CES/SEM.52/15*
30 December 2003

ENGLISH ONLY

STATISTICAL COMMISSION and UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT (UNCTAD)

INTERNATIONAL TELECOMMUNICATION UNION (ITU)

UNESCO INSTITUTE FOR STATISTICS (UIS)

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD)

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (Eurostat)

Joint UNECE/UNCTAD/UNESCO/ITU/OECD/Eurostat Statistical Workshop:
Monitoring the Information Society: Data, Measurement and Methods
(Geneva, 8-9 December 2003)

Event related to the World Summit on the Information Society

PEOPLE IN THE INFORMATION SOCIETY: INDIVIDUAL AND HOUSEHOLD USE AND PENETRATION OF ICTs METHODS*

Keynote paper

Mr. Huub Meijers, Researcher
MERIT & International Institute of Infonomics, University Maastricht, The Netherlands

[*] Due to the late submission, this paper could neither be translated nor reproduced and has been posted on Internet as submitted by the author.

Using ICT to measure ICT use: facts and fictions

Prepared for the Statistical Workshop on:

Monitoring the Information Society: Data, Measurement and Methods

Geneva, 8-9 December 2003

Session VI: People in the Information Society:

Individual and household use and penetration of ICTs: Methods

By

Huub Meijers

Draft version

Researcher at
MERIT & International Institute of Infonomics
University Maastricht
PO Box 616
6200 MD Maastricht
The Netherlands
t: +31 43 3883882 f: +31 43 3884905
e: huub.meijers@merit.unimaas.nl i: www.merit.unimaas.nl & www.infonomics.nl

Introduction

The development and diffusion of new Information and Communication Technologies (ICTs), which has accelerated in the last decade, and the rapid adoption of Internet technologies and its related applications demands for new and accurate statistics on the Information Society. The fast adoption of these new technologies is, however, not a global phenomenon: many large regions seem to be excluded from this development but also in leading countries there are many who are not included in this process, either forced by circumstances or voluntary. From policy perspective there is a high demand for clear insights into these developments and it is the demand to statistical offices to provide these insights, or at least to provide data that lead to a better understanding of the processes at hand. ICTs also allow for other ways to measure these developments because of their own automation and communication possibilities. In this sense data collection and surveying can be regarded as a structured form of communication with individuals, households, firms and other organizations and the question is whether ICTs can facilitate this process. The main question addressed in this paper is about new methods to collect data concerning the Information Society and it has a special —but not exclusive— focus on measuring ICT use by individuals and households. To elaborate on this question, first a brief overview of relevant policy questions will be portrayed since addressing policy questions and related research questions is the main point of departure. In order to do so, a short overview of recent developments of the Information Society is needed since that guides the policy demands in many cases. In this overview, the distinction between aggregate data (here called macro data) and micro data proves to be important. Second a short overview of current non-survey based methods is provided to give an overview on what already is established, including the shortcomings of these approaches. Finally some promising new methods are presented including an overview of their gains and shortcomings. It is concluded that many policy questions, like the background of the digital divide, both at an international and at a national level, still demands classical surveys. It is demonstrated that using ICT to measure ICT is a fact but using ICT to measure non-ICT use is still an unresolved paradox, a fiction.

What do we want to measure: policy relevance

The main goal of statistical offices is to provide and distribute objective statistical information and analysis, which is needed to support political processes and to improve the decision-making process and research work. The main focus is on politicians, government, and administrative agencies, but statistical offices also serve business and industry as well as citizens and non-governmental organisations. This paper concentrates on statistics that are needed to support the political process, including the provision of statistical data for research that is aimed at the facilitation and provision of (scientific) foundations for this political process. So the production of policy-oriented statistics is the scope of this paper. This implies that first we have to determine what the policy questions are, second how these policy questions can be answered and finally what statistics are needed. After a brief description of these steps, this paper gives some possible extensions to measure the Information Society.

The Information Society is however above all an evolving society and the related policy issues move along with this development. Looking from this perspective it is more than illustrative to briefly describe the main development of the Information Society as it has developed in the past as this gives insights in changing policy goals and thus changing demands for statistics. The focus will be on the most advanced nations and a broader perspective follows afterwards.

A technical view could be a logical one since the Information Technology revolution finds its origins in the speed at which technology develops. Although a technology-driven perspective touches upon statistical problems (see

Box 1 for an example), a more socio-economic or socio-technical perspective is more fruitful in this setting since that will reveal the different phases in society which are more relevant for policymakers, given the focus on individuals and households taken in this paper.

Box 1 *Heterogeneous IT goods and statistics.*

> Goods and services clearly develop and change in quality over time and it is important to include these differences in our statistics, certainly when describing the Information Society. Joan Robinson's remark in the grand capital debate: "A spade is a spade is a spade" (free after Robinson) does not automatically translate into "a computer is a computer is a computer". Owing a computer does not say much about the capacity of it and of its possible impact on human behaviour and activities. A modern 8086-based stand-alone microcomputer with 640K memory and two 5¼ -inch floppy disks from the early 1980s is not comparable to a manifold Gigahertz computer equipped with huge amounts of memory, huge hard drives, and amazing graphics and sound possibilities that is hooked up to the Internet. The latter is in terms of capacity even outperforming the fastest mainframe computer in the old days. Moreover, it is not only an evolving path over time, it also leads to heterogeneous IT equipment (and related products) at one point in time. This clearly poses the problem of incorporating quality characteristics into statistics as aggregation problems arise and as measured quantities are less (internationally) comparable. This has been recognized by many scholars and is included in (IT-specific) surveys, asking for various characteristics of information technology goods. In making (International) comparisons, these characteristics are often disregarded, however.

