that certain traffic is being prioritized (and other traffic muffled), let alone that traffic is being monitored.[15]

The fact that the laborious process of reaching agreements with individual ISPs has to be repeated on a case-by-case basis is particularly troublesome in the light of the low external validity of network-centric measurements. External validity could, after all, be improved by measuring at (several) multiple points throughout the network.
In terms of costs, in the case of deep packet inspection, for each particular application that is investigated tailor-made modules have to be made. This is similar to spyware but now involves a rather complex process of re-engineering digital footprints. Consequently, the development costs of deep packet inspection software are relatively high compared to spyware.[16] The same largely applies to operational costs because the scope of deep packet inspection is substantially larger than the scope for spyware (and certainly for traffic monitoring). The number of footprints constantly grows and the footprints themselves constantly change. As a result, the software needs to be

[13] It is usually not possible to cover all the traffic of an ISP. Furthermore, most ISP's focus on different market segments and the specific characteristics of the users are unknown.

[14] Many ISP's do in fact already deploy deep packet inspection but "for technical reasons only" (namely the optimization of Internet traffic). For example, some network administrators limit the amount of P2P traffic between 06.00h and 20.00h to 50% of the maximum capacity of the traffic.

[15] The strong resistance of ISPs is in fact the reason that we were not able to conduct network-centric measurements ourselves during this project. However, such measurements have been done by third parties in other settings (e.g., Ellacoya, Ipoque, Hitwise).

[16] However, similar to spyware the same repository of footprints can be deployed everywhere on the Internet. Thus from a technical point of view (but less so from a practical point of view, see before) potential for upscaling is high and the high initial costs can be recouped relatively easy.

frequently updated.[17] Given the massive amounts of data involved and the fact that the data has to be processed in real-time, the use of deep packet inspection often requires heavy deployment of (expensive) hardware.[18]

3.5 Site-centric measurements

At least in theory, a spider can also access every online data source that can be accessed by a user. Spiders are, therefore, widely used for the automatic retrieval of data. The main feature of spiders is their ability to obtain very specific data. They mirror, in fact, spyware at the server side. However because spiders are not bound by specific applications (but rather to specific sources of information) they are more flexible than spyware. Spiders can be used to find information on almost every given topic on the Internet. The scope of a spider can be either very broad (covering multiple sites) or very narrow (covering a specific part of a specific site). Well-known examples of the first type of spiders are search engines like Google. A similar "broad" type has been used in the software case (crawling websites from product software companies). Spiders targeted to one specific site have been deployed in several of the other case studies, for example in online market places (marktplaats.nl), housing (funda.nl), social networking (hyves.nl), and music (bol.com and the P2P application Limewire).[19]

Regardless whether a spider covers a small part of many sites, or a large part of a few sites, massive amounts of data have to be processed as well as filtered in a meaningful way. In general, it pays to have some intelligence already built into the spider to reduce the efforts needed for filtering and diagnosis of the (large amounts of) raw data afterwards. Spiders are, however, notoriously bad in interpreting especially richer types of data. This means that the development and fine-tuning of an "intelligent" spider (e.g., an adaptive agent) takes a lot of time and (costly) human effort. Thus, whereas the development costs of simple spiders are relatively low, similar costs for more sophisticated spiders can be much higher. In the case of targeted spiders, development costs cannot easily be recouped because they are tailor-made for a specific setting (a particular site in a particular language) and, therefore, do not scale well.

In terms of operational costs, much depends on the dynamics of the content. When the content or structure of a website changes, the software has to be adapted. Again, this often involves a lot of skilled manual efforts. Next to this, most webmasters are not particularly fond of visits by crawlers and they will try to block access to and/or mask the content of their website or databases (see next paragraph). Thus, besides the autonomous fine-tuning of the software and the constantly changing content, the spider administrator is often also involved in a technological rat race with the website administrators.

Privacy is less of a concern in the case of site-centric measurement then in the case of user- or network-centric measurements. Information that is made accessible to regular users is

[17] But note that in most cases the business model of the firms that sell deep packet inspection products (such as Ellacoya or Ipoque) is based on the primary use of deep packet inspection (namely optimization of traffic flows at an ISP). The re-use of the data for statistical purposes is considered as a by-product and priced as such – the costs for development and operations are already made anyway. Hitwise is an exception since it sells the data generated by deep packet inspection, not the product itself.

[18] Much depends on the specific way in which the measurement application has been designed. In more sophisticated designs, scaling is partly dealt with at the software level. In this way, traffic flows of up to 10 Gigabit per second can be handled with relatively cheap off-the-shelf dedicated hardware.

[19] See Annex 3 for the stylized results of the eight case studies that have been conducted during this project.

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

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

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

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

4.1 Introduction

As outlined in the introductory chapter, the best way to test the usability of the Internet as a data source for developing new and alternative indicators is to build case studies and perform some small-scale experiments in each. Six of a total of eight case studies were performed in areas, which are typical blank spots for currently available statistics on the EDE domains. They include music, Internet TV, webstores and electronic marketplaces, product software, social networking and online gaming). Two were performed in more established markets (housing and pig market) that have been affected by digitalisation, but where statistics are readily available.

In each of these eight markets we assessed the usability of Internet as a data source for developing new, extra and substitute indicators by analysing readily available statistics, conducting interviews within the industry (especially – potential – data providers). In most of the markets we conducted small-scale experiments using spider technology (for more details on these see Annex 1).

4
Feasibility of using Internet as a data source: eight case studies

The selection of case studies was made to ensure we included examples of markets that differed on three dimensions, that is by the:

1. degree to which these markets were "digitalized"[22];
2. type of market differentiating between Business-to-Business; Business-to-Consumer and Business-to-Consumer markets or combinations of these[23];
3. typical examples of "old" economy and "new" economy markets and phenomena[24].

This is schematically presented in Figure 4.1 below.

In Annex 3 we provide an overview of the results from these 8 case studies. The complete case studies are included in a separate report, but since this is in the Dutch language we decided to make two of the eight case studies available in English, one about on online TV (Brennenraedts et al., 2008) and the case study on gaming (te Velde, 2008).

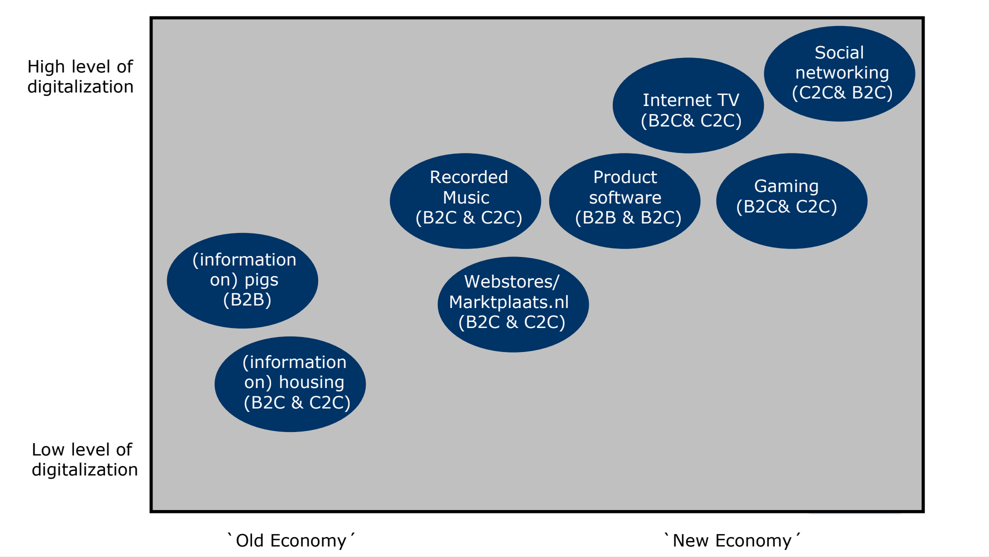
In this chapter we point to the added value of this set of eight case studies (section 4.2). Subsequently we reflect on the main lessons learned (section 4.3) and the wider economic trends and impacts reflected in the case studies (section 4.4).

