

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’

D0.4 Final technical report

April 2014

Document Service Data

Type of Document	Deliverable		
Reference:	D0.4 Final technical report		
Version:	3	Status:	Draft
Created by:	Michalis Petrakos, Photis Stavropoulos	Date:	30/4/2014
Distribution:	European Commission – Eurostat/G4, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	31/12/2013	Initial release
2	20/3/2014	Completion of the sections that were missing from version 1.
3	30/4/2014	Revision of sections concerning deliverables D3 and D4 and of the respective annexes

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

1. Executive summary.....	4
1.1. Conceptual framework and recommendations for internet data – based ICT statistics	4
1.2. Consultations with statistics authorities, business web sites and individual Internet users.....	7
1.3. Feasibility of internet data – based ICT statistics	8
1.4. Pilot testing of specific internet data – based ICT indicators	9
1.5. ‘Cookbook’ for internet data – based ICT statistics	12
1.6. Feasibility of big data as a source for the production of official statistics.....	12
1.7. Outline of procedure for the accreditation, by producers of official statistics, of big data sources as input data for official statistics	14
1.8. Conclusions.....	15
2. Introduction	17
3. Conceptual framework and recommendations for internet data – based ICT statistics....	17
3.1. IW4OS: A novel conceptual framework	17
3.2. Mapping current ICT statistics against the Internet as a data source	21
3.3. Considerations about the future	27
4. Consultations with statistics authorities, business web sites and individual Internet users	31
4.1. Consultation with statistical authorities	31
4.2. Consultation with business web sites	32
4.3. Consultation with individual Internet users.....	32
5. Feasibility of internet data – based ICT statistics.....	33
6. Pilot testing of specific internet data – based ICT indicators.....	35
6.1. Pilot survey of Internet usage by individuals	35
6.2. Pilot survey of the characteristics of the web sites of business enterprises	39
7. ‘Cookbook’ for internet data – based ICT statistics.....	43
8. Feasibility of big data as a source for the production of official statistics	46
8.1. Vessel movement data from the Automatic Identification System (AIS)	46
8.2. Real estate classified advertisements	47
8.3. Social media message data	47
8.4. Credit card transaction data (Visa Europe)	48
8.5. Government financial transparency portal data	48
9. Outline of procedure for the accreditation, by producers of official statistics, of big data sources as input data for official statistics	49

10. Conclusions	51
11. References.....	52
12. Annex – Technical deliverables of the project	54
12.1. D1 - Definition of Internet data-based indicators part I	55
12.2. D1 - Definition of Internet data-based indicators part II	117
12.3. D2 – Results of the feasibility analysis	146
12.4. D3 – Results of the testing of the two methods	229
12.5. D6 – Cookbook for the implementation of new methods and indicators at national level ..	297
12.6. D4 – Feasibility analysis of selected data repositories for official statistics	350
12.7. D5 – Accreditation procedure for statistical data from non-official sources	398
12.8. D7 – Powerpoint presentation to the Information Society Working Group	433

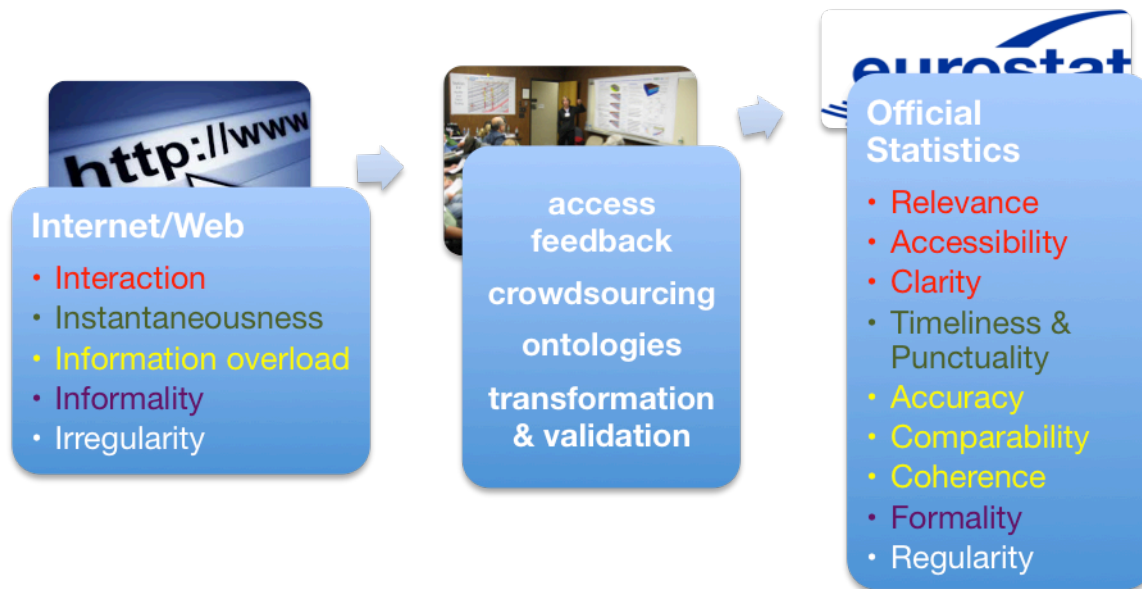
1. Executive summary

The present report summarizes the results of project ‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’, which have been presented in detail in specific technical deliverables.

1.1. Conceptual framework and recommendations for internet data – based ICT statistics

In this activity we carried out two related investigations. The first one was theoretical and examined the place of the Internet in everyday social and economic life and the data that are generated by the interactions between individuals and enterprises in the Internet. The second one examined the extent to which data collected by software monitoring users’ devices and by crawlers extracting content from enterprise web sites can substitute or extend the current ICT surveys. In addition, and although beyond the scope of the project, it examined the new ways opening for production of official statistics based on the proliferation of data in the Internet and on data from the “Internet of things”.

IW4OS: A novel conceptual framework



The new sources and forms of data in the Web are raising imperative questions to Official Statistics. The envelope question is which methods should be changed or even introduced to let Official Statistics retain their character, but at the same time exploit the emerging potential of online contexts. The proposed conceptual framework for Internet and Web as data sources

should facilitate the orchestration of their main characteristics with the approach of Official Statistics.

Interaction

The traditional triptych of producers-exchange-consumers has been replaced by the presumption model where consumers contact producers directly or can act, at the same time, as producers. These new modes of human interaction and production could be incorporated in providing more accessible and relevant Official Statistics to the users.

Instantaneousness, Information Overload, Informality & Irregularity

Web 3.0 technologies, such as Semantic Web have been engineered to provide assistance to locate information by human and machine-based tools. Existing ontologies and vocabularies have been expanded to handle online statistical information and mainstream statistical standards.

The transition from Official Statistics obtained by real world data through surveys and personal communication with individuals, to a new era of indicators computed complementarily or even solely from Internet and the Web is not easy or obvious. We have to study in depth and understand the universe of Internet and the Web as an extremely complex system in order to fully utilize it for obtaining Official Statistics through the proposed conceptual framework.

Mapping current ICT statistics against the Internet as a data source

The project examined the variables collected in the current ICT surveys and identified those on which data can be collected from the Internet. As a rule of thumb, questions related to matters of access cannot be answered from the Internet. To the extent that their measurement is important, the availability of computers, desktop or portable, mobile phones and other ICT devices cannot be known from the Internet. To some extent, this is an oxymoron and reminiscent of the digital divide: knocking at the door of the “haves”, you cannot find the “have-nots”. Generally, the Internet as a data source is ideally situated for the measurement of indicators of use. Lastly, the Internet is not meaningful for content concerning the views by individuals or businesses of their experiences or any subjective assessments and opinions.

Considerations about the future

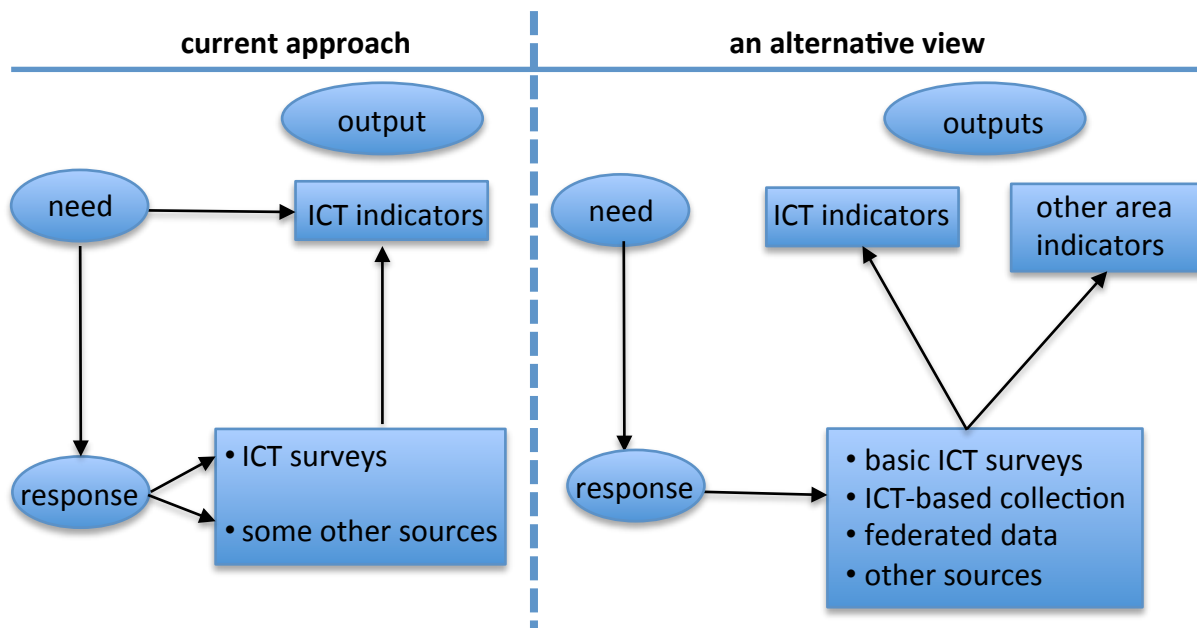
The Internet of things

The Internet today provides access to continuously increasing amount of information universally, at any time and from any device. In the evolving Internet of Things (IoT) landscape, any device equipped with sensors is essentially an information warehouse, capable of collecting and transmitting real-time data originating from and interacting with the surrounding environment (people, places and things).

Sensor devices and social interactions along with powerful applications can provide data for calculating various indicators related not only to ICT use and their social impact but also to other financial and social indicators related to either individuals or enterprises. Sensor data can be used for official statistics related to agriculture, forestry, environment, urban traffic and accidents, travels, health services, tourism, natural disasters, etc. Interaction of sensors with humans through applications converting sensor data to natural language expressions and social media is a potentially interesting perspective for validating the quality of data. On the other hand, this potential source of official statistics requires powerful technological infrastructure.