Following Ian Miles (2000), the Information Society can be described along four different phases, which coincide roughly with the last four decades. In the early days, computer, telecommunication and media were evolving seemingly independent and no sign of convergence could be observed. In this first phase, called the "Islands", computers were not connected and these machines were mainly, if not all, mainframes and minicomputers. The computer facilities were small (although their size was big), few and isolated and to use them a fair amount of technical know-how was required. In the end of the 1970s, governments recognized the importance of the IT industry and supported it by special programmes, i.e. military or national plans for IT. Individual and household use was absent until the end of the 1970s when (some) hobbyists started to build self-assembly based computers. The telephone infrastructure was developing; the exchange was based on electromechanical principles and the user interface far from mobile. Although fax-machines and its underlying technology were available for a long time, the adoption did not begin until the early 1980s, the second and so-called "Archipelago-phase". During this phase, the (personal) computer appeared in many different sizes and it had a revolutionary impact on the organisation of its use. Although the use was initially restricted to some particular (professional) groups, more and more people had access to the computer and the development of software like the early spreadsheet and word processing programmes fuelled its diffusion. Also the first, two-way based communications systems appeared which made the very first steps towards mass communication. Developments in microelectronics had a huge impact on different devices: telephone connections started to become digital, the bulbs in the TV-sets were replaced by transistors and many different new devices emerged (microwave, answering machine, video-recorder etc. etc). Email became increasingly popular but the hassle of finding the appropriate addresses blocked its real break-through. Although technically less sophisticated, the facsimile diffused at an enormous speed. Moreover, deregulation of the telecom market started and national R&D programmes for computer and telecommunication gained importance. Statistics were mainly focussed on business use and on measuring (macro-) economic effects of ICT investments. The productivity paradox debate gained momentum and statistical flaws (due to rapid changes in quality of both inputs and outputs) was one of the possible explanations for this paradox.

The third phase is characterized by convergence and interconnection, i.e. the archipelago's became connected and turned into the "continent". Not only desktop computers (and mainframes) but a wide variety of other devices (laptop's, handhelds and PDA's, mobile phones etc.) crowd this continent. Mainly two developments characterized the changing Information society in the 1990's: the success of the Internet and of mobile communication. It is also in this period that the terms "Information Society" and "Information Society Technologies" are adopted to indicate the society-wide implications of these technologies. All kinds of "killer applications" like email, the Web, chatting, sms etc., speeded the diffusion of computing and communication devices, firmly being pushed by strong deregulations of telecom markets resulting in rapid falling prices. The last phase, starting around the turn of the century is characterized as "the Ecosystem" and indicates the further convergence and consolidation of the "continent" were wireless communication through Bluetooth, WiFi (or Wlan), and similar technologies connects the growing number of devices that surrounds us. Expanding adoption of new organisational and behavioural patterns to benefit from the technology-oriented developments and integration of these devices and applications in more and more activities will finalize the transformation form an industrial era towards the Information era.

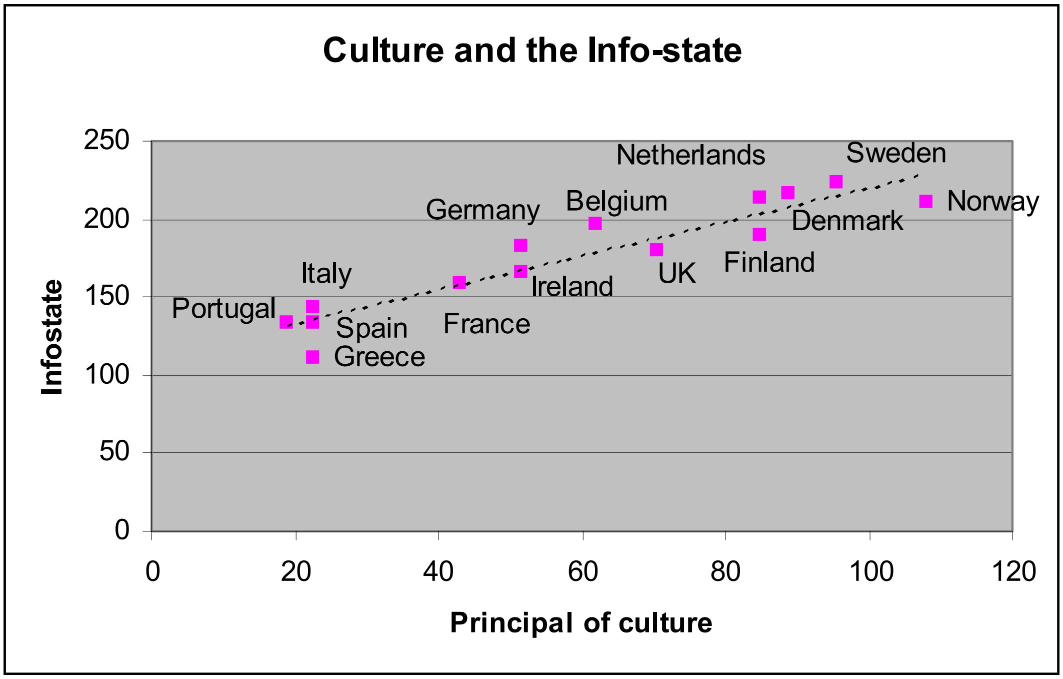
This very brief —advanced Information Society based— overview of main developments in the Information Society —which obviously oversimplifies and is too short to portray all developments— gives a broad background to understand various policy demands and the development of these demands over time. In the early years the policy focus was on ICT investments by firms and, more important, the development of ICT producing industry. Demand for specific statistics was fairly low. In the second phase policy attention was mainly focused on R&D and on special ICT related initiatives implying a demand for statistics on various types of R&D expenditures and ICT investment as well as a demand for evidence of the effects of such policies. The scope was limited mainly to the economic sphere and international comparison gained in importance, albeit at a low pace. The third phase broadened the scope of ICT applications and individual and household use became a policy issue, mainly because of the potential skill enhancing effects of private ICT use and of the importance of lifelong-learning programmes. Also the unequal utilization of ICT within and between countries posed policy questions, which had associated implications for statistics. Not only economic but also social and organisational effects were taken into account. Finally the last phase shows an expanding Information Society with many devices and applications that surround both businesses and individuals and policy questions focus more and more on the use of ICT in specific sectors, on specific applications, on the effects of ICT as well as on the effects of ICT policy. Demand for statistics becomes more and more detailed and there is a clear need to investigate causal relations both among the users —why do they use, what and how do they use and what are the consequences— and non-users —why don't they use, what are the consequences— and within all sectors in our societies. The explanations become more multidisciplinary and cultural aspects become more important next to socio-economic aspects. This concerns not only the impacts of ICT use, it certainly is important in the explanation of users versus non-users as for instance is experienced by two Italian students researching Internet and ICT use:

"Idiosyncrasy is a possible reason for low Internet use and high mobile phone use. Italy is a country in which personal relationships form the cornerstone of life and its daily transactions. The Internet is a global communication network for both individuals and business, where the personal loses its priority...the anonymity of the Internet would render them almost powerless and be counter to their culture" (Leo and Gabriele, 2000).

An important concern in researching the diffusion and impact of ICT is the type of (policy) question one wants to answer. Comparing ICT use between countries or regions clearly can be done at an aggregate level. Such macro studies reveal insights in a country's relative position and also can shed light on relations between former actions or policies and the country's performance. Examples are for instance studies on the relation between productivity and the investment in ICT, as is shown by many scholars such as Collechia and Schreyer (2002), Daveri (2002), Jorgenson (2000), to mention just a few. Although macro studies and the statistics involved are often based on surveys, the typical characteristic of these statistics is that they do not need to stem from omnibus surveys and only rely on aggregate figures.

Other good examples of such macro statistics are Sciadas (2003) and Sibis (2003), both defining a Digital Dive Index, which is based on a basket of relevant indicators. Sciadas (2003) measures the Digital Divide Index along the degree of a country's Info-state. The Info-state is an index based on Info-density and Info-use, which are both defined as indices of several underlying components. Info-density is a measurement related to ICT stocks whereas Info-use is more related to the consumption flow of ICT. The work of Sciadas is focussed on a worldwide scale of analysis, which clearly restricts the number of variables that can be included. Sibis (2003) shows a similar analysis but their framework is focussed on the EU and includes many more variables, including those related to the eEurope initiative, so including indicators for e-Government, e-Health, e-Commerce, e-business etc. Note that both indicators are based on a mix of data, some obtained from non-survey measurement, some via surveys.

Figure 1. Culture and the Info-state in Europe



Above we already mentioned the increasing importance of social and cultural dimensions that are included in the research on the adoption and impact of Information Society Technologies. As an (simple) example for a macro analysis on this issue Figure 1 shows the relation between Cultural aspects —measured by leisure activities— and the use and adaptation of ICT in 14 European Countries. In this graph a principal component measure of 13 different aspects of leisure activities is plotted against the Info-state as measured by Sciadas (2003).[1] The graph suggests that there is indeed a relation between culture and the adoption of Information Society Technologies. However, a more thorough analysis is needed because of the possible

[1] The principal component is based on 13 aspect of leisure: (based on a one-dimensional
Princals solution): the importance attached to leisure, satisfaction with leisure activities, expenditure on leisure, alcohol consumption, attendance at performances, trips abroad, participation in football, book reading, television viewing, cinema attendance, sales of cds, museum attendance and preference for domestic
music. The data are taken from SCP (2000).

endogenous relations between income, education etc. and leisure activities and the relative position of the Info-state.

To conclude, Macro indicators are important and can reveal several relations such as the above-mentioned relation between ICT investments and productivity or the hypothesised relation between Culture and the adoption and use of Information Society Technologies. They are also highly useful in describing a country's or regions position vis-à-vis other countries or regions, as for instance is done in many Digital Divide-related documents. However, since aggregate data do not allow for microanalysis by definition, and since statistical objects are (in general) not homogeneous, this also puts restrictions on the range of possible exploitation of aggregate data.

Aggregation of non-homogeneous data always describes some average of the sample and disregards within-sample variation. If carried out for more than one variable, also the relation between variables can disappear (or even wrongly appear) due to aggregation. An example of such aggregation problem is examined in the productivity paradox discussion in which no relation between aggregate productivity and aggregate investments in ICT were found whereas such relation was found by micro-studies. A clear positive aspect of aggregate data is that they can come from unrelated sources, i.e. there is no need for comprehensive (and expensive) omnibus surveys to produce aggregate statistics. Constructing comparable aggregate data on a wider, more global, scale can be achieved more easily, as is shown by Sciadas (2003) whereas the underlying surveys may differ between countries. This also makes the construction of aggregate time-series much more straightforward. Moreover, as we will see below, such statistics can also be based on non-survey data collection.

It is obvious that micro data can reveal much more detail than is ever possible by using aggregate data. Relations between ICT use and variables like income, employment status, and level of education etc. can only be established using micro data. In principle such analysis would also be possible using macro data in a cross-country setting —there is high correlation between income per capita and individual ICT use— but since many other economic, social and cultural variables also have an important role in this relation, macro data often don't allow for such more precise analysis.