4.2 Assessing the value added by using IaD methods

In the case studies we systematically looked at the following four categories:

- Current stats & indicators: here we assessed the statistics and statistical indicators which

Figure 4.1: Cases selected

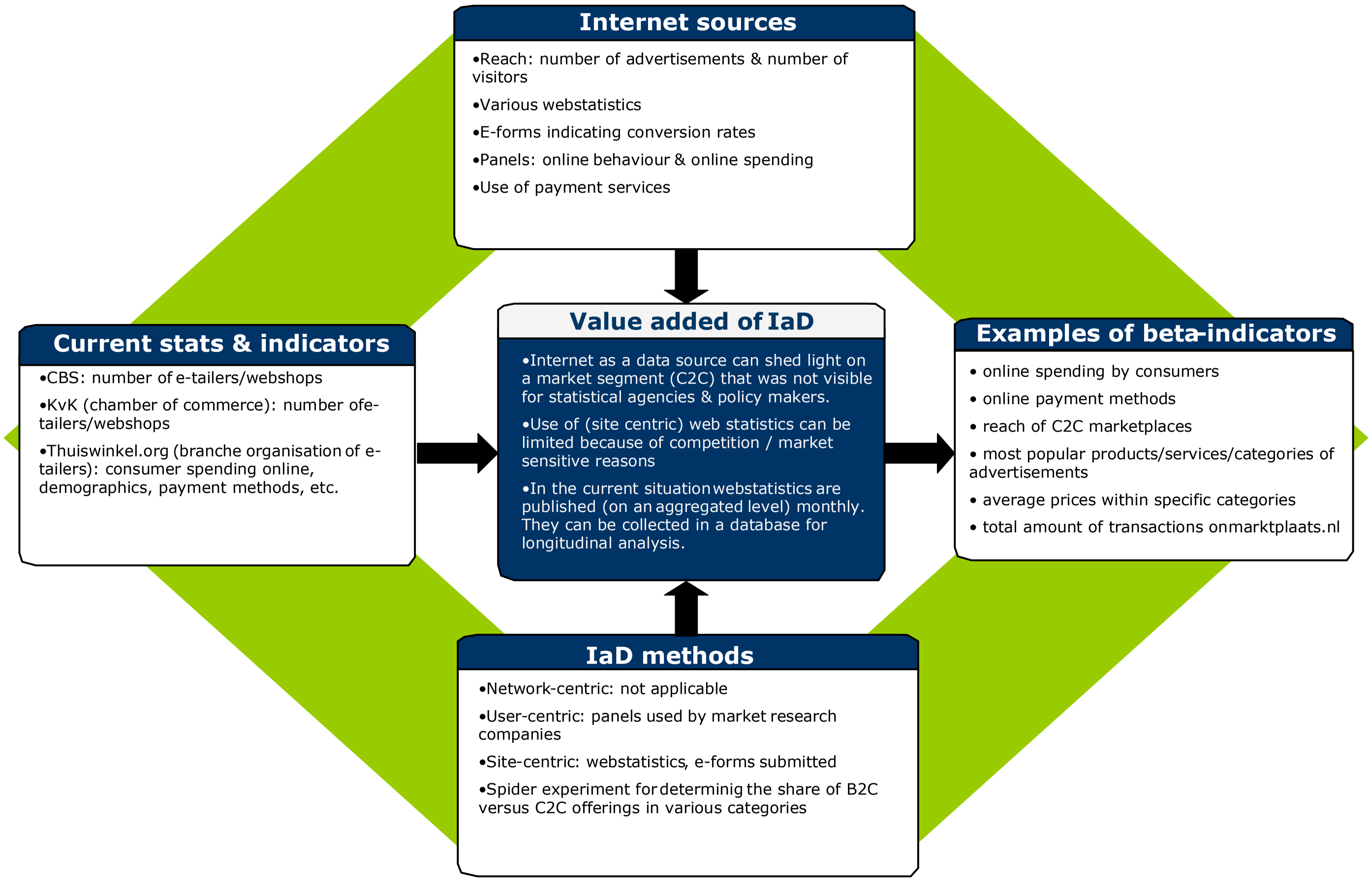

High level of digitalization
Low level of digitalization
Social networking (C2C& B2C)
Internet TV (B2C& C2C)
Recorded Music (B2C & C2C)
Product software (B2B & B2C)
Gaming (B2C& C2C)
(information on) pigs (B2B)
Webstores/ Marktplaats.nl (B2C & C2C)
(information on) housing (B2C & C2C)
`Old Economy´
`New Economy´

[22] Ranging from analogue product and digital information (1) to an intermediate level of digitalization where at least the market function was (partly) digitalized (2) to all digitalized markets where the product, the information and the marketplace as well as the relevant business processes were digitalized (3).

[23] We are aware of markets and applications where government plays an important role, but these were omitted, as it was impossible to cover all combinations.

[24] There is a certain bias towards 'New Economy' markets as these are typically markets regarded as "blank spots" in, for example, the Statistics Netherlands publication on the Digital Economy. Both principal and members of the steering committee had a preference for trying to cover those "blank spots".

Figure 4.2: 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 aggregated 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

are readily available in the market under investigation, looking both at official statistics (as produced by national bureaus of statistics) as well as other intermediary type organizations, industry associations and larger players in that specific market;

- Internet sources: By drawing a figure with typical concentration points of electronic information per market (see section 2.4) we tried to assess where electronic data streams might be available in the market;
- IaD methods: on the basis of the overview of possible IaD methods as presented in chapter 3 we established what type of measurement was feasible in that particular market. In most of the case studies we ran small spider experiments to see whether we could collect some relevant data ourselves and to see what sort of practical barriers we encountered.
- Beta-indicators: on the basis of the three steps outlined above, we were able to provide examples of what we labelled as "beta-indicators". It is important to note that the list of beta-indicators that we came up with per case study is a combination of suggestions for statistics and indicators that can be collected by using IaD (but which we did not collect) and indicators that we collected experimentally.

On the basis of these four categories we drew up overview schematics for each of the eight cases. We then formulated the value added or key conclusions regarding the feasibility of using Internet as a data source for providing basic characteristics, structure, conduct and performance of that specific market. An example for the webstore/(marktplaats.nl) case is given below (figure 4.2). Similar schemes are included for all 8 cases in Annex 3.

The practical usefulness of Internet as a data source differs considerably between the eight cases, as shown in Figure 4.5 later on in this section. But, in almost all cases, it provided insights into (sub) markets that had either never previously been described statistically or only partially. IaD especially proved helpful in describing recent trends and issues in these markets. We observed the speed at which relevant, detailed and sometimes near real-time data can be provided on existing, but especially, new phenomena in the emerging digital economy. At the same time it proved difficult to meet the established statistical standards used commonly by statistical bureaus, in terms of robustness, validity and representative sample size. The trade off between relevance and statistical quality is real. The case studies also illustrated the methodological and practical pros and cons of the individual methods (see Chapter 5).