Dis-assembling and re-assembling

The new situation calls for new models. We need to realize at a deeper level that the “whole” questionnaire-based approach will have to be broken down to pieces that fit the new reality. The following schematic displays simply what all this means.



Under the habitual approach, data needs (typically advocated by policy makers) were met through a survey or an administrative source. This essence of the approach connecting new needs to eventual statistical answers is depicted on the left-hand side of the schematic.

Today more options for responding become available (right-hand side). Additional possibilities open up, which may turn orthodox processes upside-down. It is possible that in the process of tapping the new resources to answer a defined set of questions, answers to totally different questions can be fetched. Moreover, identifying the data that can be collected from where they exist and communicating such information to the demand side their thinking may be influenced in a way that they modify the questions asked.

1.2. Consultations with statistics authorities, business web sites and individual Internet users

Consultation with statistical authorities

Discussions about the feasibility of Internet-data based methods were held with four National Statistical Institutes (NSIs) from the European Statistical System (ESS). The discussions revolved around the experiences they might have had with such methods and around their opinions about these methods in general.

The picture that emerges from these discussions is firstly one of “no objection” to the new methods. Most NSIs view the new methods favourably as production tools. They experiment with them and assess them with the same procedures they assess the quality of production processes. They are concerned about the accuracy of their results but in most cases they find it satisfactory, while they recognise the gains in timeliness they offer.

The legal setting is not clear for any of the NSIs. It is not clear to them if the consent of individuals or enterprises whose data are collected or of the owners of the data is sufficient to make the methods “legal”.

Consultation with business web sites

A sample of 61 randomly selected websites was used in order to investigate, via a questionnaire, whether they are willing to accept and implement the proposed new method of data collection. We have prepared a questionnaire, which outlined the proposed method and indicators and posed five questions in order to collect their opinions about them. Out of the 61 selected websites that were contacted, 27 (44,3%) websites’ owners replied, 16 (26,2%), refused to take part and 18 (29,5%) never replied.

Of the 27 that did offer their responses, almost half would accept an automatic data collection system but they require some bilateral agreement before they do so. So a large part of websites (about half or more) will refuse cooperation or not reply at all and those that can potentially agree see themselves as partners and not just respondents and require bilateral cooperation agreements rather than self-imposed rules and commitments from the National Statistical Institute.

Consultation with individual Internet users

In this section we examined attitudes of individuals towards a system of data collection for statistical purposes from their day to day activity. Most of the respondents (38/40 i.e. 79%) did not have reservations and 10 (21%) provided some. Confidentiality was the main concern and most users stated anonymity as a condition for accepting software installation. Overall, we found that most users want to cooperate and will do so if they are satisfied that their privacy and

anonymity will be preserved and their use of their devices will not be affected in a substantial way. Incentives may help to further increase cooperation

1.3. Feasibility of internet data – based ICT statistics

Two separate production processes, one web site-centric and the other user-centric have been examined:

- the production of statistics on the characteristics of business web sites, based on data collected with the help of crawlers or search engines that rely on earlier crawling from the said web sites.
- the production of statistics on the use of Internet by individuals, based on data collected with the help of monitoring software installed on the users' devices.

The two processes have been examined from several angles.

Technically they are both feasible. Software components are available in several forms and the software technologies needed for development from scratch are commonplace. The capacities needed for development and maintenance are quite easy to find in the job market even if not already available to the NSIs.

The two processes diverge in the conclusions about their methodological feasibility. They both produce very relevant, timely and rich-in-detail statistics. Compared to the current ICT surveys the web-site centric process has a much narrower scope: it substitutes and expands a small subset of the current survey's indicators, while the user-centric process can reproduce most current indicators. The user-centric process thus also offers great savings in response burden. Both have accuracy issues: the web site-centric one suffers from measurement errors, in its keyword-based implementation and possibly by non-response. The user-centric one mainly suffers from non-response, manifested as refusals to participate or switching off of the monitoring software occasionally.

The two processes also achieve different cost-benefit balance. The web site-centric process seems to have too high costs for the benefits it offers, especially if one takes into account that it covers a small subset of current indicators and has reduced accuracy. The user-centric approach seems to be more expensive than the current ICT survey but reduces response burden and production times considerably. Unfortunately there was no detailed cost information about these processes or the current ICT surveys so as to make a more precise assessment.

The processes are compatible with current European legislation, as long as NSIs inform explicitly individuals and enterprises about the collected data and the uses they will be subjected to and they obtain the sample units' consent. In principle the processes do not differ from traditional surveys that collect sensitive business or personal data.

In user centric approach we found that most users want to cooperate and will do so if they are satisfied that their privacy and anonymity will be preserved and their use of their devices will not be affected in a substantial way. Incentives may help to further increase cooperation. Regarding the site centric approach, a large part of websites (about half) will refuse cooperation and those that can potentially agree see themselves as partners and not just respondents and require bilateral agreements rather than self-imposed rules and commitments from the National Statistical Institute.

Overall, the user-centric process is the more feasible of the two. It can replace the current ICT survey to a great extent for a not much higher cost. The same cannot be said for the web-site process. As envisaged it collects a small subset of the current survey's indicators. A variation, namely the collection of data from enterprise servers, which was outside the scope of the project, can supplement this process and can deliver a much larger set of highly relevant ICT and other enterprise data.

1.4. Pilot testing of specific internet data – based ICT indicators

Two separate pilots were implemented, one targeting individuals and the other the websites of enterprises. Each pilot is the subject of a separate section of this chapter.

Pilot survey of Internet usage by individuals

Statistical indicators produced

Three indicators have been produced by the pilot survey referring to specific types of activity:

1. Share of users that have engaged in each type of online activity
2. Percentage of time online that users devote on average to specific types of activities.
3. Amount of time that users devote on average per day to specific types of online activities.

Sampling

Due to difficulties in obtaining a proper random sample from the Hellenic Statistical Authority a non-random sample was used. A panel of persons compiled by a Greek market research company for use in opinion surveys was used as sampling frame. Its members were offered a monetary incentive of €30.00 each. Due to this cost, as well as the cost of the monitoring software it was decided to restrict the sample to 150 persons and devices. After three reminders, 145 persons accepted to participate. Due to the difficulties in installing the software or due to second thought perhaps, we finally managed to enlist only 48 persons in the sample.

Software tool

The software selected for monitoring and recording the users' activities was the online parental controls service Qustodio¹.

Implementation

Implementation lasted one month. The first 10 days were spent deploying the software to the sample members. The remaining days were spent on collecting usage data. During the course of the collection, users were sent an email questionnaire requesting some demographic data and also some Internet usage data. These data were combined with those collected by Qustodio.

Conclusions

The types of activities can be discerned at great detail and therefore rich classification can emerge for statistical use. Moreover, the fact that data are recorded at great detail also allows the change of the classifications to fit changing statistical needs. In addition, historical data can be converted easily to the new classifications. The variations of usage time can be observed and also reported to the desired degree of temporal detail. Finally, data can be combined and jointly analysed with data collected with regular questionnaires.

On the other hand the method has disadvantages. The most serious is the lack of trust from individuals towards the producer of statistics. The possibility of a financial incentive should not be ruled out by NSIs. The chosen software cannot work on devices with the iOS operating system, i.e. iPhone and iPad. This excludes a substantial share of the target population from the survey. An additional problem in the pilot study was the lack of transparency of the measurement process implemented by the tool. An NSI must not accept this; it should have complete knowledge of what each measurement means. Care is therefore needed in the selection of the software tool; the development of bespoke solutions might be necessary.

Overall, the use of activity monitoring software shows great promise as a data collection tool and the ESS should carry out additional investigations of the statistical methodology and practical arrangements needed for its incorporation in regular statistical production.

Pilot survey of the characteristics of the web sites of business enterprises

Statistical indicators produced

All indicators that have been produced in the pilot survey are of the sort "Percentage of enterprises whose website ..." and they refer to whether the site provides specific types of information, uses particular types of technologies or offers certain facilities to its users. An enterprise's website has been defined as the set of pages whose addresses start with the same single URL that characterizes the enterprise.

¹ www.qustodio.com

Sampling

Due to difficulties in obtaining a proper random sample from ELSTAT the project team decided to resort to a non-random sample. It was drawn from a list of enterprises, which contains contact details of Greek enterprises that have received in the past European funding for research. The total list contains 1777 enterprises. A random sample of 281 enterprises was drawn from this list.

Software tool

The tool used was Google's Custom Search Engine (CSE), instead of any specific crawling utility. It provides an interface to the user in order to specify a list of sites and a list of keywords to search for in these sites. The indexing by Google has already carried out crawling of sites and therefore we implicitly relied on crawling too.

Implementation

The collection of the data relies on the use of keywords. Each of the indicators is viewed as resulting from answering "Yes" to a question asking whether the website has / provides / uses / offers the mentioned type of content or facility. Instead of asking questions we specified a number of keywords relevant to each indicator. Appearance of even one of these in at least one page of a website was considered as a "Yes" to the corresponding fictional question.

The CSE returns a list of URLs (pages, within each website) where any of these keywords has been found. Therefore, if for example site www.agilis-sa.gr contains in four of its pages the keyword "telephone" and in three more (possibly overlapping) it contains the keyword "tel", the results will list seven URLs with the keyword found in each one attached to them. Post-processing was therefore carried out with a text parser which grouped such findings into a single "hit" per indicator and website.

Conclusions

The results of the particular approach chosen are not very encouraging. The data returned by the search engine contain many spurious findings while on the other hand several occurrences of the site characteristics in which we were interested went un-noticed. This is a deficiency of keywords.

Detection capabilities could possibly improve with keywords in the national languages of each country, with linguistic analysis of a site's content, with searching for keywords in the site's HTML source code (when retrievable) and with image analysis or image search (to identify key icons; e.g. the logos of Facebook or Twitter).

Besides site features that are manifested through keywords that cannot be specific enough there are other features which are not connected to verbal aspects of the sites. For example, video thumbnails may be the links to Youtube videos, without any keywords. Furthermore, web

analytics may be deployed on a site invisibly to its visitors. Such features require the utilisation of tools that detect technologies rather than keywords.

Based on the results of the pilot study it can be inferred that the developed methodology for collecting data from enterprise web sites does not produce statistics of high enough quality. A more extended appraisal of the method, which will encompass aspects of multilingualism, extraction of source code and detection of technologies, is needed for a more informed decision about its usefulness.

1.5. ‘Cookbook’ for internet data – based ICT statistics

The ‘cookbook’ is a guide for the application of Internet-data based methods for the production of official statistics. Its audience are the producers of official statistics. The guide borrows its structure and some of its content from Eurostat’s “Methodological manual for statistics on the Information Society”². More specifically, for aspects of the production methods, which will be implemented in the same manner as in the current households and enterprises ICT surveys (e.g. sampling enterprises from the business register of the NSI) the guidelines were copied from the current manual. Even then however, minor changes were made in order to discuss possible difficulties that will be faced by the new methods. A considerable part of the cookbook however consists of original material drafted by the project team.