An important question is whether ICT related surveys should be included in omnibus surveys or whether they can be used stand-alone. From a historical perspective it is evident that it is fairly impossible to predict future policy and research questions and unexpected relations can become increasingly important. This clearly favours the inclusion of Information Society Technology questions in omnibus type of surveys.[2] The negative aspects are of more practical nature, albeit not less important. Cost of surveys and a limited length of surveys (and decreasing response rates) are the main reasons to choose for stand-alone questionnaires. Also the demand for international comparable statistics plays an important role since high quality omnibus surveys are simply not possible in every region of the world. Socio-economic and cultural differences between world regions also imply that relevant variables and thus survey questions have to differ to depict the regions' specific characteristics. It is obvious that whenever possible, whether included in omnibus or stand-alone surveys, a set of comparable questions should be included in both to allow for international comparison and benchmarking.

By way of example, examine the digital divide in an advanced country, the USA. In their telephone survey based research among US citizen Lenhart, et al. (2003) found that 42% of the Americans are consider themselves as not using the Internet. Of

[2] The US Bureau of the Census in cooperation with BLS conducts monthly a Basic Current Population Survey, focussing mainly on labour market issues, employment etc., but includes at regular intervals a supplement on computer and Internet use.

these non-users, 20% are “Net Evaders” (Those who live with someone who uses the net), 17% are “Net Dropouts” (having used the Internet but dropped out due to several reasons (most of technical nature), and 24% are the “Truly Disconnected” (no experience with the Internet at all).[3] This example demands for further investigation and it is very likely that a basket of socio economic and cultural aspects can explain (at least a part) of the digital divide. This also demands for integration in other, broader, surveys like the CPS.

"Bridging the digital divide requires more than simply offering computers and Internet access. Technological fixes won't close the divide unless they take into account the social reasons why people aren't online," Patrick Moorhead, GCAB chairman, said in a statement." For companies that are increasingly focused on global emerging markets, understanding socioeconomic factors impacting technology adoption in the various regions is absolutely crucial." (Chen and Wellman, 2003).

Initiatives such as the OECD’s "Working Party on Indicators for the Information Society" try to establish a basic questionnaire that could be used by individual countries to overcome the shortcomings of comparable statistical data in the Information Society. The European Commission and EUROSTAT also indicate that: “in order to understand the socio-economic impacts of IST, indicators on availability, penetration, activities and use, particularly by households, are needed. It is argued that special user surveys should be undertaken” (cf. EPROS 2000, p 11). Since future questions are not know yet, an integration in or at least the provision of possible connections with omnibus surveys should be considered, although a full integration may not be possible on practical grounds.

[3] For other example see for instance Pastore (1999) and Wyatt et al. (2002).

Some examples of IST data and methodologies

Concerning the use of the Internet, many statistical agencies report on the number of individuals and households on line in their statistics and they usually measure Internet access on the basis of surveys of households and individuals. Surveying ISPs is another way statistical offices use to collect information on Internet use by and these surveys provide a wide range of information, for example on type of subscriber (business, household, government), type of technology used (dial-up, ADSL, cable, WAP, etc.), and sometimes even the length of connection and volume of data downloaded.

Although additionally information on Internet subscribers by country can be obtained from reports of the largest telecommunication carriers the increasing use of other technologies such as cable, satellite and Wlan reduces the coverage of such statistics and more carriers (so not only telecommunications) should be included.

One of the first established data on Internet use is the host-count. Initially this was measured by the count of unique IP numbers corrected for those hosts that did not react on a ping-request, i.e. those hosts that are not ‘alive’. This measure is improved several times and also some variations are made such as the web-host count, i.e. those Internet hosts for which the IP number resolves into a domain name and which serve World Wide Web content. A brief overview of such host counts as is used by the OECD is given in Box 2. Also the allocation of hosts to countries has been improved since the Generic Top Level Domains, gTLD such as .com, .edu, .net can be allocated to the country’s IP address blocks but also to the residence of the owner of a registered domain. Of course this is restricted to those hosts for which the address resolve into a domain name. Institutions listed in Box 2 are continuously improving this measure.

Closely related to the host count is a measure of the quality of connections to the hosts. An example of such approach is the PingER project, see Cottrell and Matthews (2003). This project measures Internet reliability and quality (speed and average packet loss) by using a fairly simple and lightweight tool and can give a timely and global overview of the quality of the Internet infrastructure at low costs.

Box 2. Measuring Internet hosts and servers

The number of Internet hosts is one of the most commonly used indicators of Internet growth. It includes any computer system connected to the Internet (via full-time or part-time, direct or dial-up connections), although some systems may not be accessible owing to technologies such as firewalls. Hosts can thus be thought of as an indicator of the minimum size of the public Internet.

Surveys of Internet hosts are undertaken by several entities. Every six months, Network Wizards, on behalf of the Internet Software Consortium (ISC), carries out the longest running host survey. RIPE conducts monthly surveys of Internet hosts for countries in their region. A third source of statistics is NetSizer's Internet Sizer from Telcordia Technologies which provides daily updates of the number of Internet hosts based on a random sample of IP addresses throughout the day. Telcordia provides hosts by country as well as by top-level and second-level domains. Hosts by country are computed by redistributing the hosts with three letter domains (e.g. .com, .net, etc.) to individual countries and then adding them to the hosts by two-letter country domains.

Netcraft surveys Web servers in order to provide information about the software used on computers connected to the Internet. The data can be used to estimate the number of active Web sites under each domain, as well as the number of Web sites in each country by distributing gTLD and ccTLD registrations according to the country allocation of IP address blocks.