Before discussing in section 4.3 the general lessons learned we highlight some of the results of the case studies schematically.

Figure 4.3 gives an idea of some of the practical data collected in the individual case studies. The figure also shows whether the indicator is completely new (i.e. not described statistically before), is an extra indicator (i.e. a more detailed indicator in a market that has been already described to some extent using regular statistics) or a substitute indicator (i.e. an indicator replacing an existing indicator derived from regular statistics).

Figure 4.3: Some examples of IaD indicators (some experimental) in the 8 cases and their refresh rate

Case study	Examples of practical data collected using IaD (not exhaustive)	Type new/extra/ substitute
Webstores/ marktplaats.nl	• Marktplaats.nl has 7,5 million unique visitors over the month August 2007 and a total number of visits of 72,6 million in that month	New
	• In the first 6 months of 2007 6,2 million people in the Netherlands bought something online; for a total amount of 1,84 billion Euros	Extra
	• Total value of transactions on marktplaats.nl in 2006 is 4.7 billion Euros	New
Online music	• Approximately 70% of the Internet traffic consists of P2P. Music files make about 1% to 3% of this stream. [25]	New
	• The three most downloaded music files in Germany are "Mia – Stille Post", "Amy Winehouse- Back to Black" and "Top 100- Hotpop charts 2007" [26]	New
	• Limewire users share mostly audio files (90%). Video makes up 5% of the files and 5% contains of a broad set of file types	New
Internet-TV	• Approximately 70% of the Internet traffic consists of P2P. Music files make about 60% to 80% of this stream. [27]	New
	• The three most downloaded video files in Germany are "Next", "Fantastic Four – Silver Surfer" and "Spiderman 3" [28]	New
	• Online broadcasting of proceedings by city councils in the Netherlands is conducted by 50% of the municipalities with more than 200.000 inhabitants.	New
Online Gaming	• Total global value of transactions in sales of virtual objects by dedicated firms (secondary real money trade, RMT)	New
	• Profile of hardware (graphics card) and software (operating system) used by users of the online game Half-Life	Extra
	• Total number of active users for any major massive multiplayer online role playing game (World of Warcraft, Guildwars)	Substitute
Social networking	• 74% of Dutch Internet users has a profile on one or more social networking sites	New
	• Hyves.nl had 5.3 million profiles in December 2007	New
	• On the basis of a spider experiment, the number of active users on Hyves is 3.8 million (69%)	Substitute
Product software	• From the portal www.softwaregids.nl, the websites of 741 product software companies can be stored (spidered) for retrieval and analysis	New
	• The year of foundation can be retrieved from the websites of 33% of these 741 product software companies, showing that the average company has been active for 13.9 years	Extra
	• The majority of the retrieved/spidered product software companies use the words "partner" (67%) or "independent" (50%) on their website	New
Housing	• 1,813,623 visitors visited Kadaster Online in October 2007 (by type of user, including private households)	New
	• Funda (largest Dutch housing site) has 3 million unique visitors during the month March 2007	Extra
	• On October 1st 49.551 houses were on sale via marktplaats.nl (real time)	New
Pigs	• Market price at 10:43 at the online pig exchange Teleporc	Substitute
	• Average growth per animal per day	Substitute
	• Average asking price for pig breeding rights (weekly basis)	Substitute

[25] Ipoque (2007) Internet study 2007.

[26] Ipoque (2007) Internet study 2007.

[27] Ipoque (2007) Internet study 2007.

[28] Ipoque (2007) Internet study 2007.

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

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

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

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

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

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

Figure 4.4: Examples of practical research questions addressed and resulting proposals for beta-indicators in the 8 case studies

Case study	Research questions	Example of beta-indicator (not exhaustive)
Webstores/ marktplaats.nl	• What is the reach of marktplaats.nl? • What is the share of B2C offerings versus C2C offerings? • What is the total amount/value of transactions?	• Online spending by consumers • Online payment methods used • Reach of C2C market places • Most popular products/services/advertisements (cat) • Average prices within specific categories • Total amount of transactions on marktplaats.nl
Online music	• Which music is made available for (illegal) file sharing? • What is the popularity of (illegal) P2P transfer for recorded music? • Current average price of a CD or music-DVD?	• Size of music catalogue on offer • Real time average prices • Real time insight into shared music • Share p2p music in total web traffic • Most popular files transferred
Internet-TV	• Who makes use of Internet-TV? • What are properties of the offer of Internet-TV? • Share of Internet-TV in the total amount of Internet traffic?	• Real time indicators on consumer behaviour • Total offer of video content (amount, type, size) • Percentage video (http, P2PTV) of the total Internet traffic
Online Gaming	• How has the number of players of online games developed over time and what is the current reach? • What are the three largest online games (MMORPGs, casual games) – in the Netherlands and worldwide? • How do these large online games perform in terms of market share and turnover?	• Total numbers of active users (better estimates) • Total size of Real Money Trade/Virtual objects • Average hardware profile used by gamers (lead users for graphic cards) • Percentage of MMORPG players spending over five hours a day online (indicator for gaming addition)
Social networking	• What is the reach (number of users) + development in time of SNS? • What are the three largest SNS (NL and worldwide)? • What are the effects of SNS on the economy?	• 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
Product software	• What is the size and composition of the Dutch product software sector in the Netherlands according industry statistics and the existing product software portals? • How many websites from product software companies can be spidered in order to retrieve company features such as number of employees, year of foundation and the presence of business software terms?	• Number of websites of product software companies in the .nl domain that can be spidered for retrieval and further (trend) analysis • Age of Dutch product software firms • Terminology used on product software company sites towards customers and the business/industry
Housing	• Can Cadastre Online be used as a proxy for development of the property market (prices, market demand, number of transactions, transaction speed)? • Can Funda.nl be used for assessing the developments on the housing market (prices, market demand, transactions), housing preferences potential buyers? • Is there a real trend towards selling houses without a broker/estate agent ?	• Number of requests for information at Cadastre Online by type of user, type of products, etc. • Popularity of various types of mortgages • Average housing price/m2 per housing type, region, etc. • Number of houses for sale through self service sites
Pigs	• How will the demand price for one kilo of pig meat on the Dutch market develop in the very short term? • 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 – more demanding, less fat/more meat –? • What are the trends in the geographical location of new pig farms (esp. where are the multiple storey mega pig pens established?)	• Real time price development of pigs at online auctions • Current stock of pigs at various stages of the production cycle • Average % of fat in meat • Comparison in price trends for pig breeding rights between regions (e.g., shift away from South – where most pigs are bred at the moment).