1.6. Feasibility of big data as a source for the production of official statistics

The potential of big data as a source of official statistics was examined. Of particular interest were the so-called ‘federated open data’ which are (big) data from business or the public sector, generally not accessible by the public, but shared in an agreed and defined way with the producers of official statistics. Five specific ‘use cases’ were examined:

- Vessel movement data from the Automatic Identification System (AIS)
- Real estate classified advertisements
- Social media message data
- Credit card transaction data (Visa Europe)
- Government financial transparency portal data

Vessel movement data from the Automatic Identification System (AIS)

There is high potential in using AIS data in the production of maritime transport or emission statistics. A potential data source for obtaining AIS data is MarineTraffic³. Although some data

² Eurostat (2013) Methodological manual for statistics on the Information society, v. 3. Luxembourg: Eurostat.

³ www.marinetraffic.com

about vessels' characteristics may be missing or may not be readily available, these can either be estimated or obtained from an international database on vessel characteristics.

Real estate classified advertisements

There is a high potential in using Internet advertisement in the production of current statistics on the housing price index and Purchasing Power Parities (PPPs) related to rental and owner occupied housing. Moreover, there is some potential to using Internet advertisement in production of the owner occupied housing sub index of the Harmonised Index of Consumer Prices (HICP) although there are differences in concepts. It is however unlikely that data from Internet advertisements can replace the rent surveys for the HICP completely.

Social media message data

Social media provide useful input data for the production of subjective indicators, which are used in the current statistics. They provide sentiment information, however it is important to highlight that those sentiments cannot replace the existing official statistics and its indicators. The measures of sentiments and their scoring can be used complementarily to official statistics and provide us with useful trends over time as well as with comparisons between the different European countries.

Credit card transaction data (Visa Europe)

There are a lot of benefits from using Visa's data in the production of consumption expenditure statistics. Currently, the Household Budget Survey (HBS), which produces similar data, is carried out at an informal basis every five years. In fact, Visa Europe compiles an index, named "EU Consumer Spending Barometer"⁴, using real-time card transaction data. It is worthwhile using Visa as a source, in a complementary way, for the production of flash estimates about the structure and amount of consumption expenditure. However, it is important to highlight that an index similar to Visa's Barometer, cannot replace the existing official statistics and its indicators.

Government financial transparency portal data

The Greek government's transparency portal was examined. A huge amount of data on public expenditure is available through this portal. Data can be retrieved and processed for statistical purposes as they are publicly available and contain fields that can be linked to statistical classifications. On the other hand, there are several data entry errors and shortcomings in the current portal that was prepared as a pilot. Most of them are expected to be solved with a new version, expected in September 2014. Moreover, there are important impediments in terms of coverage; only expenses that require decisions are included. Therefore, the source cannot be used on its own but it can be used as a supplementary source so as to reduce the burden to public administration entities and to substantially improve timeliness. There are some specific areas however, where coverage is complete or near complete (e.g. public procurement, R&D spending); there the portal can serve as a primary source for statistics.

1.7. Outline of procedure for the accreditation, by producers of official statistics, of big data sources as input data for official statistics

In this activity we proposed a procedure that NSIs pondering whether to use big data sources as input in the production of official statistics could employ to accredit such sources. Our work was based on the analysis of the available recent literature on topics such as quality of statistics in general and quality of administrative data sources in particular.

The actual **accreditation procedure** evolves in a step-wise fashion. It consists of *five stages* with gradual assessments involving indicators measured through scales and hard data:

Stage 1: Initial examination of source, data and metadata. An early assessment of the data, the metadata and the source.

Stage 2: Acquisition of data and assessment. This stage entails negotiations with the source with a view to acquire a set of files or file extractions adequate for rigorous testing. The primary objective is to clarify whether the source is willing and able to deliver files or extractions at the record level, as well as keep open a communication channel during the testing process.

⁴ http://www.visaeurope.com/en/newsroom/all_reports/european.aspx

Stage 3: Forensic investigation. This stage requires a fair amount of work by the NSI. It is divided in four distinct phases: i) producing a clean microdata file (halfway through which we meet a decision point); ii) using the file to produce and analyse aggregate statistics iii) producing pilot new outputs or using the file in the production of existing outputs, and; iv) assessing the capacity of the existing statistical tools to handle the new data.

Stage 4: NSI decision. This stage is dedicated to the assessments necessary for a corporate decision to be made based on as much information and knowledge as possible. It can sub-divided in four distinct phases: i) an itemisation of the exact uses of the new data and their impacts; ii) a top-level cost-benefit analysis, which focuses on the financial picture; iii) assessment of the risks that need to be undertaken and managed by the NSI, iv) assessment of the feasibility of incorporating the new source into the gamut of the NSI's statistical operations from a legislative and socio-political point of view

Stage 5: Formal agreement with source. This final stage involves high-level negotiations with the source as an institution to secure cooperation and arrive at a formal and comprehensive agreement.

1.8. Conclusions

Two possible Internet data-based methods, one user-centric and one site-centric were examined. They are in line with the current proliferation of data in the Internet and with the necessity of using them for statistical production (at least not ignoring them without having examined them first).

The analysis of their feasibility demonstrated that they can produce very relevant statistics, with rich detail and in a much more timely manner than the current ICT surveys can manage. Their accommodation in the legal context of privacy and personal and corporate data is also feasible. The methods present problems too. They can lead to high refusal rates, especially in the case of individuals. The site-centric method tested in the pilot surveys suffers from problems of accuracy too. The user-centric method appears more costly than the current ICT survey but at least it can reduce considerably the response burden and processing time. The site-centric method's costs cannot be offset by the benefits it brings.

Moreover, we examined the potential of producing official statistics, in any domain, based on big data repositories. Five specific use cases were examined, providing data relevant to diverse domains such as: transport, environment, consumer sentiment, government finances, housing prices, consumer expenditure. In all cases large amounts of relevant data are available, at different degrees of openness. These data can produce, on their own or in combination with statistical data, existing statistical indicators (i.e. they can replace them) or new ones.

The presence of these potential data sources means that NSIs are “suddenly” confronted by a pool of sources much wider than the current one. In order to be able to shift through them and

identify those suitable for statistical production the project has proposed an outline of an accreditation procedure.

2. Introduction

The present report summarizes the results of project ‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’, which have been presented in detail in specific technical deliverables.

3. Conceptual framework and recommendations for internet data – based ICT statistics

In this activity we carried out two related investigations. The first one was theoretical and examined the place of the Internet in everyday social and economic life and the data that are generated by the interactions between individuals and enterprises in the Internet. Based on its results, we examined what Information Society statistics can be produced based on these data.

This second investigation branched out into two directions. The first one examined the extent to which data collected by software monitoring users’ devices and by crawler extracting content from enterprise web sites can substitute or extend the current ICT surveys. This would provide the context for the feasibility analysis and pilot surveys described in sections 4, 5 and 6 of the present report. The other direction is more far-reaching, beyond the scope of the present project and examined the new ways opening for production of official statistics based on the proliferation of data in the Internet and on data from the “Internet of things”.

3.1. IW4OS: A novel conceptual framework

The new sources and forms of data in the Web are raising imperative questions to Official Statistics. The envelope question is which methods should be changed or even introduced to let Official Statistics retain their character, but at the same time exploit the emerging potential of online contexts. The proposed conceptual framework for Internet and Web as data sources should facilitate the orchestration of their main characteristics with the approach of Official Statistics. This framework, the Internet and Web for Official Statistics framework (IW4OS) is presented in Figure 1.

Interaction

At the current Web 2.0 era, users are the protagonists of the online ecosystem because they can easily edit, interconnect, aggregate and comment online content as never before. Most of these opportunities can also be engineered in the personal level. The traditional triptych of producers-exchange-consumers has been replaced by the prosumption model where consumers contact producers directly or can act, at the same time, as producers. Web 2.0 enables interaction and crowdsourcing through openness, peering, sharing and acting globally (Tapscott & Williams, 2008).

These new modes of human interaction and production could be incorporated in providing more accessible and relevant Official Statistics to the users. For instance, social media can serve both as pools for data collection and data publication in order to get direct feedback from the online users about the usefulness of indices.

Instantaneousness, Information Overload, Informality & Irregularity

Web 3.0 technologies, such as Semantic Web (Berners-Lee, 2006) and Linked Data (Bizer, Heath, & Berners-Lee, 2009) have been engineered to provide assistance to locate information by human and machine-based tools. Existing ontologies and vocabularies have been expanded to handle online statistical information and mainstream statistical standards (e.g. Data Cube vocabulary (Cyganiak, Reynolds, & Tennison, 2012), Linked SDMX data (Capadisli, Auer, & Ngomo, 2013), etc.).

The most important aspect of the proposed analysis is to identify an effective set of transformation and validation rules that will enable the timeliness, punctuality, accuracy, comparability, coherence, and eventually, formality of IaD sources.

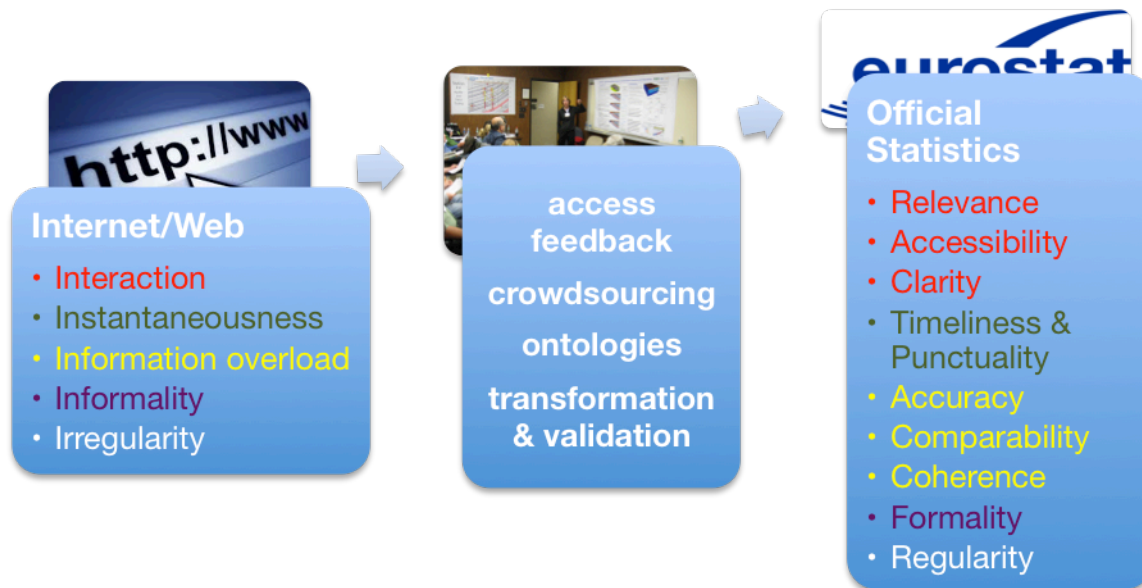


Figure 1. Internet and Web for Official Statistics framework (IW4OS) is designed to orchestrate the main characteristics of the online ecosystem and Official Statistics.