—A host is a domain name that has an IP (Internet Protocol) address "record" associated with it.

—Internet Protocol (IP) addresses are the numbers used to identify computers, or other devices, on a TCP/IP network.

—Servers are computers that host World Wide Web content.

—A top-level domain name (TLD) can either be a country code (for example .be stands for Belgium) or one of the generic top level domains (a so-called gTLD such as .com, .org, .net).

Source: OECD 2002, page 40.

Using ICT to measure ICT use

The basic question underlying this paper is whether new methods can be developed to collect data on the Information Society. In this paper, this question is interpreted as "Can we use ICT to measure ICT use?". As noted above, the difference between macro and micro data has important consequences for the data collection process and for the usability of the resulting datasets. In this overview of possible new measurement methods, these methods are classified according to their macro versus micro nature. To be more precise on this, macro data can very well be based on individual/personal data but there is no relation with other variables, the data are collected in a stand-alone setting and relations with other variables can only be modelled at an aggregate level, like is done above for the relation between culture and Info-state.

Macro data

Above I already listed some existing non-survey methods to collect data by using ICT. Here we describe some new directions that can be explored. These new directions describe mainly ideas to be explored in further research and are not fully available yet. More important, the data and methods listed are not specific targeted at household or individual use but in some cases households or individuals can be separated.

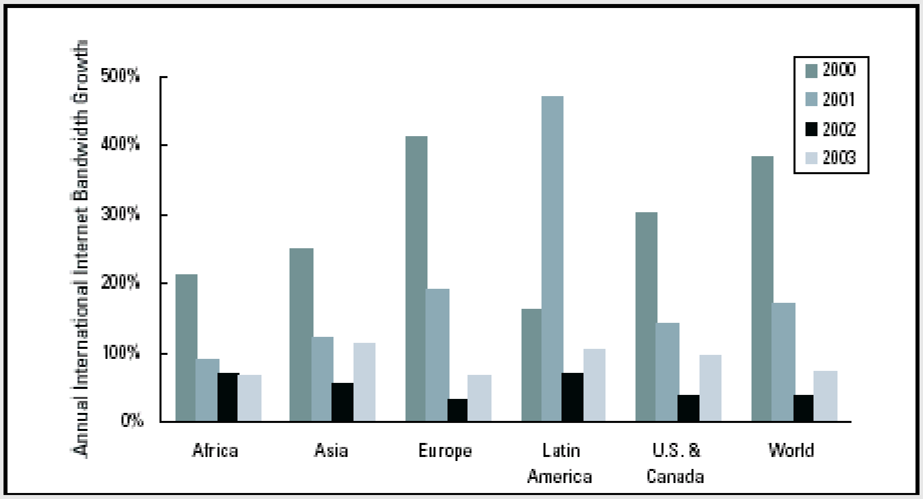
International bandwidth

A relative new indicator on the size of the Internet is the bandwidth measure as is used by e.g. Sciadas (2003) and also described by Abramson, (2000). The basic approach is to determine the bandwidth of Internet Exchanges and main Internet Service Providers. This gives insights in the global Internet infrastructure and can be used as a proxy for actual traffic. Figure 2 shows the Internet bandwidth growth by region as collected by Telegeography[4], showing a reacceleration of bandwidth growth after a slowdown of the growth rate until 2002. Note that average bandwidth growth is still 38 percent in 2002 and is growing to 74 percent in 2003. Of course much more detailed data can (and are) be gathered from the underlying data, including actual traffic and prices. Below, we will argue that such analysis can be extended to the type of traffic and even can —at least in principle— be extended to individual use.[5]

[4] See www.telegeography.com, a research division of PriMetrica Inc., a private Internet research company

[5] Individual refers here to a specific IP-address and is not necessarily related to persons, but could be ralted to persons to some extent after some corrections.