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

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

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

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

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

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

Figure 4.5: Summarizing overview case studies: market characteristics, number of concentration points and added value of IaD

	nature product (digital or not)	Level of digitalization value chain	Level of regulation	Level of market concentration	Geographical scale	Availability concentration points	Value Added IaD (sec)
C2C marketplaces	□	■■■	■■	■■■■■	■■■■■	■■■■■	■■■■■
Houses	□	■■■	■■■	■■■■	■■■■■	■■■■■	■■■
Pigs	□	■■■	■■■■■	■■■■	■■■■■	■■■	■■■
Internet TV	■	■■■■	■■■	■■■	■■	■■■	■■■■
Online Gaming	■	■■■■■	■	■■■	■■	■■■	■■■■
Social networking	□	■■■■	■	■■■■	■■■■	■	■■■■
Recorded music	■	■■■■	■■	■■	■	■■	■■■
Product software	■	■■■■	■■	■	■■■	■■	■

4.3 Lessons learned from the case studies

IaD helps in signalling new trends, developments and phenomena. We saw that IaD is not really a substitute for existing data collection methods and indicators, but instead offers great opportunities for signalling new trends, developments and phenomena that cannot be tracked using existing methods and statistics. There are important developments in markets that are unknown at present, not being seen or measured by statistical agencies and policy makers. When we compare most traditional statistical work with looking through the rear-view mirror of a car, then we believe our research gathered information of a more forward looking nature: what are the new developments online and how are they influencing the economy? The function of IaD-based measurement can be used as an early warning mechanism. If we had used network-centric measurements during the last three years, we could have predicted the success of video sharing sites or specific P2P protocols in sharing music online. An important lesson, therefore, is to look at new methods and new data sources in terms of a trade-off between statistical quality demands and the relevance of signalling new developments in the EDE.

The mix of IaD methods that can be applied will vary widely between markets and industries. An important second lesson we learned after conducting eight case studies is that there is no "one size fits all" approach that can be applied in the majority of markets or industries. In fact, each product, service, each specific economic activity in the value chain and each market has its own digital footprint and therefore provides very specific opportunities for using this digital data source for new, extra or substitute indicators. The mix of methods that can be used will vary widely between markets and industries (making comparison between markets and industries very difficult).

Digital footprints may be available for markets not directly associated with the EDE. A third lesson learned is that the level of digitalization is not necessarily determined by the degree to which a product or service is digital. There is a clear difference between the degree to which a product or service is digital and the availability of digital information on a specific product or service. For this reason, it could well be the case that some industries and markets associated with the \`old economy´ can be tracked more effectively using the Internet as a data source than products and markets associated with the ´new economy´ which are highly digitalized themselves. The two most traditional markets in our selection of case studies, i.e. the housing market and the pig market, are increasingly dependent on electronic information services and marketplaces in order to function properly.[30] Housing sites have developed such widely used, powerful applications that they offer opportunities for statistically describing the development of the housing market. The market for pigs can be monitored in almost real time, not because the product itself is digital, but because the information streams in the market are highly digitized.

The level of market digitalization (or information thereabouts) is dependent on some basic market characteristics. A fourth lesson learned (see also Figure 4.5) is that the usability of the concept of Internet as a data source is, naturally, determined by the availability of digital sources. This level of digitalization is dependent on various aspects. First of all, the type of product or service and the type of market are important factors to be taken into consideration. More specifically, various market characteristics such as its market, concentration level (dominant players), frequency of transactions, level of governmental regulation and the market scale (national/international)[31] are important.

Industry itself has already started to mine digital footprints. A fifth lesson is that in some markets, digital footprints are already substantial and their use is on the rise. In a several markets individual actors or market research bureaus have started mining these digital footprints (or are about to start). This implies that market players themselves – with or without the help of market research firms - will increasingly provide statistics about their markets and industries. In the online gaming industries, for instance, some firms directly measure the number of players using specific titles. In the software market for farmers, specialised suppliers of administrative software register all kinds of administrative information that can be used for assessing the pig market. In markets where, typically, one or a just a few players are dominant (social networking, property markets), some simple web statistics may already provide significant insight into the market or industry whereas traditional statistical bureaus do not yet have indicators available.

EDE markets are more fuzzy and diffuse. A sixth lesson is that the case studies revealed that some economic activities associated with the emerging digital economy – mostly new

[30] More generally, market places seem to offer good opportunities for assessing the trends in some markets such as the type and number of goods sold through marktplaats.nl, trade facilitated through marktplaats.nl or the split between (semi-) professional and amateur traders.

[31] IaD methods are probably less suited for describing those markets that are highly internationalized. Some markets such as the online gaming market and the online music market are internationalized and it is therefore harder to gather data and develop national statistics.

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

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

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

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

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

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

costs and the higher frequency at which data are available when compared to the "regular statistical indicators". Therefore IaD statistics can – in the absence of established indicators – play a role in especially those areas. In markets that are much better covered (pigs, housing) IaD proved to be instrumental as well, mainly because quite a number of indicators are almost always readily available, more frequent and the methods for collecting these are less invasive (opportunities for reducing the administrative burdens). So opportunities for substitution are most real in established markets and industries. Internet as a data source can be very helpful in mapping typically the rise of these new type of applications and new type of industries and markets and so point at economic activity that represents a considerable economic value in itself, but is not yet or only marginally covered by regular statistics.

4.4 Wider economic trends and impacts

Several of the typical economic trends associated with the EDE (see Section 2.2) are substantiated by the various case studies. Peer-production (as a result of lowering entry barriers) and the increased transparency of markets are the most dominant of these trends. Peer-production, co-creating consumers or a more active role of users in economic production is already reality and most evident in the C2C marketplaces, recorded music, Internet TV, and social networking cases. In these examples, consumers produce at least some of the content and participate, at least partially, under conditions they set themselves. But also, in housing and the pig industry, users play a more profound role. Users have largely taken control themselves in their quest for a new house. Some even decide to sell their property without a broker/estate agent. They may have opted for new types of broker services where users have a much more active role in the whole process of selling a house. In the pig market users produce all kinds of market information themselves collectively, which can then be used in their work processes. Market transparency has increased considerably mainly due to the availability of these new services and advanced search engines. Electronic marketplaces give a detailed insight into the supply and demand of product and service categories, which can be often specified in great detail. Markets for housing, product software (due to self-registration systems) recorded music and pigs have become much more transparent.

These two trends highlight related developments such as the rise of very specialized and sometimes obscure offerings on the Internet (such as fly fishing trips to Eastern Siberia, or first prints of Dutch literature before 1800). These highly specialized markets indicate the "long tail" is real. Further they lead to new business models for services offered. The winning business model in the recorded music market has changed dramatically and has considerably affected traditional music recording companies. In the property markets the housing sites have mushroomed and these have triggered traditional brokerage firms to step forward with new service offerings. The advertising on social networking sites can pinpoint specific target groups very effectively and this has had consequences for traditional media. Also, social networking sites are entering the arena of job search and labour market communication – and this too will have an effect on traditional players in this market.

In addition, the case studies provided helpful cues in signalling wider socio-economic impacts of the further digitalization of the various markets analyzed. The case studies do seem to point to a substantial rise in new types of economic activities. The estimated turnover realized through the leading C2C marketplace in the Netherlands is possibly most illustrative here. But consider also the rise of these C2C marketplaces and what it will mean to the development of electronic payment systems, the new types of logistics that emerge (micro logistics), the impact on regular retailing and so on and so forth. Some of the cases also point to new forms of illegal or at least "grey" economic activities such as the use of P2P networks for distributing audio and video.

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

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

5 Strategic implications and relevance of using IaD

5.1 Introduction

It is evident from the preceding chapters that Internet as a data source provides relevant sources of information for various markets with specific characteristics. It is clearly relevant for politicians, policy makers (see section 5.2.), researchers, statisticians, market research firms and industry associations in the private sector. The clearest added value of IaD is that it provides relevant insight into markets and phenomena where the established statistical agencies have no information.

Using IaD may help in spotting and capturing relevant trends that are unknown and/or are not adequately described by existing statistical sources and methods. These new developments and economic activities impacting on our economies are still in the "blind spot" of policy makers, researchers and statisticians. IaD can provide the first insights, a snap shot of the emerging digital economy in a quick and timely (near real time) fashion. In some cases it can even act as a substitute for existing indicators and data collection methods, leading to reduced administrative burdens and consequently lower costs.