Based on the past experience in developing Internet and Web standards, these rules should not be all-encompassing from the beginning, but will better follow the “divide-and-conquer” and the procrastination principles. First, the general problem will be demarcated in smaller sub-problems (e.g. IaD for specific indices in ICT statistics) and second, according to the procrastination principle that can be summarized in the phrase “do not do anything that can be done later by

users ” most problems confronting the IaD approach can be solved later by other researchers and users of statistics.

The transition from Official Statistics obtained by real world data through surveys and personal communication with individuals, to a new era of indicators computed complementarily or even solely from Internet and the Web is not easy or obvious. We have to study in depth and understand the universe of Internet and the Web as an extremely complex system in order to fully utilize it for obtaining Official Statistics through the proposed conceptual framework.

The Web 2.0 economy in a nutshell

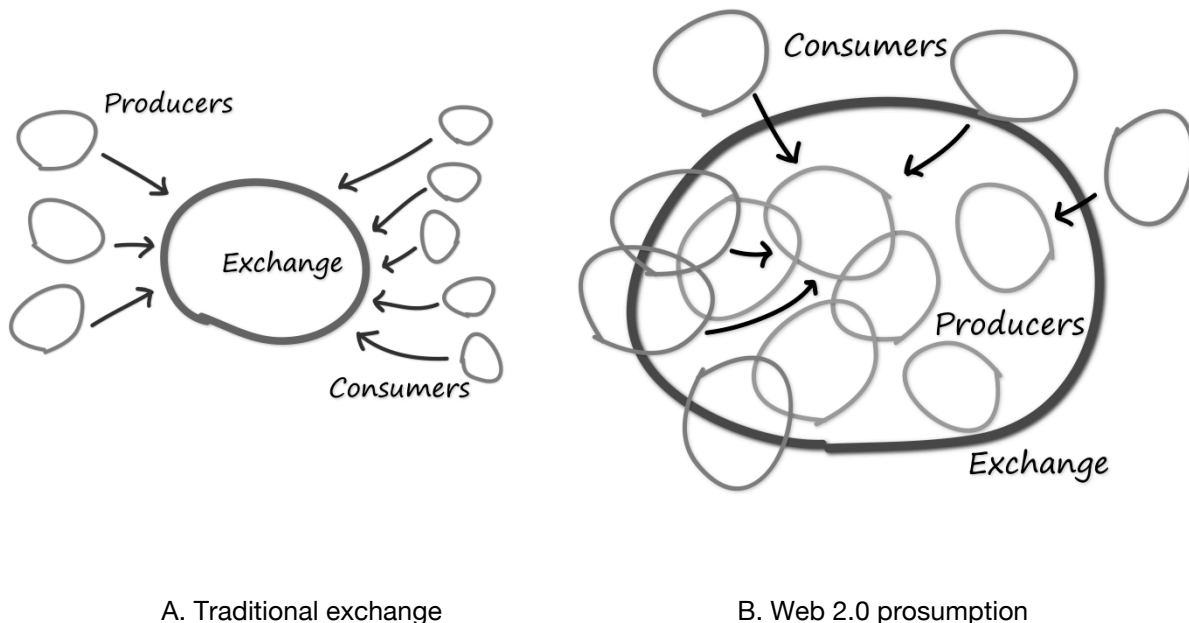


Figure 2. The traditional triptych of producers-exchange-consumers has been updated to the presumption model where consumers also contact with producers directly in global scale or/and become producers (Vafopoulos, 2011a).

Today, during a minute, online users send more than 204 million emails, make 6 million page views in Facebook, watch 1.3 million video clips on YouTube, listen to 61,000 hours of music on Pandora and spend approximately \$83,000 in Amazon⁵. In 2012, only Americans spent 74 billion minutes, or 20 percent of their time, on social networks (Nielsen/Incite's Social Media Report for 2012). This figure could be also interpreted as productivity cost of workplace interruptions that the research firm Basex puts at \$650 billion a year.

⁵ <http://abcnews.go.com/blogs/technology/2013/03/what-happens-in-1-minute-on-the-internet/>

These fundamental changes in preferences are supported by new types of consumption and production (e.g. Peer communities), new service sectors (e.g. Software as a Service) and the transformation of existing industries (e.g. mass media). The resulting reconfiguration in the triptych of production-exchange-consumption is based on a radical change in the fundamentals of the economy that the Web brings (Figure 2). Basically, the online ecosystem brings a major new source of increasing returns in the economy: more choices with less transaction costs in production and consumption.

This source of value arises from the orchestration of digital and network characteristics of Web goods and services. More choices in consumption range from larger variety of available goods, to online consumer reviews and ratings. This updated mode of connected consumption allows consumers to make more informed decisions and provides them with stronger incentives to take part in the production and exchange of mainly information-based goods. On the other hand, the provision of more choices with less transaction cost in consumption does not always come without costs. The leading native business model in the Web is the forced joint consumption of online information and contextual advertisements in massive scale. Also several cases of users' personal data abuse have been reported.

Consumption in the Web economy becomes more energetic and connected blurring the borders between production-consumption and (re-) brings in the fore the idea of prosumption. Moreover, the recent emergence of “social commerce” (Stephen and Toubia, 2010) as a consumer-driven online marketplace of personalized, individual-curated shops that are connected in a network, demonstrates the volatile boundaries among production, exchange and consumption in the Web.

Turning to the production side, many business operations went online and became less hierarchical, niche online markets and services have emerged and traditional industries revolutionized.

The change in user preferences, expectations and behaviour in our networked world is tightly related to the rise of Peer Production communities. Facebook, YouTube, Wikipedia, Twitter and LinkedIn are top in the list of the most popular websites and worth several billions of dollars.

In particular according to (M. Vafopoulos, 2011a):

“Peer Production is the creative process of user communities, which collaborate, mainly in the Web, to produce sharable goods. These communities enjoy open access to the means of production, share information about inputs and outputs and create pooled knowledge in order to increase the efficiency of future production. In Peer Production communities private information and preferences are revealed and aggregated without frictions, through explicit (e.g. voting, ranking, pricing) and implicit (e.g. tags, reputation) information sharing mechanisms. Because of the fact that information and preferences are public, transparent choice of inputs and outputs is an efficient coordination of rights assignment mechanism. Contrastingly, in traditional business, private hierarchical structures are designed to minimize coordination costs. Peer Production

communities could be more efficient than firms or markets if they can operate under less coordination costs in atomizing production. In this context, entrepreneurs have begun to exploit distributed economies of scale in Peer Production on industries with high coordination costs (e.g. social networking, freelancers markets) by providing production platforms.”

If we want to generalize, Peer Production pervades both the private and the public domain and the demand-supply dichotomy by introducing a the third mode of production, a third mode of governance, and a third mode of property (Bauwens, 2006).

In this context, a further investigation is needed on the potential ways of incorporating this new complex and dynamic reality in Official Statistics. IW4OS model offers a fertile ground for evaluating existing tools and methodologies and testing new state-of-the-art approaches.

3.2. Mapping current ICT statistics against the Internet as a data source

Top-level trade-offs

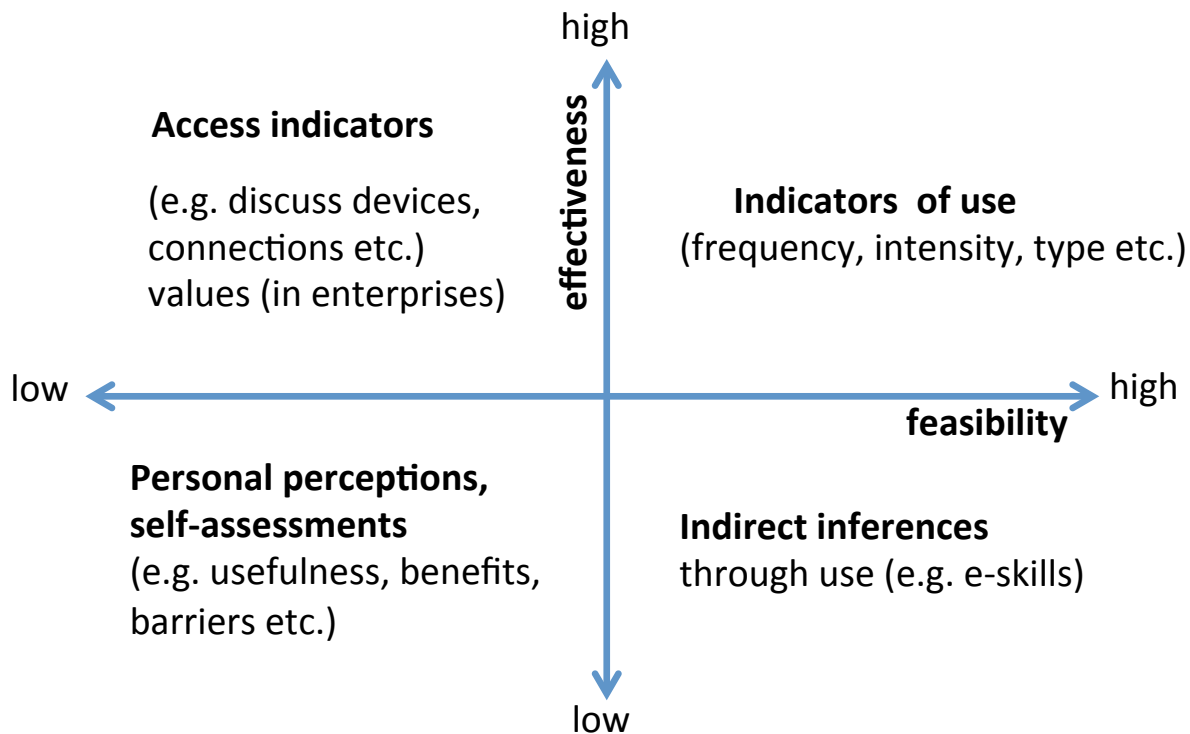


Figure 3. Feasibility and effectiveness of different categories of measurement.

As a rule of thumb, questions related to matters of access cannot be answered from the Internet. To the extent that their measurement is important, the availability of computers, desktop or

portable, mobile phones and other ICT devices cannot be known from the Internet. To some extent, this is an oxymoron and reminiscent of the digital divide: knocking at the door of the “haves”, you cannot find the “have-nots”. Generally, the Internet as a data source is ideally situated for the measurement of indicators of use. Lastly, the Internet is not meaningful for content concerning the views by individuals or businesses of their experiences or any subjective assessments and opinions.