Figure 2. International Bandwidth Capacity Growth by Region 2000-2003



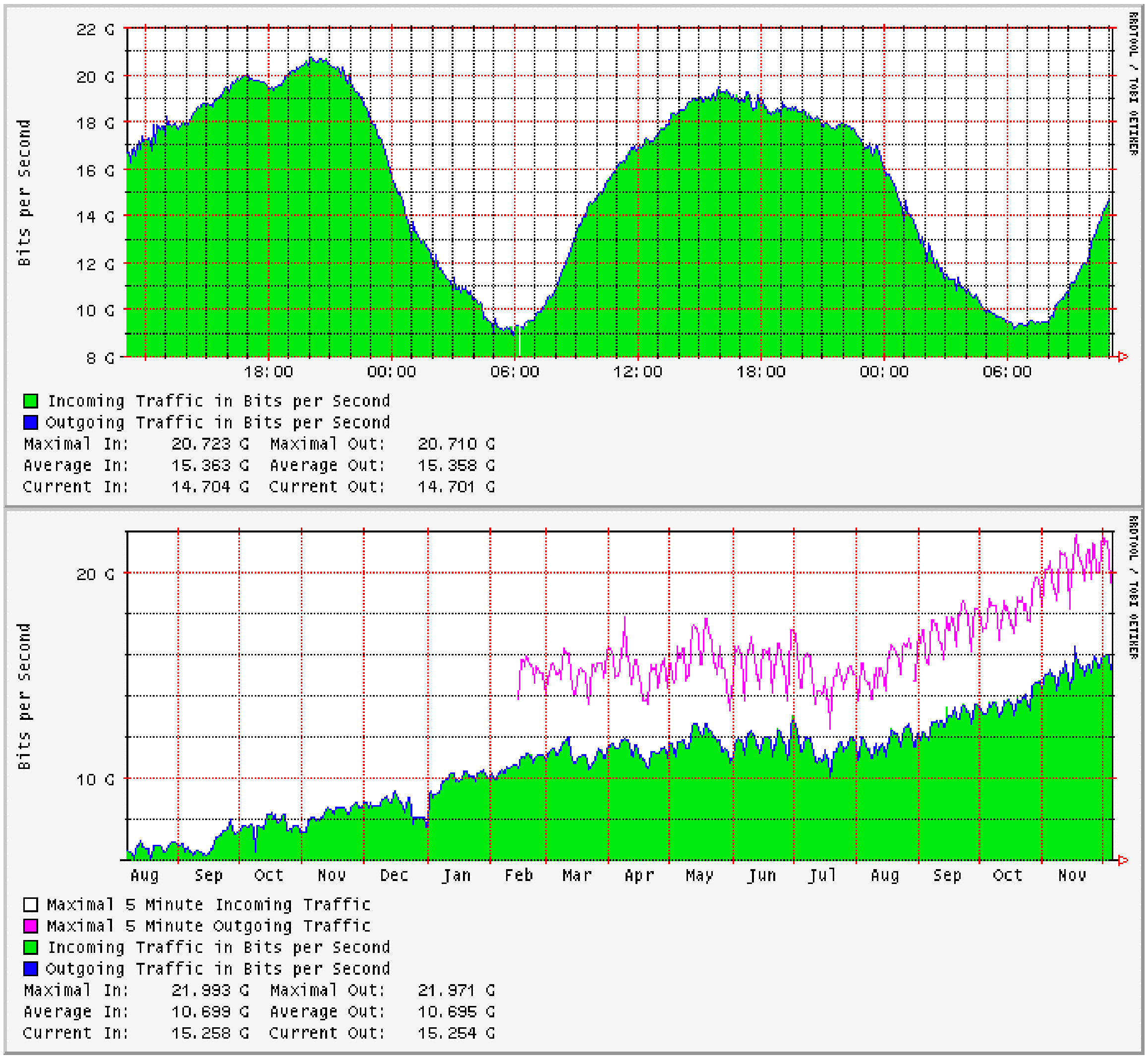
Notes: Data as of mid-year.

Source: TeleGeography research

© PriMetrica, Inc. 2003

So bandwidth capacity and actual Internet traffic can be analysed worldwide, giving an indication of actual use. Of course this worldwide traffic is based on traffic of individual Internet Exchanges, such as the traffic at the Amsterdam Inter Exchange displayed in Figure 3, which shows the remarkable amount of traffic varying between 10 and 20 Gigabit *per second* at just one Internet Exchange.

Figure 3. Hourly (top) and monthly (bottom) traffic statistics at AMS-IX



New possibilities

This section briefly describes three new measurement techniques were the focus is on Internet use. Two can be used to collect macro data whereas the last one collects data at a micro level. A method to measure non-users and their motivations and characteristics is not included such that classical (omnibus) surveys are still needed to capture and to fully understand the digital divide.

Detailed Traffic analysis

One line of further investigation is to analyse traffic data at the main Internet routers, i.e. the Internet Exchanges. This is in someway an extension of the International bandwidth measure in the sense that not only capacity or actual traffic is measured, but also provides an overview of type of traffic (type of packets and type of applications), origin and destination IP addresses and whether a packet is a request or a reply.[6] The packet header reveals the used protocols, mainly being TCP/IP for Internet traffic, but more important also reveals the port number which resolves into the application used, such as HTML traffic (HTTP), file transfer, mail traffic, peer-to-peer network traffic, etc. etc. So from the packet headers one can derive a wealth of information without the need to look into the data part of each packet. Such statistics could reveal:

- IP numbers: the amount of traffic as well as the type of traffic that originates from each single IP-address can be traced. Existing, but still limited, databases contain the location of IP numbers and this also reveals the location of the sender/receiver, at least to some extent. In almost all cases the country of origin can be determined and in many cases also a more specific region/location is known. Moreover, combining the IP numbers with information from Internet Services Providers, this even can give insights in the nature of the sender/receiver, i.e. being an individual, firm, education or research institute etc. as well as its (exact) location.
- Type of use. Each application uses its unique port number and this port number is enclosed in the packet header. This reveals the type of packet, being part of an email transfer, a web-content transfer, a peer-to-peer file transfer etc. Also whether it origins from a request (so being a request in the case of web-

[6] As defined by the Internet standards, all data are transformed into packets of limited size and each packet contains information in the packet header. So each packet contains a header part and a data part.

transfer) or from a reply (the web-server that sends the information to the requesting person) can be easily traced. Combining information from IP number and type of use clearly offers new insights in the use of the Internet, not only by perceived or estimated behaviour, but by actual traffic.

Gains and Shortcomings

Main problems of this approach are privacy issues and practical problems that arise when retrieving information from Internet Exchanges. Since the Internet Exchanges will protect their clients, they will initially be concerned about the information that is retrieved (including privacy-related issues) and whether it will harm their clients. Another problem is more of technical nature. Internet Exchanges such as the Amsterdam Internet Exchange (AMS-IX), the main Internet Exchange in the Netherlands to which all ISP's are connected, has currently peak traffic of about 20 to 25 Gigabit per second, as is shown in Figure 3. Assuming an average packet size of 600 bytes, which is about the average size of HTTP packets, this implies traffic of about 5 million packets per second. Since such amount of data cannot be handled as a continues stream, sampling techniques have to be employed to reduce the amount of data to a useable but still statistical meaningful size.