Some of the material we collected in the case studies can be used as beta-indicators. Beta-indicators are fine for characterising relatively new phenomena for which no statistical indicators exist yet. The statistical quality is in general poor, though some of the beta-indicators can be developed into indicators that meet the kind of quality criteria that are normally used by statisticians (validity, robustness, year-on-year availability, etc.). It may also be possible to address the issue of poor statistical quality by validating beta-statistics with the help of existing statistics.

This chapter deals with the strategic implications and relevance of the various IaD methods. First we will elaborate on the policy implications and the potential of IaD (section 5.2.). Subsequently we will make some practical suggestions for additional research and experiments (section 5.3). The statistical pros and cons of IaD are discussed in Annex 4.

5.2 Usability from a policy-makers' perspective

As a collection of methods, IaD has proven to be helpful in mapping new trends, phenomena and markets. The arrival of IaD has given us the opportunity to signal and measure completely new phenomena that have either been ignored or only been very marginally measured in existing statistics (e.g. Social Networking, C2C marketplaces, Internet TV). This was also what was expected by the organisation that commissioned this research project, namely the Netherlands Ministry of Economic Affairs. The Ministry understands the link between ICT and economic growth and sees the measurement of the wider impact of ICT as a key challenge.

One of the points of departure of the IaD project was to fill in the blank spots within the yearly statistical publication The Digital Economy as published by Statistics Netherlands. Some of these blank spots have now been filled with new indicators or (suggestions for) beta-statistics. Relevance and timeliness are important features of these beta-statistics (pro). However (con) we have had to deal with poor statistical quality. The relevance of IaD methods for policy makers thus depends on a trade off between the new insight and early warning function (pros) on the one hand and the practical and statistical problems we have encountered on the other hand (cons).

The most important strategic implication from a policymakers' perspective is that IaD can signal new developments. It can serve as an early warning system for developments that could have a profound impact on the economy. For instance, if we had been conducting network-centric measurements on a large scale at several Dutch ISPs over say the last three years, we could have predicted the enormous success of video sharing sites or specific P2P protocols in sharing music online and the rise of the popularity of online gaming (massive multiplayer online role playing games and casual games). Given the dynamic of the digital economy, policy makers can no longer solely rely on established statistical agencies, as the speed, timeliness and flexibility of these agencies falls short. Instead of fully relying on the studies of market research agencies to describe these new phenomena (often also with poor statistical quality), IaD can provide an alternative source of information.

Another expectation on the part of the organisation that has commissioned this research project was that some existing indicators and statistics could be substituted by cheaper and less invasive methods. The government is, of course, very keen on more efficient methods of data collection. Exploring ways to reduce administrative burden is high on the political agenda in the Netherlands. In this context it would mean that Statistics Netherlands or other government agencies substitutes part of its active data collection (e.g. via traditional surveys) with passive data collection (via one or more of the

IaD methods). However, one important lesson we have learned from this research project is that the potential for the substitution of existing indicators and statistics should not be overestimated. In most of the case studies, statistical demands such as internal and external validity, scalability and longitudinal use were not met. Also, accessibility of digital sources has proven to be a practical, rather than a technical challenge in many of the cases. Privacy has proven to be a major issue. Closely related are proprietary matters. Some highly relevant data sets are within closed networks (extranets owned by large companies), zealously guarded by commercial third parties (market research bureaus such as Nielsen or Hitwise), or somewhere in-between (semi-commercial registers such as Cadastre).

In general, the potential usability of IaD, as a means for developing new (beta) statistics or substituting existing statistics is at its highest when:

- Value chains or parts of the value chain are highly digitalized;
- Products are digital and/or information on the product is highly digitalized;
- Markets are dominated by a few players;
- Market players are highly transparent (e.g. when web statistics are publicly available);
- When markets are highly regulated and administrative burdens are high;[33]
- Government registers are highly digitalized and of good quality (in the Netherlands the so-called basic registers are an interesting resource from an IaD perspective);
- Online activities are the subject of research;
- The subject of research is highly dynamic (and measurements on an annual basis are not sufficient) and/or real time information on the market is required;
- Various methods (user-centric, site-centric, network-centric) can be combined;

On the downside, the potential usability of IaD is limited when:

- Privacy issues and/or legal restraints regarding specific methods are concerned;
- Data on sector level are needed. Most IaD methods are aimed at product categories or markets and less so at sets of firms or organizations[34];
- Statistical quality demands are high (validity, transparency);
- National markets and phenomena are concerned as IaD often has difficulties in dealing with country specific measurements[35];
- International comparability is required as standardisation of definitions and measurement methods is generally lacking in the use of IaD methods. [36]

Recent discussions in the Netherlands and at a European level on matters such as privacy (is an IP number to be considered personal informa-

[33] This results in a somewhat strange trade off. Some forms of regulation result in, for example, registrations which are mostly administrative burdens. At the same time these registrations may be used to for statistical purposes. Attempts to reduce administrative burden may then result in less scope for using Internet as a data source!

[34] Additionally most activities associated with the EDE are more fluid, involve players over various industries and it is therefore difficult to link these one to one on existing sectoral typologies. Further, few markets can be measured using a standardized IaD approach, reducing the opportunities for intersectoral comparability.

[35] This is however not due to the measurement method but inherent to the phenomena that is being measured. Thus describing such phenomena at a national level might not be possible in the first place.

[36] This is however not a structural matter but rather a temporary issue which we need to overcome collectively and to which Statistical Bureaus can contribute considerably.

tion or not)[37], net neutrality[38] and safety on the Internet[39] are of course not much of a stimulus for IaD methods. They may be seen as too intrusive and not offering enough respect for personal integrity. In other words: not everything that is technically possible is also desirable from a political or societal point of view. Eventually, the usability of IaD very much depends on the legal and political choices to be made on these matters, which clearly is a delicate balancing act.

5.3 Future agenda: further research and experimentation

It is clear that using IaD is still in its infancy and that there is a need for further experimentation and research. We found that there is no "one size fits all" approach that can be used in any market or in every industry. In fact, each product, service, or specific economic activity in the value chain and each market has its own set of concentration points of digital footprints and therefore provides very specific opportunities for using these digital data sources for developing new, extra or substitute statistical indicators. These opportunities should be explored further on a case-by-case basis. The R&D project that we have conducted could be a starting point for new experiments.

We think there are several ways in which policymakers can contribute to this further exploration of opportunities and stimulate the further assessment of the usability of Internet as a data source. To elaborate on this further, we have formulated a number of policy recommendations:

1. A new publication could be initiated in which new phenomena in relation to the emerging digital economy are covered by using the so called "beta statistics";

2. A network of researchers, market research agencies, policy makers and statisticians could be set up as a means to contribute to this publication and to share their knowledge and experiences;

3. As an addition to this network, a clearinghouse for Internet statistics could be established. This clearinghouse should enable policy makers to ask questions to the specialists in the network on possible data sources for their specific subject, provide a quality check on statistics and indicators as developed by commercial third parties (market research companies) and assist in mining existing data. This clearinghouse can also address the problem of definitions and standards with regard to Internet statistics;

4. Statistics Netherlands (CBS) is possibly best equipped to develop into a key player in this research and this network, but they will have to be stimulated or commissioned to do so. Statistics Netherlands has obvious advantages in comparison to market-led parties as it:
   - has the scale and expertise for developing and collecting statistical indicators (sunk costs);

[37] Illustration here is the recent debate between Google and European regulators (see for instance http://www.washingtonpost.com/wp-dyn/content/article/2008/01/21/AR2008012101340.html)

[38] The pending Internet Freedom Preservation Act (H.R. 5353) is specially meant to anchor the net neutrality principle in the US constitution. This would block applications that discriminate between various Internet protocols, kill the market for traffic optimization and thus effectively the deployment of network-centric measurement in the US.