The following schematic helps visualize such top-level trade-offs. It shows the degree of feasibility of different categories of measurement against their effectiveness. The latter is defined roughly, as a combination of the desirability of continuing to have such measures and/or their expected fitness for the uses intended vis-à-vis the traditional methods.

Indicators about individuals

The following table lists the variables included in the questionnaire of the 2013 round of the survey on ICT usage in households and by individuals. Comments are provided for their amenability to measurement by device monitoring software.

Table 1. 2013 Households ICT survey questionnaire and comments about amenability to measurement through device monitoring software.

HOUSEHOLDS QUESTIONNAIRE		Comments
Module A: Access to Information and Communication Technologies		
A1	Do you or anyone in your household have access to a computer at home?	The availability of computers or other peripheral ICTs, particularly as stand-alone not necessarily connected to the Internet, can only be answered by household members.
A2	Do you or anyone in your household have access to the Internet at home?	
A3	What types of Internet connection are used at home?	Possible to capture this information. The data will be superior to data collected from the existing questionnaire since the technicalities surrounding the type of connection is somewhat problematic for respondents.
A3a	broadband	
A3b	ISDN, dial-up or other narrowband	
A3c	Wired fixed (cable, optical fibre, Ethernet, PLC, etc.)	
A3d	Fixed wireless (satellite, public WiFi)	
A3e	mobile phone network (at least 3G, e.g. UMTS) via a handset	
A3f	mobile phone network (at least 3G, e.g. UMTS) via a card or USB key	
A3g	Dial-up access over normal telephone line or ISDN	
A3h	Mobile narrowband connection (less than 3G, e.g. 2G+GPRS, used by mobile phone or modem in laptop)	
A4	What are the reasons for not having access to the Internet at home?	
A4a	Have access to Internet elsewhere	Information about opinions can only be collected with a questionnaire.
A4b	Don't need Internet (because not useful, not interesting, etc.)	
A4c	Equipment costs too high	
A4d	Access costs too high (telephone, DSL subscription etc.)	
A4e	Lack of skills	
A4f	Privacy or security concerns	
A4g	Broadband Internet is not available in our area	
A4h	Other	
Module B: Use of computers		
B1	When did you last use a computer (at home, at work or any other place)?	These data cannot be obtained from the Internet. Even if it were possible to use federated data, inferring the usage of computers from the Internet is not necessarily optimal, if the intent of the indicator is to capture non-networked use. On the other hand, a good case can be made that such indicators were designed for a time past, when computers were differently understood than today – in that case, such indicators may not serve much purpose in the future.
B2	How often on average have you used a computer in the last 3 months?	
Module C: Use of the Internet		
C1	When did you last use the Internet	Information captured by this question can now become much more granular, in a way that was not possible through the questionnaire. Having access to real-time use information, there is no need to be restricted in knowing the respondent's last Internet use within the last 3 months or a year. Among other uses, analytical profiles of different user groups according to the frequency and intensity of use can also be constructed.
C2	On average how often did you use the Internet	Information captured by this question can now become much more granular, in a way that was not possible through the questionnaire. Having access to real-time use information, there is no need to be restricted in knowing the respondent's frequency as daily, weekly etc. Superior and very detailed information can be obtained, including the exact number of sessions, their duration, their distribution in the course of a day or week, any differences between week days and weekends, and many other angles that can be supported by the collected data on usage. Recall issues among respondents, and the burden imposed on them from the need of detailed answers will be avoided. Among other uses, analytical profiles of different user groups according to the frequency and intensity of use can also be constructed.
C3	Where have you used the Internet in the last 3 months (using a computer or any other means)?	These questions cannot be answered from the Internet.
C4	Do you use any of the following mobile devices to access the Internet away from home or work?	
C4a	Mobile phone (or smart phone)	
C4a1	via mobile phone network	
C4a2	via wireless network (e.g. WiFi)	
C4b	Portable computer (e.g. laptop, tablet)	
C4b1	via mobile phone network, using USB key or (SIM) card or mobile phone as modem	
C4b2	via wireless network (e.g. WiFi)	
C4c	Other devices	
C4d	I don't access the internet via any mobile device away from home or work	
C5	For which of the following activities did you use the Internet	Information captured by this question can now become much more granular, in a way that was not possible through the questionnaire. Generally, taxonomies on the types of use can be as long as their designers desire them, as there is no real end to the amount of detail. Practically, prioritization takes place and choices are made, constrained by scarce resources and concern for response burden. Question C5 contains 6 main categories, which in turn include 17 individual types. With an appropriate design, data on those and many additional types of user activities, not in the current question, can be obtained from the Internet. A clarification is in order: Internet-based use will not be able to capture "use for private purposes", as is currently asked in the questionnaire.
a	Sending / receiving e-mails	
b	Participating in social networks	
c	Reading online news sites / newspapers / news magazines	
d	Seeking health-related information	
e	Looking for information about education, training or course offers	
f	Finding information about goods or services	
g	Downloading software (other than games software)	
h	Posting opinions on civic or political issues via websites	
i	Taking part in on-line consultations or voting to define civic or political issues	
j	Doing an online course (in any subject)	
k	Consulting wikis to obtain knowledge on any subject	
l	Looking for a job or sending a job application	
m	Participating in professional networks	
n	Using services related to travel or travel related accommodation	
o	Selling of goods or services, e.g. via auctions	
p	Telephoning over the internet / video calls (via webcam) over the internet	
q	Internet Banking	

Table 2. 2013 Households ICT survey questionnaire and comments about amenability to measurement through device monitoring software (continued).

MODULE D: Use of e-government		
D1	Did you contact or interact with public authorities or public services over the Internet	Capturing user traffic statistics on the Internet can identify such websites, but more specificity and structure would have to be imposed – with the potential of not only obtaining comparable information but improving on its degree of detail and the eventual interpretability. For instance, the open-endedness of what constitutes the public sector can be improved – and it may not be the same across countries. Ministries of national governments can be specified and tracked, as can departments of municipal governments and universities. Moreover, depending on the methodology of the exercise (discussed in Section 10), the period of the last 12 months (used to avoid issues of seasonality) may be possible or may need to change to the period of monitoring, particularly if more than one collection cycles occur during one year.
a	Income tax declaration	
b	Downloading official forms	
c	Submitting completed forms	
D2	Did you use websites of public authorities or public services for any of the following	
a	Income tax declaration	
b	Claiming social security benefits	
c	Requesting personal documents (passport, ID card or driver's licence) or certificates	
d	Public libraries (availability of catalogues, search tools)	
e	Enrolment in higher education or university	
f	Notification of change of address	
D3	Have you experienced any of the following problems when using websites of public authorities or public services for private purposes in the last 12 months?	At the same time, the specific transactions asked are not likely to prove possible.
a	Technical failure of website	
b	Insufficient, unclear or outdated information	
c	Support was needed but not found (on-line or off-line)	
d	Other	
D4	Are you satisfied or dissatisfied with the following aspects public authorities or public services in the last 12 months?	
a	Ease of finding information	
b	Usefulness of the information available	
c	The information provided on the progress, follow-up of the request	
d	Ease of using services on the website	
D5	Did you contact public authorities or public services using methods other than websites for private purposes in the last 12 months?	These questions do not render themselves to Internet measurement.
a	yes, by telephone (excluding SMS)	
b	yes, by e-mail	
c	yes, in person, by visits	
d	yes, by other means (e.g. post, SMS, fax)	
e	no	
D6	What were the reasons for not submitting completed forms to public authorities' websites for private purposes in the last 12 months?	
MODULE E: Use of e-commerce		For e-Commerce, the information asked in all questions can generally be captured from the Internet. On the other hand, it may not be possible to differentiate between free and bought services/content consumed online.
E1	When did you last buy or order goods or services for private use over the Internet	The time period of the Internet-based data will coincide with the time period of the study.
E2	What types of goods or services did you buy or order over the Internet	This question conceptually fall under the previous discussion regarding the open-endedness of such taxonomies (food, medicine, clothes, hardware, software, hotels etc). Answers to those and transactions for many other products may be obtained – subject to additional back-end work. In other words, the compilation of data will not be automatic. Information will have to be mined not only for the sites visited but specific pages of such sites, and this must be compared against appropriate lists of businesses that must be created. It may well be that we end up with more specific product groupings than in the existing questionnaire, but that would not represent a loss!
a	Food or groceries	
b	Household goods (e.g. furniture, toys, etc)	
c	Medicine	
d	Films, music	
e	Books, magazines, newspapers (including e-books)...	
f	e-learning material	
g	Clothes, sports goods	
h	Video games software and -upgrades	
i	Other computer software and -upgrades	
j	Computer hardware	
k	Electronic equipment (incl. cameras)	
l	Telecommunication services	
m	Share purchases, insurance policies and other financial services	
n	Holiday accommodation (hotel etc.)...	
o	Other travel arrangements (transport tickets, carhire, etc.)	
p	Tickets for events	
q	Other	
E3	Were any of the following products that you bought or ordered over the Internet downloaded or accessed from websites rather than delivered by post etc	The same more or less applies to this question, although in this case categorization will be more obvious – and can be more specifically itemized than the existing question since films/movies can be separated from music, books from newspapers and the like.
a	Films, music	
b	(Electronic) books, magazines, newspapers, e -learning material	
c	Computer software (incl. computer and video games and software upgrades)	
E4	From whom did you buy or order goods or services for private purpose over the Internet	Answers to this question can also be obtained.
a	National sellers	
b	Sellers from other EU countries	
c	Sellers from the rest of the world...	
d	Country of origin of sellers is not known	

Table 3. 2013 Households ICT survey questionnaire and comments about amenability to measurement through device monitoring software (continued).

MODULE F: e-skills		
F1	Which of the following Internet related activities have you already carried out	Answers to all the 9 categories under this question can be obtained. As has been already explained, using search engines, sending e-mails, creating a Web page, making Internet telephone calls, uploading games and many more activities will be captured.
a	Using a search engine to find information	
b	Sending e-mails with attached files (documents, pictures, etc.)	
c	Posting messages to chatrooms, newsgroups or an online discussion forum	
d	Using the Internet to make telephone calls	
e	Using peer-to-peer file sharing for exchanging movies, music, etc.	
f	Creating a web page	
g	Uploading text, games, images, films or music to websites	
h	Modifying the security settings of internet browsers	
i	None of the above	
F2	Do you judge your current internet skills to be sufficient?	The self-assessment questions F2 and F3 cannot be answered. It is entirely possible, though, for a skills index to be created based on the profile of usage of individuals if desired.
a	To communicate with relatives, friends, colleagues over the internet	
b	To protect your personal data	
c	To protect your private computer from virus or other computer infection	
F3	Do you judge your current computer skills to be sufficient if you would need to take up a new job on the labour market or change your job within a year?	
MODULE G: Socio-demographic background characteristics		
G1	Age	These personal details can only be collected from the individual.
G2	Sex	
G3	Country of birth	
G4	Country of citizenship	
G5	Legal marital status	
G6	De facto marital status	
G7	Educational level	
G8	Employment situation	
G9	Occupation	
G10, G11	Region of residence (NUTS1, NUTS 2)	
G12	Geographical location	
G13	Degree of urbanisation	
G14	Number of members in the household	
G15	of which, number of children under 16	
G16	Household income	

Indicators about enterprises

The following table lists the variables included in the questionnaire of the 2013 round of the survey on ICT usage and e-commerce in enterprises. Comments are provided for their amenability to measurement by crawlers.