Concerning costs of such approach, the initial costs will be fairly high: establishing contracts with Internet Exchanges, creating and extending databases of IP numbers, creating software to analyse and aggregate raw data etc. However, once established the marginal costs are fairly low since the whole process can be automated. Of course costs for maintenance of software and databases.

Furthermore, interpretation problems arise since such traffic measures do not reveal the actual but will be a proxy for it. The main issues are:

- Internet traffic does not have to pass an Exchange. For instance all traffic within a network, i.e. from one user to another using the same ISP will obviously not pass the exchange. The use of proxy server also limits the measured traffic since cached (web) content will be delivered directly and no request will pass the exchange. The same applies for tunnelling techniques (such as VPN) such that the correspondents between an IP-address and the MAC address is lost.[7]
- An IP number does not necessarily reveal individuals or households. The popularity of Internet Café's and other public or semi-public Internet providers clearly disturbs the one-to-one relation between an IP number and a person or household. Individuals also use the Internet at work for private business and this information cannot be traced.

A gain for such approach is the potential worldwide coverage such that the number of users can be estimated at various levels of (geographical) aggregation and Internet use can be monitored worldwide in a harmonized way. This does not imply, however, that there are there are no biases in these data since use of proxy servers, for example, can differ between regions. Another gain is that, if corrected for biases indicated above, such measure reveals details that cannot be measured by surveys or other methods, except the Netratings approach as discussed below.

Web log Analysis

A collection of Web documents requests are kept as log files on all Web servers worldwide, giving a huge potential for further analysis and for data collection, not only for individual website and exposure analysis but also —if aggregated— for statistics on a broader scale. Web logs in general keep track of the requested document

[7] A MAC (Media Access Control) address is unique for every Network Interface Card.

or file, the IP number of the computer/device requesting the file, the operating system that is used on this device, type of web browser and time and date of the request. Moreover, search-engines keep also track of requested information, i.e. the items searched. These logs are already frequently used to measure website exposure but using these data to collect aggregate information at a wider scale is still missing.

A promising way to use these data is to collaborate with popular services like Google, Microsoft, Yahoo etc. in addition with some country-specific popular services and to collect the information from their web logs. This of course requires corporation of these services but once established analysis of simple items as IP numbers can reveal (or proxy) worldwide Internet penetration. Log files of search engines also discloses information about the type of information that is searched for. A simple example is the so-called Zeitgeist project of Google, which shows per region the most popular search items. Extension of such analysis and combining search requests with IP-address information reveal detailed information of the use of these services. Such analysis can also be extended towards (major) web shops, Internet portals etc.

One of the major gains is the relative low cost involved in such process whereas the coverage is (almost) worldwide from the outset. It is also limited to web sites visits and does not include other Internet activities such as email, file transfer etc. Furthermore, the usual problems of matching IP-addresses to individuals remain, as well as non-logged usage due to proxy servers. Note however that proxies almost never keep the results from search engines in their caches implying that all traffic is forwarded to the search engines. Masking IP-addresses by using proxy servers, tunnelling techniques and the use of NAT still poses yet unresolved problems to relate MAC addresses, so individual devices, to IP addresses. On the other hand, it is fairly easy to determine whether an IP address refers to a proxy in most cases, making it possible to treat that information separately.

Micro data

Micro data obviously contain responses on a coherent set of variables for each statistical subject. As noted above, they contain a wealth of information that cannot be retrieved from aggregate (macro) data whereas evolving policy questions typically demand for such detailed information. Surveys are the most traditional and most widely used technique, which are carried out by face-to-face interviews (traditional or Computer Aided Personal Interviews), telephone based interviews (traditional and Computer Aided Telephony Interviews), mailed questionnaires (postal or by email) and online questionnaires (website based). They all have advantages and disadvantages such as response rates being the highest for fact-to-face interviews and poor for postal questionnaires, maximal length of the questionnaire (long for face-to-face interviews), costs (high for face-to-face, low for online surveys) and statistical properties like geographical spread, stratification problems, selection bias etc. The choice of employed methodology depends highly on the nature of the survey, an analysis of non-ICT-users concerning their motivation being a non-users clearly cannot be carried out online and similar argument hold true for other Information Society Technologies such as (mobile) telephony penetration.[8]

A fairly new method to measure individual computer activity as is done by Nielsen//NetRatings by using a fairly simple computer program that runs on the computer of the surveyed person and that monitors sites the user visits and programs he or she uses. At this moment NetRatings provides worldwide web site ratings based on a sample of over 225,000 individuals in 26 countries. From these data one can measure actual use and type of information that is retrieved from the web. Also asking them to send a floppy disk containing the program's log files at regular time intervals,

[8] For a comparison between telephony based surveys and online survey see for instance Sanders at al. (2002).

e.g. monthly, can monitor non-connected persons to measure computer use in general. So the coverage is not necessarily restricted to Internet users and also non-connected computer users can be included. Collecting other detailed information about the surveyed individuals and households creates a database with a wealth of information. Note that detailed computer and Internet use from home as well as from the workplace can be measured separately. Apart from being a computer or Internet user, such a survey has to set up carefully to overcome geographical clusters of respondents, stratification problems etc. etc. and more research is needed on these issues to create statistical reliable data.

The gains of such approach are the relative low costs (as compared to face-to-face or telephone based surveys) and the possibility to measure more detailed user activities in an objective manner, which is less or even not possible with other surveying methods, and this method thus leads to new statistical information. Information concerning the most frequently hosts that are visited also gives information for the above-mentioned web log analysis approach since first, the most popular websites result from this information and thus can be included if needed and second, it gives information about the population that is ignored by restricting that approach to the most popular websites, for instance the number of people that never use Web-applications like the visiting the World Wide Web but restrict their use to for instance email.