[39] In Germany, the Federal Constitutional Court has recently declared that 'cyber spying' violates individual's right to privacy and could only be used in exceptional cases (see for instance, http://news.bbc.co.uk/2/hi/europe/7266543.stm).

- can link data and indicators derived from IaD measurements and validate these using regular statistics[40];
- can guarantee privacy if needed;
- may use its judicial status (i.e. SN-law) to enforce co-operation of data providers;
- has the international network for international benchmarking, exchange of expertise and setting standards and developing international guidelines;

5. To start exploratory talks with organisations and companies that can contribute to this R&D network: payment service providers (representing a concentration point in almost every market), ISPs, Google, Microsoft, leading SNS, etc. These owners of promising data sources will have to be convinced that they have a role to play in producing more current statistics. This might involve new coalitions and some persuasion, since most information providers are stakeholders and are probably not very willing to cooperate;

6. In markets where ICT services are already highly concentrated or dominated by just one or a few companies (e.g. Agrovision in the pig breeder market), these companies can be approached in a similar way for their cooperation or they can be paid for the use of their data sources (on an aggregated level so there is no conflict with privacy regulations);

7. The government can anticipate the use of digital sources for statistical purposes when developing or implementing their own registers and ICT projects. The Personal Internet Page (PIP), the digital client dossier in the realm of work and income (DKD) and the electronic patient dossier (EPD) are all examples of major ICT projects that can be used for generating statistics;

8. Since many of the new developments we have studied (social networking sites, Internet TV, P2P music sharing, online gaming) are by no means limited by national boundaries, an international perspective is almost inevitable. Within OECD, Eurostat and other relevant transnational agencies, these new IaD approaches and methods should be discussed, experimented with and definitions and standards agreed upon. If possible, new lines of research should be opened up, best practices shared and new indicators (that are based on the use of IaD methods) should be used in regular statistical publications and policy documents.

In addition to these policy recommendations, we have formulated some suggestions for experiments based on the eight case studies we have performed (Figure 5.1.). When starting a follow-up R&D project on this subject, a specific budget for experimenting and developing beta statistics should be allocated.

[40] This can be done by comparing the development of time series generated by both types of methods. The assumption is that matching time series should be directly correlated but that the data generated by IaD methods generally precedes the conventional time series. Thus a structural change in the first type of time series should be followed by a similar change in the second type of time series. Note that in the case of diverging trends, no conclusion can be drawn with regard to the (lack of) validity of either of the two time series – either the data based on the IaD methods is flawed (e.g., measures a phenomenon which is not really there – type I error) or the traditional methods no longer reliably cover developments in the emerging digital economy (e.g., not measuring a phenomenon which is actually occurring – type II error). The fact that no clear conclusions can be drawn with regard to the overall validity is that the correlations are made without an underlying causal model.

Figure 5.1: Some suggestions for additional experiments

Case	Some suggestions for additional research & experiments		
	User-centric	*Network-centric*	*Site centric*
Pigs	Comparing the quality of Agrovision's statistics with regular statistics (LEI)	Not relevant	Investigate why sales prices at the online pig auction Teleporc are significantly higher than regular prices
Houses		Not relevant	Real-time measurement of (ratio between) asking price (Funda) and transaction price (register NVM) Map rise of C2C (or B2C disguised as C2C) market for housing
Webstores/ online market places	Set up panel to monitor online activity, more specifically buying and selling activities on leading C2C marketplaces (Marktplaats)	Not relevant	Collect real time data on online transactions (in collaboration with Thuiswinkel.org) Map financial flows of online transactions (via Equens/ Interpay)
Product software		Measure development/market share of SaaS-applications (e.g., Citrix)	Assess usability of self-reporting sites (e.g., softwaregids.nl) to describe Dutch software producing market
Recorded music	Measure use of P2P applications for sharing music files	Measure share of (illegal) music downloads in P2P-traffic (compared to sales via traditional and online retail). Measure rise (and fall) of Digital Rights Management (DRM) on the Internet	Assess quality of data of existing data from commercial third parties (esp. BigChampagne, also used by OECD) Measure occurrence (or not..) of "long tail" effect at music sites (e.g., iTunes, bol.com, Freerecordshop) Map market for "free music" (MySpace)
Internet TV	Monitor use of video and IP-TV applications (and compare with data from traditional panels, e.g., Stichting Kijkonderzoek)	Measure global market share of YouTube in Internet traffic (assess validity earlier measurements, e.g., Ellacoya) Measure extent of (illegal) video downloads on most popular P2P-networks (follow-up Ipoque measurements)	Follow development of Uitzendinggemist.nl traffic
Online Gaming	Assess re-use of user-centric data gathered by the game distribution platform Valve	Measure the development of Second Life versus RuneScape (a popular MMORPG) – both applications use fixed ports (thus are relatively easy to detect) and are comparable in size	Deploy spiders to map the market for virtual objects a.k.a. Real Money Trade or RMT) via professional sites which sell such items ("gold farms" such as IGE and THsale)
Social networking	Set up panel to monitor online activity, more specific time spent and applications used on leading SNS, amount and nature of content that is contributed, etc.	Measure usage patterns in popular SNS sites (e.g., Hyves) and chat applications (e.g., MSN Messenger)	Spider user profiles on SNS sites (development over time, demographics, etc.)

6 Conclusions

Background

The starting point of this project was the notion that organisations and individuals increasingly leave behind a so-called digital footprint. This builds up during various economic and social activities mediated in whole or in part over the Internet. The challenge for this project was to assess the feasibility of mining these footprints to describe socio-economic phenomena. The project also looked at ways of developing new data and indicators for the emerging digital economy (which might replace existing indicators).

The research questions were twofold: what methods can be used and what can be measured with them? We identified several methods of using the Internet as a data source spread over three inherently different types of measurements i.e. user-centric, network-centric and site-centric. In the course of eight case studies, we looked for data and indicators (new, extra and substitutes, if relevant) to characterize the markets concerned. We also assessed the usability of the IaD concept. This process enabled us to identify many advantages and disadvantages of the Internet as a data source.

In this concluding chapter, we review the main findings presented in previous chapters.

Relevance

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

The clearest added value is that IaD has been shown to provide insight into markets and phenomena in areas where the established statistical agencies have no information. On the basis of IaD, we gathered information that was previously unknown. IaD provides new or enhanced insight into relevant economic and social developments, in a quick and timely (almost real time) fashion. In some cases, it can act as a substitute for existing indicators and data collection methods, leading to reduced administration and lower costs. Even if the statistical quality of the data collected with IaD methods is poor, it is better to have measured relevant developments partly or badly, than not to measure them at all.

Relevant, in this case, means that these new developments can generate new economic activities (e.g. new services) with a considerable economic value. Policy makers currently largely ignore these new developments and new economic activities when using established statistical indicators. In one of the case studies we were able to make a calculation of the total amount of transactions mediated by a leading Dutch C2C online marketplace (marktplaats.nl). On the basis of their web statistics, we have calculated that the total value of transactions in 2006 is €4,7 billion, which represents a very substantial part of online consumer spending in the Netherlands.