Table 4. 2013 Enterprise ICT survey questionnaire and comments about amenability to measurement from the enterprise's website with crawlers.

	ENTERPRISE QUESTIONNAIRE	Comments
	Module A: Use of computers and computer networks	
A1	Did your enterprise use computers?	These questions cannot be answered directly from the website.
A2	How many persons employed used computers at least once a week?	
A3	Did any persons employed have remote access to the enterprise's e-mail system, documents or applications (via fixed, mobile or wireless connection to the Internet)?	
	Module B: Access and use of the Internet	
B1	Did your enterprise have access to the Internet?	This question cannot be answered directly from the website. An enterprise can have a website, e.g. hosted by a third party, without itself having access to the Internet.
B2	Did your enterprise have the following types of external connection to the Internet?	
a	DSL connection	
b	Other fixed broadband Internet connection	
c	ISDN connection or dial-up access over normal telephone line	
d	Mobile broadband connection via a portable device using mobile telephone networks (so called 3G or 4G)	
d1	via portable computer	
d2	via other portable devices like Smartphone, PDA phone	
e	Other mobile connection	
B3	What was the maximum contracted download speed of the fastest Internet connection of your enterprise?	
B4	How many persons employed used computers with access to the World Wide Web at least once a week?	
	Mobile connection to the Internet for business use	
B5	Did any persons employed have portable devices provided by the enterprise, that allowed a mobile connection to the Internet for business use?	These questions cannot be answered directly from the website.
B6	How many persons employed had a portable device provided by the enterprise, that allowed a mobile connection to the Internet for business use?	
B6*	Estimate the percentage of the total number of persons employed which had a portable device provided by the enterprise, that allowed a mobile connection to the Internet for business use.	
	Use of a website or home page	
B7	Did your enterprise have a Website or Home Page	These questions can be answered directly from the website.
B8	Did the Website or Home Page have any of the following:	
B8a	Online ordering or reservation or booking, e.g. shopping cart	
B8b	A privacy policy statement, a privacy seal or certification related to website safety	
B8c	Product catalogues or price lists	
B8d	Order tracking available on line	
B8e	Possibility for visitors to customise or design the products	
B8f	Personalised content in the website for regular/repeated visitors	
B8g	Advertisement of open job positions or online job application	
	Use of the Internet in contact with public authorities	
B9	During 2012, did your enterprise use the Internet for interaction with public authorities to:	This can be obtained by an employee's computer connected to the Internet, and therefore server data, but not through the enterprise's website.
B9a	obtain information from public authorities' websites or home pages	
B9b	obtain forms from public authorities' websites or home page, e.g. tax declaration	
B9c	submit completed forms electronically, e.g. forms for customs or VAT declaration	
B9d	declare VAT completely electronically without the need for paper work (including electronic payment, if required)	These questions cannot be answered directly from the website.
B9e	declare social contributions completely electronically without the need for paper work (including electronic payment, if required)	
B10	During 2012, did your enterprise use the Internet for accessing tender documents and specifications in electronic procurement systems of public authorities	
B11	During 2012, did your enterprise use the Internet for offering goods or services in public authorities' electronic procurement systems (eTendering)	
B11a	in your own country	
B11b	in other EU countries	
	Use of Social Media	
B12	In January 2013, did your enterprise use any of the following social media	All this content can be captured from server data too but unlikely to be had from the website – unless the traffic went through there.
B12a	Social networks	
B12b	Enterprise's blog or microblogs	
B12c	Multimedia content sharing Websites	
B12d	Wiki based knowledge sharing tools	
B12e	Did not use any of the above or used them only for posting paid adverts	
B13	In January 2013, did your enterprise use social media to:	
B13a	Develop the enterprise's image or market products (e.g. advertising or launching products etc.	
B13b	Obtain or respond to customer opinions, reviews, questions	
B13c	Involve customers in development or innovation of goods or services	
B13d	Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non governmental organisations, etc.)	
B13e	Recruit employees	
B13f	Exchange views, opinions or knowledge within the enterprise	
B14	Did your enterprise have a formal policy for using social media?	It can be captured if posted on the website.
	MODULE C: Electronic Invoicing	
C1	In January 2013, did your enterprise send electronic invoices?	It is possible that these indicators can be had from the website – with the exception of the part of C1 that refers to e-invoices not suitable for automatic processing, which can in fact be sent by enterprises that do not have a website. From a technical standpoint, reference to "standard structure" implies the use of technologies whose implementation requires adherence to common standards by developers, and which typically refer to an underlying web infrastructure. However, it is also possible that such applications reside on other enterprise servers and may therefore not be retrievable from the website.
C1a	e-invoices in a standard structure suitable for automatic processing, e.g. EDI, UBL, XML	
C1b	Electronic invoices not suitable for automatic processing, e.g. emails, email attachment in PDF format	
C2	In January 2013, did your enterprise receive e-invoices in a standard structure suitable for automatic processing suppliers or customers	

Table 5. 2013 Enterprise ICT survey questionnaire and comments about amenability to measurement from the enterprise's website with crawlers (continued).

MODULE D: Automatic share of information within the enterprise		
D1	In January 2013, did your enterprise use an ERP software package	As ERP and CRM software are not typically stored on websites, this information cannot be had.
D2	In January 2013, did your enterprise use CRM software to manage:	
D2a	the collection, storing and making available information about customers to various business functions	
D2b	the analysis of information about customers for marketing purposes.	
MODULE E: e-Commerce		
e-Commerce purchases		
Web sales		
E1	During 2012, did your enterprise <i>receive</i> orders for goods or services placed via a website	This information can be obtained from the website. However, the comments in section C for the availability of such data on the web or other enterprise servers apply here too.
E2	Please state the value of the turnover resulting from orders received that were placed via a website (in monetary terms, excluding VAT)	This information cannot be obtained from the website.
E2*	Please indicate an estimate of the percentage of the total turnover resulting from orders received that were placed via a website	
E3	In 2012, did your enterprise <i>receive</i> orders placed via a website by customers located in the following geographic areas	This information can be obtained from the website. However, the comments in section C for the availability of such data on the web or other enterprise servers apply here too.
E3a	Own country	
E3b	Other EU countries	
E3c	Rest of the world	
E4	Please provide a percentage breakdown of the turnover from orders received that were placed via a website by type of customer	This information cannot be obtained from the website.
E5	Did any of the following obstacles limit or prevent your enterprise from selling via a website?	
E5a	The enterprise's goods or services were not suitable for web sales	
E5b	Problems in web sales related to logistics (shipping of goods or delivery of services)	
E5c	Problems in web sales related to payments	
E5d	Problems in web sales related to ICT security or data protection	
E5e	Problems in web sales related to the legal framework	
E5f	The cost of introducing web sales was, or would have been, too high compared to the benefits	
EDI-type sales		
E6	During 2012, did your enterprise <i>receive</i> orders for goods or services placed via EDI-type messages	This information cannot be obtained, so long as the EDI systems are proprietary and not on the Internet – and thus not on the website either.
E7	Please state the value of the turnover resulting from orders received that were placed via EDI-type messages (in monetary terms, excluding VAT)	
E7*	Please indicate an estimate of the percentage of the total turnover resulting from orders received that were placed via EDI-type messages	
E8	In 2012, did your enterprise <i>receive</i> orders placed via EDI-type messages by customers located in the following geographical areas	
E8a	Own country	
E8b	Other EU countries	
E8c	Rest of the world	
e-Commerce purchases		
E9	During 2012, did your enterprise send orders for goods or services via a website or EDI-type messages?	This information can be obtained from the website.
E10	During 2012, did your enterprise <i>place</i> orders for goods or services via a website	
E11	During 2012, did your enterprise <i>place</i> orders for goods or services via EDI-type messages	This information cannot be obtained from the website – unless EDI systems happen to be through the website.
E12	Please indicate the value of orders that were sent electronically in relation to the total purchases' value (in monetary terms, excluding VAT)	
E12*	Please state the value of the purchases resulted from orders placed electronically (in monetary terms, excluding VAT)	The information that does not refer to EDI-type messages can be obtained from the website. About EDI-type messages please see the comment to questions E11, E12.
E12**	Please provide an estimate of the percentage of the total purchases that resulted from orders placed electronically	
E13	In 2012, did your enterprise <i>place</i> orders via a website or EDI-type messages to suppliers located in the following geographic areas?	
E13a	Own country	
E13b	Other EU countries	This information cannot be expected to be obtained from the website.
E13c	Rest of the world	
Background information		
X1	Main economic activity of the enterprise	This information cannot be expected to be obtained from the website.
X2	Average number of persons employed	
X3	Total purchases of goods and services (in value terms, excluding VAT)	
X4	Total turnover (in value terms, excluding VAT)	

3.3. Considerations about the future

Facets of the Data era

It looks like that the discussions about the “social media era” surrender their position to the “data era”. For clarity, let us first provide our approach to some basic definitions about the different facets of data.

Inferred data

As inferred data could be considered the data that are collected through the “traditional” crawling and scrapping processes of unstructured and/or semi-structured of webpages. Usually, inferred data are stored in RDBMS and analyzed by “data mining” approaches.

Big data

Big data is popular term that has various definitions and views. According to Wikipedia big data is considered to be a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

Open data

Again from Wikipedia: “Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other "Open" movements such as open source, open content, and open access.”

Linked data

Linked Data enable the creation of better and massive services for data re-usage, driving existing infrastructure in its full potential. For government bodies, Linked Data adoption is focused on open, transparent, collaborative and more efficient governance. For enterprises, the core issue is about effective knowledge management and the implementation of new business models that initiate more energetic involvement and collaboration between producers and consumers (Vafopoulos, 2011a).

Linked Data is an attempt to simplify and spread horizontally throughout the Web the network externalities that exist in Web 3.0. Specifically, two sources of value have been identified for Linked Data technology. First, it enables users to build bidirectional and massively processable interconnections among online data and second, these data are critical enablers for existing infrastructure in the government and business spheres (Vafopoulos, 2011b).

Thus, Big data is more about scale, Open data about access and Linked data about the use of data (small or big, open or closed).

Federated open data

The present project is focused in «Federated open data» as have been defined by (Glasson et al., 2012). Federated open data is the counterpart (or supplement) of the so-called “open data” of governments. It refers to a shared sub-set of Big data from private sector entities, which will be “open” for use by NSOs.