A difficulty is that only computer users can be surveyed and that inclusion of public Internet Points such as Internet café's will probably call for large resistance. Also possible biases have to be researched since some specific groups of Internet users may refuse to be monitored.

Conclusions

Measuring the Information Society is measuring a moving target. Policy questions evolve as Information Society Technologies further diffuse, as the variety of these technologies increases and as they become more and more important in societies. Not all regions are gaining from these developments and also in within each region, there are many who are excluded, voluntary or forced by circumstances. Politicians, governments, and researchers, they all demand for statistical data to learn the actual state and to find evidence of possible explanatory theories, next to evidence on the consequences of Information Society Technologies, in economic terms, but also the social and cultural aspects.

The basic question of the paper is whether there are new methods to measure the Information Society and this question is interpreted as: "Can we use ICT to measure ICT". From an analysis of IST developments and the development of policy questions, it is concluded that aggregate statistics are needed but that micro data are in many cases crucial to get a real understanding of the Information Society. This paper presents some promising new methods including an overview of their gains and shortcomings. The focus is on Internet use leaving aside many other artefacts of the IS. In some way the presented methods are extensions of current techniques and concern traffic analysis, web content analysis and automated computer based measuring. Both traffic analysis and web content analysis will gives insights in aggregate statistics on like who is connected (how many people per country or region), what type of information is requested, what is the Internet used for. The automated computer based measuring methods is already practised by private companies but not by statistical offices. This method seems very promising and offers detailed micro data on the actual use of IST.

So new methods to create data at an aggregate as well as at a micro level are presented and more research is needed to get insights in their potential and their pros and cons vis-à-vis existing methods. Measuring characteristics of the non-users by IST seems, however, fairly impossible. So many policy questions, like the background of the digital divide, both at an international and at a national level, still demands classical surveys. It is demonstrated that using IST to measure IST is a fact but using IST to measure non-IST use is still an unresolved paradox, a fiction.

Bibliography

Abramson, Bram Dov, 2000, “Internet globalization Indicators”, *Telecommunications Policy*, vol. 24, pp. 69-74

Chen, Wenhong and Barry Wellman, 2003, *Charting and Bridging Digital Divides: Comparing Socio-economic, Gender, Life Stage, and Rural-Urban Internet Access and Use in Eight Countries*, Global Consumer Advisory Board, http://www.amd.com/us-en/assets/content_type/DownloadableAssets/FINAL_REPORT_CHARTING_DIGI_DIVIDES.pdf, accessed 1 December 2003.

Cottrell, R. Les and Warren Matthews, *Measuring the Digital Dived with PingER*, paper presented at the Round Table on Developing Countries Access to Scientific Knowledge, Trieste, Italy, http://www.ejds.org/meeting2003/ictp/papers/Cottrell-matthews.pdf, accessed November 2003

Lenhart, Amanda, John Horrigan, Lee Rainie, Katherine Allen, Angie Boyce, Mary Madden, and Erin O’Grady, 2003. "The Ever–Shifting Internet Population: A New Look at Internet Access and the Digital Divide," The Pew Internet & American Life Project, at http://www.Pewinternet.org/, accessed 4 November 2003

Leo, A.D., & Gabriele, R. (2000). Information Technology in Italy - Internet Diffusion. Kogol School of Business, American University. http://www.american.edu/carmel/rg4784a/page3.cfm

Miles, Ian, 2000, Rethinking Organisation in the Information Society, *paper presented at the SOWING Conference in Karlsruhe*, November 2000. http://www.eforesee.info/malta/ianmiles2.pdf, accessed 3 November 2003.

National Telecommunications and Information Administration (NTIA), 2002, *A Nation Online: How Americans Are Expanding Their Use of the Internet*, http://www.ntia.doc.gov/ntiahome/dn/Nation_Online.pdf, accessed 15 November 2003

OECD, 2002, *Measuring the Information Economy*, OECD, paris

Pastore, Michael, 1999. "U.S. Internet audience growth slowing," at http://cyberatlas.Internet.com/big_picture/geographics/print/0,,5911_246241,00.html, accessed 2 November 2003.

Sally Wyatt, Graham Thomas, and Tiziana Terranova, 2002. "They came, they surfed, they went back to the beach: Conceptualizing use and non–use of the Internet," In: Steve Woolgar (editor). *Virtual society? Technology, cyberbole, reality*. Oxford: Oxford University Press

.

Sanders, David, Harold Clarke, Marianne Stewart, Paul Whiteley and Joe Twyman. "The 2001 British Election Study Internet Poll: a Methodological Experiment." Paper prepared for delivery at the 2002 Annual Meeting of the American Political Science Association, Boston, August 29-September 1, 2002. http://apsaproceedings.cup.org/Site/abstracts/040/040001TwymanJoe0.htm

Sciadas, George (editor), 2003, *Monitoring the Digital Divide….and beyond*, Orbicom, http://www.orbicom.uqam.ca/index_en.html, accessed 18 November 2003.

Social and Cultural Planning Office of the Netherlands (SCP), 2001, *The Netherlands in an European perspective, Social and Cultural Report 2000*, The Hague, http://www.scp.nl/english/publications/english/20010426/download/scr2000.pdf, accessed November 2003.

Statistical Indicators Benchmarking the Information Society (SIBIS), 2003, *Matching up to the Information Society An evaluation of the EU, the EU Accession Countries, Switzerland and the United States*, an EU IST funded project, www.sibis-eu.org, accessed 20 November 2003.