Potential

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

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

The usability of IaD methods is also high when:

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

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

Forward looking: early warning function or real time monitoring

IaD-methods are helpful in mapping new trends, phenomena and markets that cannot be tracked using existing methods and statistics. IaD gave us the opportunity to signal and measure completely new phenomena associated with the emerging digital economy. They have not been previously measured or only very marginally in existing statistics (e.g. Social Networking, C2C marketplaces, Internet TV, etc.). We compare most traditional statistical work with looking through the rear-view mirror of car. On the other hand, during our research we gathered information of a more forward-looking nature: what are the new developments online and how are they influencing the economy?

We believe the function of Internet based measurement can be of use as an early warning mechanism. For example, if we had used network-centric measurements over the last three years, we could have predicted the enormous success of video sharing sites or specific P2P protocols in sharing music online. Timeliness is, therefore, a very important feature of IaD methods. Sometimes it is even possible to monitor certain markets in real time (pigs, housing, etc.).

Opening up the black box of the digital economy

There are several important lessons learned on the development/deployment of the digital economy when experimenting with assessing the feasibility of IaD methods. They include:

1. The notion of digital footprints is not limited to digitalized products and services, but applies to a wider set of markets and industries. The availability of digital information on products and markets has proven to be far greater than digitalization of the product itself. For that reason it is often the case that certain industries and markets associated with the "old economy" (houses, pigs) can be tracked more effectively using the Internet as a data source than products and markets associated with the "new economy" which are highly digitalized themselves (online gaming, online music).
2. Information on goods and services firstly represents an economic value in itself. This in turn gives rise to new, often innovative economic activities with new business models and hence competition patterns within markets (e.g. funda.nl, marktplaats.nl, product software distributed electronically). Many of these new economic activities are increasingly hard to trace using existing statistics and categorizations (online music is a fine example here, as it is hard to delineate to which industry it actually belongs). The "long tail" phenomenon is real.
3. Peer-production and the increased transparency of markets are the most dominant and typical economic trends associated with the emerging digital economy (see section 2.2). Consumers and users add value in the form of user generated content.
4. In several markets, user generated content has started to mix with traditional economic production (online music, Internet TV, housing) and, as a result, barriers between the social and economic realm are blurring (SNS, C2C marketplaces). In fact the emerging digital economy affects time consumption patterns and logistics. For instance, think of the micro logistics as a result of the massive number of transactions between individuals through Marktplaats.nl. This increases the bargaining power of consumers (more transparency, fast sharing of information) and citizens alike.
5. IaD can shed light on – and therefore confronts us more explicitly with - the darker side or grey zone of the emerging digital economy. This includes illegal downloading, privacy issues, tax evading commerce, cybercrime and so-called "shameful content" (e.g. pornography). The use of IaD can provide more insight into their magnitude.

Challenge for traditional statistical agencies

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

On-line stores are no longer limited to the retail industry. Music is sold and distributed through a wide number of channels by an increasing number of actors. C2C informal "markets" now coexist next to B2C markets. An important lesson here is that many of the respondents interviewed during case studies, do not recognise themselves in the existing statistics. They ask for new indicators and revised definitions of products and markets. This point also signals an important problem for established statistical agencies. If they do not succeed in capturing dynamic, relevant developments in the emerging digital economy, they risk being overshadowed by those that do. In other words, statistical agencies will have to come up with new methods such as IaD in order not to run the risk of disintermediation, something that has seriously affected the record companies in the music business.

Substitution and statistical usability

The potential of IaD-methods for substitution of existing indicators and statistics should not be overestimated. The idea of the IaD-project was that some existing indicators and statistics could be substituted by cheaper and less invasive methods. An important lesson we have learned during this research project is that this is only feasible in a limited number of cases.

The value added of IaD methods regarding substitution is highest where

- the indicators derived are readily available,
- where there is a need for more frequent statistics
- and where there is clearly a need for less invasive ways of gathering data.

In most of the case studies, statistical demands such as internal and external validity, scalability and longitudinal use were not met.

The statistical usability of IaD is the result of various trade-offs. All IaD methods share the distinctive trait that they are based on the non-reactive observance of spontaneous behaviour, whereas traditional methods (e.g. postal or telephone surveys) are based on the observance of reactive, provoked behaviour. This has important consequences for the statistical usability. In general, IaD methods perform better than traditional methods (panel surveys) in terms of efficiency, objectivity and reliability but worse when it comes to internal and external validity. The threats to internal validity are less severe when the statements only refer to data traffic – and this is precisely the part of the new economy that is not well covered by traditional methods.

When it comes to statistics on the actual use of the Internet, data collected by IaD methods are probably more relevant than data gathered by traditional surveys. They are definitely more current, which is a major issue in the highly dynamic emerging digital economy and technically easier to access. However, when the scope of the statements is broadened (and thus moves away from the hard data), the validity becomes questionable.

Beta-indicators

IaD methods may lead to beta-indicators for the emerging digital economy. These indicators are a new category of socio-economic indicators for measuring the emerging digital economy more closely. They produce results with a clear early warning quality, but without the same statistical rigour or quality obtained from established indicators as published by statistical bureaus.

In our view, there is a constant need for more experimental or "blue skies" indicators for assessing the prevalence, size and ultimate impact of the emerging digital economy. We acknowledge that, from this pool of beta-indicators, only a few will make it into mainstream established statistical indicators. Nevertheless, others remain interesting statistical indicators even though they have a lower statistical quality. They are too good to ignore, especially in areas where no other higher quality indicators are available. Some will simply fade away as better statistical indicators surpass them or it turns out they are measuring only a temporary hype (which is often difficult to assess in advance). Conversely, data generated by IaD methods might be used to calibrate (adjust) or even validate (confirm or reject) established traditional statistics.[41]

The way forward and role for Statistics Netherlands

The use of IaD methods clearly rides on the waves of the trend towards increased digitalization. The reach of the digital domain is still expanding and also covers an increasing part of the traditional economy. The number of nodes in the network (be it organisations, individuals or devices), and the frequency and variety of interaction between these nodes, will continue to grow. This greatly improves the possibilities to observe the behaviour of individuals or firms through the digital footprints they leave behind on the Internet. In this respect, IaD methods are highly efficient. They measure at the heart of the matter, namely changing patterns in data traffic. The fact that there is a true need for new types of statistics is witnessed by the rapid grow of various "Internet-based statistics" that are offered by commercial third parties. The quality of these commercial statistics is often rather dubious and unknown at best.

In our policy recommendations we have made a plea for a clearinghouse function for these Internet-based statistics. Statistical agencies like Statistics Netherlands could provide this function, but they will have to be stimulated or commissioned to do so.

They are best positioned and well equipped to develop into a key player in this type of research. Running such clearinghouses brings obvious advantages when compared to market-led parties. These include the:

1. scale and expertise for developing and collecting statistical indicators (sunk costs);
2. possibility to validate data and indicators derived from IaD measurements using regular statistics;
3. possibility to guarantee privacy if needed;
4. a judicial status they might want to use to enforce co-operation of data providers
5. the international network for international benchmarking, exchange of expertise and setting standards and developing international guidelines.

Further research and experiments

It is clear that using IaD is still in its infancy. We believe there is a strong need for further experimentation and research. Each product, service, specific economic activity in the value chain and each market has its own digital footprint. They have their own typical concentration points and, therefore, provide very specific opportunities for using IaD for indicator development or for the substitution of existing statistics.