As we indicated earlier, the proposed framework is meant to be applied to specific statistical indicators. The first implementation concerns data collection and analysis for ICT statistics.

The Internet of things

The Internet today provides access to continuously increasing amount of information universally, at any time and from any device. In the evolving Internet of Things (IoT) landscape, any device equipped with sensors is essentially an information warehouse, capable of collecting and transmitting real-time data originating from and interacting with the surrounding environment (people, places and things). These types of data are invaluable for official statistics since they contain information about the everyday life of individuals and communities and environment.

There is a growing need and interest in this regard by the Commission highlighted in its report “Internet of Things in 2020: A roadmap for the future”, where the key topics identified were the “smart living” and “mastered continuum of people, computers and things”. There are a growing number of innovative social and human-centric application areas, including social networking, smart metering, smart data collection, environmental models and so on. It is clear that with the growth of Web 2.0 and the social media, a wide sharing of information and know-how is held and such social networking activities can be properly harvested for the benefit of official statistics.

However, data streams generated from sensors are not readily usable for computation of indicators. Applications which are able to exploit IoT data streams and at the same time capture social pulse are necessary. Social pulse can be captured from the Web 2.0 new generation of applications and particularly Location-based Social Networks (LBSNs), which enable users to publish their actual “real time” geographic location online. Recent advances in mobile and sensor technologies provide new possibilities for supporting services and users supporting activities that can be distributed and incorporate different physical and environmental sensory data.

Therefore sensor devices and social interactions along with powerful applications can provide data for calculating various indicators related not only to ICT use and their social impact but also to other financial and social indicators related to either individuals or enterprises. Sensor data can be used for official statistics related to agriculture, forestry, environment, urban traffic and accidents, travels, health services, tourism, natural disasters, etc. Interaction of sensors with humans through applications converting sensor data to natural language expressions and social media is a potentially interesting perspective for validating the quality of data. In any case, this potential source of official statistics requires powerful technological infrastructure.

Dis-assembling and re-assembling

Right now we are in a transition and some crucial aspects appear blurred, inertia is still strong, we still do linear thinking, and are driven by cost efficiencies. All this is understandable, but we must develop a new mindset. It is perhaps honest to say and admit upfront that many of the

gains will be in the volume and quality of future outputs – with processes different from the ones we know (what is more difficult to come to terms with is that all those too will be evolving)! It is akin to the desktop computer replacing the typewriter many years ago – the gains did not come from replacing the typewriter with a word-processor alone, but from the fact that the desktop computer brought with it numerous new applications too transforming many processes.

The new situation therefore calls for new models. Starting with some questionnaire of the traditional type, which responded to policy, business and general societal needs and attempting to fill it through digital footprints is not the correct approach. Surely, in the interim at least, there will be questions that will render themselves to such substitution, and others that cannot. For the most part, these can be known and articulated and this activity has done so. What is more important, though, are the data and indicators that can be had, and which did not make the cut in the early wish list – either because they were not thought of at the time or because the designers of the instrument thought they cannot be had.

We need to realize at a deeper level that the “whole” questionnaire/approach we had - and went after filling it in its entirety with one process - will have to be broken down to pieces that fit the new reality. The classic “you can’t simplify the real world to fit your model” applies. These pieces then will feed not only the old “whole” but also many more different “wholes”. This is fundamentally different from the habitual, and perhaps more painstaking, but the sooner we start to develop a degree of comfort, the better.

The following schematic displays simply what all this means. We can draw lessons, albeit imperfect, from familiar examples of integrating activities, such as the System of National Accounts (SNA) satellite accounts. Under the habitual approach, data needs (typically advocated by policy makers) were met through a survey (entirely new or addition of modules to an existing one) – unless an administrative source existed (highly unlikely given that the new demands were associated with information concerning new and emerging phenomena). This essence of the approach connecting new needs to eventual statistical answers is depicted on the left-hand side of the schematic. Today, if we are to capitalize on the advent of ICT-based collection and/or federated data, more options for responding become available (right-hand side). Moreover, additional possibilities open up, which may turn orthodox processes upside-down. It is possible that in the process of tapping the new resources to answer a defined set of questions, answers to totally different questions can be fetched. As well, something much more intriguing is in the horizon: identifying the data that can be collected from where they exist (with the qualifications discussed earlier) and communicating such information to the demand side (e.g. policy makers), their thinking may be influenced in a way that they modify the questions asked. Thus, the interplay between data needs and responses elevates to another level. There may well be a link between perceived statistical needs and statistical outputs (not shown in the schematic).

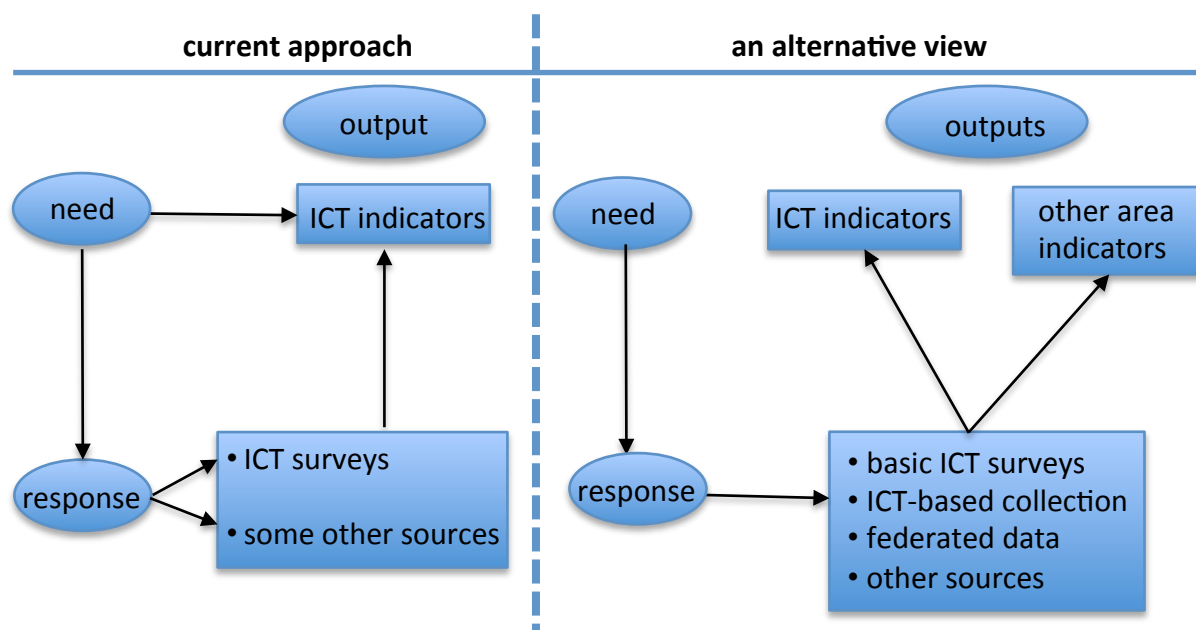


Figure 4. An alternative view of the way to produce ICT statistics.

At the same time, such transition will afford a prime opportunity to upgrade our core statistical infrastructure and prepare it for entrance to the new reality. A good example would be our registers – of businesses and of populations.

The detailed relevant results are presented in deliverables D1(a) and D1(b), which are available in sections 12.1 and 12.2, respectively, in the annex of this report.

4. Consultations with statistics authorities, business web sites and individual Internet users

4.1. Consultation with statistical authorities

Discussions about the feasibility of Internet-data based methods were held with four National Statistical Institutes (NSIs):

- Office for National Statistics (ONS), UK
- Hellenic Statistical Authority (ELSTAT), Greece
- ISTAT, Italy
- Central Bureau of Statistics (CBS), Netherlands

The discussions revolved around the experiences they might have had with such methods and around their opinions about these methods in general (irrespective of whether they have applied

such methods or not). The list of topics that was sent to the NSIs to prepare them for the discussion is shown in appendix 10.3 of deliverable D2 of the project.

The picture that emerges from these discussions is firstly one of “no objection” to the new methods. Excluding the attitude of ELSTAT, which probably is due to its lack of acquaintance with them, the other NSIs view the new methods favourably as production tools, in principle not different from the other methods they use. They experiment with them and assess them with the same procedures they assess the quality of production processes. They are concerned about the accuracy of their results but in most cases they find it satisfactory, while they recognise the gains in timeliness they offer.

The legal setting is not clear for any of the NSIs. It is not clear to them if the consent of individuals or enterprises whose data are collected or of the owners of the data is sufficient to make the methods “legal”. Some of the scraping experiments in fact have been conducted without the site owners being aware of the scraping.

Leaving legal feasibility aside, the new methods seem feasible in the context of the ESS. They should be discussed with ESS partners and be “promoted” by their exponents like any other production method. This, together with methodological support should go a long way in ensuring their adoption by the Member States.

4.2. Consultation with business web sites

A sample of 61 randomly selected websites was used in order to investigate, via a questionnaire, whether they are willing to accept and implement the proposed new method of data collection. We have prepared a questionnaire, which outlined the proposed method and indicators and posed five questions in order to collect their opinions about them. Out of the 61 selected websites that were contacted, 27 (44,3%) websites’ owners replied, 16 (26,2%), refused to take part and 18 (29,5%) never replied.

Of the 27 that did offered their responses, almost half would accept an automatic data collection system but they require some bilateral agreement before the do so. So a large part of websites (about half or more) will refuse cooperation or not reply at all and those that can potentially agree see themselves as partners and not just respondents and require bilateral cooperation agreements rather than self-imposed rules and commitments from the National Statistical Institute.

4.3. Consultation with individual Internet users

In this section we examined attitudes of individuals towards a system of data collection for statistical purposes from their day to day activity. Most of the respondents (38/40 i.e. 79%) did not have reservations and 10 (21%) provided some. Confidentiality was the main concern and most users stated anonymity as a condition for accepting software installation. Overall, we

found that most users want to cooperate and will do so if they are satisfied that their privacy and anonymity will be preserved and their use of their devices will not be affected in a substantial way. Incentives may help to further increase cooperation

The detailed relevant results are presented in chapter 3 of deliverable D2 of the project. This deliverable is available in section 12.3 in the annex of this report.

5. Feasibility of internet data – based ICT statistics

Two separate production processes, one web site-centric and the other user-centric have been examined:

- the production of statistics on the characteristics of business web sites, based on data collected with the help of crawlers or search engines that rely on earlier crawling from the said web sites.
- the production of statistics on the use of Internet by individuals, based on data collected with the help of monitoring software installed on the users' devices.

Technically they are both feasible. Software components are available in several forms and the software technologies needed for development from scratch are commonplace. The capacities needed for development and maintenance are quite easy to find in the job market even if not already available to the NSIs. The detailed relevant results are presented in deliverable D2 of the project. This deliverable is available in section 12.3 in the annex of this report.