The mix of methods that can be used will largely differ between markets and industries making comparison between markets and industries tough. However, as the digital footprints of products, markets and industries increase further over the next years, there is a clear need to further invest in experiments with IaD.

[41] See paragraph 5.3 (#4). Data generated by traditional measurement methods may no longer be valid due to the fact that the world is changing. Note that this is regardless of the reliability of those methods (which may still be high – they just correctly measure phenomena that are less and less relevant).

Role of government

Government agencies are first and foremost users of statistics, be it traditional or innovative statistics like IaD. The important question is not whether IaD methods should be used or not but rather how they should be used. Sometimes there are simply no alternatives to the use of IaD methods; existing statistics lose their relevance due to their lack of timeliness and/or because entire sections of the emerging digital economy are not covered at all.

Secondly, it can be said that IaD methods generally offer a cost-effective solution and can significantly reduce administrative tasks. Privacy, however, is an issue of growing importance in the use of IaD methods. Under certain circumstances, privacy concerns might even block the use of IaD methods altogether. Therefore, the decision to use IaD methods for either substituting or complementing existing data that is collected by traditional methods should not be taken light-heartedly. There is a trade-off between efficiency, objectivity, timeliness and cost-effectiveness on the one hand and validity and privacy on the other. Saying that a measurement is non-intrusive from a technical viewpoint may still mean that it can be highly intrusive from a privacy point of view. To boost the practical usability of IaD methods, the legal framework should be organised in such a way that the critical privacy and security issues can be resolved.

Thirdly, a very important role of government is to stimulate further research on the feasibility of IaD methods and to facilitate experiments. A network of researchers, policy makers and statisticians could set up a innovative research program and new kinds of publications on this subject could be initiated. Also, governments should anticipate the use of digital (re) sources for statistical purposes when developing or implementing their own registers and ICT projects.

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

_ References

Benkler, Y. (2006). The Wealth of Networks. How Social Production Transforms Markets and Freedom. New Haven: Yale University Press

Bethlehem, J. (2006). Representativiteit van web-surveys. Een illusie? KNAW Data Archiving and Networking Services symposium, Oktober 12, The Hague.

Biddulph, M. (2004) Crawling the Semantic Web http://www.idealliance.org/papers/dx_xmle04/papers/03-06-03/03-06-03.html

Brackstone G. (1999) "Managing Data Quality in a Statistical Agency". Survey Methodology, 25(2): 139-149.

Brackstone G. (2001) "Managing Data Quality: The accuracy dimension". Paper for International Conference on Quality in Official Statistics, Stockholm.

Brennenraedts, R. (2008). Internet as Datasource. Music: Legal, Illegal, Digital and Analogue. Utrecht: Dialogic

Brickley, D. (2003). RDF Hyper-linking http://www.w3.org/2001/sw/Europe/talks/xml2003/Overview-6.html

Brynjolfsson, E & A. McAfee (2007). Beyond Enterprise 2.0. MIT Sloan Management Review. Spring 2007.

CBS (2007). De Digitale Economie ("The Digital Economy"). Heerlen: CBS.

Choi, S-Y, D.O. Stahl and **A.B. Whinston** (1997). The Economics of Electronic Commerce. Indianapolis: Macmillan Technical Publishing

Cronbach, L. J. (1949). Essentials of psychological testing. New York: Harper & Row

Cronbach, L.J. and P.E. Meehl (1955). Construct validity in psychological tests. Psychological Bulletin 52: 281-302.

Dodds, L. (2006). Slug: A Semantic Web Crawler. http://www.ldodds.com/projects/slug/

Eurostat (2000a), "Standard quality report", Doc Eurostat/A4/Quality/00/General/Standard report.

Eichman, D. (1994). The RBSE spider: balancing effective search against Web load. Proceedings of the First World Wide Web Conference. Geneva, Switzerland.

Ellacoya (2007). Web Traffic Overtakes Peer-to-Peer (P2P as Largest Percentage of Bandwidth on the Network. http://www.ellacoya.com/news/pdf/2007/NXTcommEllacoyaMediaAlert.pdf

Ernst & Young (2007). Mediabarometer. Eyeballs & Communities

Eurostat (2000b), "Definition of quality in statistics", Doc Eurostat/A4/Quality/00/General/Definition.

Hertog,P. den, C. Holland and H. Bouwman (1999), Measuring E-commerce. Recommendations for a Dutch E-commerce monitor, published by the Netherlands Ministry of Economic Affairs, The Hague.
Ipoque (2007). Internet Study 2007. Leipzig: Ipoque.

Kaart, M, J. Vrancken & W. Vree (2007), Internet topology measurements for identifying and understanding policy issues, Info, vol. 9, no. 6, pp. 70-81

Manola, F. and E. Miller (2004). RDF Primer. W3C Recommendation. 2004.

McGuinness, D.L. and F. van Harmelen (2004). Web Ontology Language (OWL) Overview. W3C Recommendation.

OECD (2007). Participative web and User created content. Paris: OECD.

Perez, Carlota (2004), The new techno-economic paradigm and the importance of ICT policy for the competitiveness of the whole economy, presentation at the high level conference "Looking into the future of ICT", September 2004, Amsterdam.

Pouwelse, J.A. , P. Garbacki, D.H.J. Epema, H.J. Sips (2007), Pirates and Samaritans: a Decade of Internet-based Measurements on Commons-based Peer Production, Elsevier preprint.

Schmitt, N. & F. L. Oswald: " The impact of corrections for faking on the validity of noncognitive measures in selection settings". Journal of Applied Psychology, mei 2006

Segers, J.H.G. (1999), Methoden voor de maatschappijwetenschappen

Stichting Internet Reclame (2007). Establishment Survey. Baarn: STIR.

Sugden, R., R.A. te Velde, J.R. Wilson (2002), "Economic Development and Communication: Problems for Governance under Globalisation", L'institute Discussion paper 19, Universities of Birmingham, Ferrara and Wisconcin-Milwaukee.

Swanborn, P.G. (1987), Methoden van sociaalwetenschappelijk onderzoek, Meppel: Boom

Varian, C and H.R. Shapiro (1999), Information Rules. A strategic guide to the network economy. Boston: Harvard Business University Press

Velde, R.A. te (2008). Internet as Datasource. Online Gaming. Utrecht: Dialogic.

Zee, F. van der (2004). Kennisverwerving in de Empirische Wetenschappen, de methodologie van wetenschappelijk onderzoek. Groningen: BMOOO

Mr. A. (Arie) van Bellen
Director ECP.NL

Drs E. (Erwin) Bleumink
Managing director Surfnet

Dr. W.A.G.A. (Harry) Bouwman
Associate Professor Information and Communication Technology at Delft University of Technology

Drs. Th.B. (Theo) Fielmich
Senior policy advisor, Dutch Ministry of Economic Affairs.

Dr. K. R. E. (Eelco) Huizingh
Associate Professor Business Development at Groningen University

Prof Dr. Ir. W J. (Wouter) Keller (Chairman)
CEO Argitek & Free University Amsterdam

Dr. Ir J.A. (Johan) Pouwelse
Researcher on Peer-to-Peer technology at Delft University of Technology

Drs. G.H. (Eric) Wassink
Statistics Netherlands

Drs. A.C. (Marcel) van Wijk (secretary steering committee)
Statistics Netherlands

Steering Committee Members

Colophon

Published by
the Ministry of Economic Affairs.

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

The Hague, April 2008

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

Information

Directorate-General Energy and
Telecommunications

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

Publication number: 08ET12