Methodologically Feasibility

Statistics on the characteristics of business web sites. The envisaged statistics are very relevant for the measurement of the information society, since they express the sophistication of business web sites and their role in the activities of the enterprises owning them. Moreover, the use of crawlers for data collection automates their production and greatly reduces the time required for one production cycle. This leads to very timely statistics, available in very few months after the end of the reference period. Relevance and timeliness are the two great strengths of the approach.

The drawbacks of the approach are three. The reliance on keywords as proxies for the possession of the target characteristics by the web sites can cause serious bias in the statistics. Moreover, the use of crawlers for data collection may cause concerns to site owners and lead to large refusal rates and therefore unit non-response. Finally, linguistic differences between countries and varying expertise in the selection of keywords between countries may reduce the geographical comparability of the statistics.

A survey encompassing all possible characteristics of a business web site will suffer from reduced accuracy. The approach should be used only for carefully selected characteristics, which can be mapped, with an 1-to-1 mapping, to specific technologies rather than keywords. Only

then can accuracy improve to a point that the approach is appropriate for official statistics. This however requires further testing.

Production of statistics on the use of Internet by individuals. The envisaged statistics are very relevant for the measurement of the information society, since they describe, in very rich detail, the interactions of society with the Internet. The use of software allows the timely recording of activities with details that cannot be matched by traditional methods. Moreover, the processing of the data is very quick and very timely statistics can be available in very few months after the end of the reference period. Relevance, degree of detail and timeliness are the great strengths of the approach.

Non-response on the other hand is the major drawback of the approach. Monitoring software resembles spyware, which is clandestinely installed on devices and which, rightly, users have learnt to fear. Moreover, the recorded data are personal and most users do not want to share them with third parties.

The expected extent of non-response is so large that it makes the approach look impractical. Pilot studies however are not surveys run by NSIs. The latter generally have institutional credentials and legal backing to engage in data collection and should be trusted to protect the data they collect. With suitable legal arrangements to accommodate digital personal data it can be expected that the reluctance of the public to participate in surveys following this approach will decrease gradually.

Costs and Benefits

The main benefits of the Web-site centric approach are that it produces very relevant indicators in a very timely way. No monetary value can be put however on them.

Moreover, the benefits are also offset by the insufficient accuracy of the produced statistics. To our opinion the costs (especially validation effort) are too high for the obtained benefits. Unfortunately, lacking more detailed cost information, no more precise assessment can be made.

On the other hand a user centric approach can automatically obtain a large part of the information currently collected with questionnaires in the regular ICT survey and therefore it reduces response burden considerably. It can also collect information, which could not be easily collected with a questionnaire, e.g. the volumes of data received or transmitted by the individuals.

Furthermore, data are recorded with great precision because they do not depend on the individuals' recall and reporting of activities but are recorded digitally. This also enables their recording in very rich detail that cannot be matched by traditional methods: individual applications and web sites, exact recorded starting and finishing times and separate recording of activities running in parallel.

The time required for data collection is also reduced considerably due to the automation. Statistics can be available a lot faster than with traditional methods. However, it is not easy to put a monetary value on these benefits so as to juxtapose it with the costs.

We have given some indications or very rough approximations of the cost of the automated data collection. The only indication of the cost of the current ICT survey comes from the grants that Eurostat gave to national authorities. Anonymized data provided to the project team report the total data collection cost, for both the households and the enterprise surveys, over the EU in 2012 at almost 3900000 euros.

The new method can present considerable savings in data collection, if a solution that is not priced by user is adopted. On the other hand it has considerable setup costs and possibly costs for the provision of incentives. Based on the limited available data it seems that the new method is overall most costly than the current survey.

The **legal feasibility** of collecting and aggregating statistical data has been analysed based on its implications that relate both to Data Protection and Privacy regulations, and to areas of Intellectual Property Rights and particularly the sui generis Database right in the EU context.

The overall process seems to be compatible with relevant data protection and database right rules. The prior consent and permissions should comply with the abovementioned provisions. The compliance is a matter of properly drafted Terms of Service to which the end user and the companies may opt in, before the installation / operation of the data collection software to their devices or web pages. The examination of the Terms of Service by the independent Data Protection Authorities in the territories exposed to the project would also provide for an additional confirmation of the legal compatibility.

The detailed relevant results are presented in deliverable D2 of the project. This deliverable is available in section 12.3 in the annex of this report.

6. Pilot testing of specific internet data – based ICT indicators

Two separate pilots were implemented, one targeting individuals and the other the websites of enterprises. Each pilot is the subject of a separate section of this chapter.

6.1. Pilot survey of Internet usage by individuals

Statistical indicators produced

The sites that users may visit while online were grouped into approximately 50 categories by the makers of the software that was used in the pilot. The same categorisation has been used for all activities that a user may perform online.

Three indicators have been produced by the pilot survey for these types of activity:

1. Share of users that have engaged in each type of online activity
2. Percentage of time online that users devote on average to specific types of activities.
3. Amount of time that users devote on average per day to specific types of online activities.

Sampling frame

Due to difficulties in obtaining a proper random sample from ELSTAT the project team decided to resort to a non-random sample. It was felt that the actual selection of the sample, carried out in the same manner as it is done in the regular survey, does not offer any input to the testing of the automated data collection method. The novel features of the method are found in the way it measures data; they can be tested on all kinds of samples. The project team chose as sampling frame a panel of persons compiled by a Greek market research company for use in opinion surveys.

Recruitment of sample members

The panel comprises 1287 persons from the whole of Greece. Due to its small size and to the expected high rate of refusals to participate there was no random selection of sample members. The market research company considered the provision of a monetary incentive to users as paramount to soliciting their cooperation. The reward for each participating member was €30.00. Due to this cost, as well as the cost of the monitoring software it was decided to restrict the sample to 150 persons and devices. In order to attract more users we proposed to participants a “certificate of participation”, stating their involvement to the pilot. Although, it was clearly stated that this is not a typical certification, certain younger participants responded positively to this.

An information note was sent to all members of the panel informing them about the nature of the data collection, the anonymity of the data and the indicators that would be produced. Together with the note the members of the panel received a screening questionnaire that asked about the types and number of devices which they use to access the Internet. Three reminders were sent and in the end 145 persons accepted to participate.

The information that the users provided with the screening questionnaire was analysed by the team and one device per user (PC, Android smartphone or Android tablet) was selected for monitoring. In the end however, due to the difficulties in installing the software or due to second thought perhaps, we finally managed to enlist only 48 persons in the sample.

The comparison between the complete panel and the sample showed some issues. Women, young persons and students were over-represented in the sample, compared to the panel. The over-representation of the two latter categories is not surprising, given their greater familiarity

with software tools and online interaction between persons. Education levels and income classes are quite fairly represented in the sample. Finally, there is a large over-representation of central Macedonia (EL 12) probably due to the fact that the market research company is located in Salonika, in central Macedonia.

Software tool

The software selected for monitoring and recording the users' activities was the online parental controls service Qustodio⁶.

Implementation

The pilot took place in December 2013. The first 10 days were spent deploying the software to the sample members. The remaining days were spent on collecting usage data. During the course of the collection, users were sent an email questionnaire requesting some demographic data and also some Internet usage data. These data were combined with those collected by Qustodio.

The data collected by Qustodio were processed by a PHP script and were converted in a tabular format with one row per individual and date.

Conclusions

Although it has not been possible to replicate all activities that an NSI would undertake, the results of the pilot study of individuals have shown the potential of the automated recording of data.

The types of online activities of individuals can be discerned at great detail and therefore rich classifications can emerge for statistical use. Moreover, the classifications can change to fit evolving statistical needs. Even historical data can be converted easily to the new classifications.

The variations of usage time can be observed and reported to the desired degree of temporal detail. One can easily imagine charts showing the evolution of usage time or of the share of users engaging in a specific activity for any category of users recorded and over any period of time. Similarly the variation can be shown by day of the week (i.e. "average Monday", "average Tuesday", etc) or by hour of the day.

The data are also recorded with great accuracy since there is no intervention of the individuals' cognitive processes. Reduced recall of past activities, which is a common problem in questionnaire-based surveys, does not affect the measurements.

Moreover, the measurements are obtained with great speed, irrespective of the size of the sample. The initial set-up of the software can be implemented in parallel for all or almost all

⁶ www.qustodio.com

sample members. Subsequently the software operates independently on each device and therefore the procedure is easily scalable to larger samples.

The speed of data collection also allows the repetition of the survey more frequently than traditional surveys. A quarterly data collection is feasible.

In addition, the installation of the software to users' devices makes possible the retention of the selected sample as a panel, which will provide measurements for the accurate estimation of changes in Internet usage.

Finally, data can be combined and jointly analysed with data collected with regular questionnaires. In this report we have only utilised the demographic information from such questionnaires with the automatically collected data. Other data could have been used in exactly the same way.

On the other hand the method has several disadvantages. The most serious is the lack of trust from individuals towards the producer of statistics. The pilot study managed to acquire the consent of 3.7% of the online panel, which was approached by the company that had created the panel. This rate of cooperation is comparable with the rates reported by recent, similar studies⁷. In fact, the rate of 3.7% was achieved with the use of a small financial incentive. The possibility that such an incentive may have to be used should not be ruled out by NSIs.

The chosen software cannot work on devices with the iOS operating system, i.e. iPhone and iPad. This excludes a substantial share of the target population from the survey. Due to the design of iOS, this problem afflicts several tools that could be used for data collection. Care is therefore needed in the selection of the software tool; the development of bespoke solutions might be necessary.

An additional problem in the pilot study was the lack of transparency of the measurement process implemented by the tool. As shown earlier, it was not clear how usage time is defined by the makers of the software and why there were discrepancies between the reported durations of usage time (in minutes) and durations as shares of total usage time. An NSI must not accept such lack of transparency; it should have complete knowledge of what each measurement means. Time and resource constraints of the pilot study did not allow us to resolve this issue.

Finally, the software reports usage times per category of site; it does not report the times at which usage started and ended. To compute usage times for aggregates of categories, the producer of statistics must add the separate usage times. If the categories of sites have been used concurrently, which is very likely, the aggregated times will overestimate the true usage times.

⁷ 5.8% of the chosen sample according to 'Bouwman, H., Heerschap, N., de Reuver, M. (2012) Mobile handset study 2012. The Hague: Statistics Netherlands' (p.10); 3.8% of the sample according to 'European Commission (2012) Internet as a data source. Luxembourg: Publications Office of the European Union.' (p. 148).

This is another point that shows the need for complete knowledge and control of the measurement process by the NSI.

Overall, the use of activity monitoring software shows great promise as a data collection tool and the ESS should carry out additional investigations of the statistical methodology and practical arrangements needed for its incorporation in regular statistical production.

6.2. Pilot survey of the characteristics of the web sites of business enterprises

Statistical indicators produced

All indicators that have been produced in the pilot survey are of the sort “Percentage of enterprises whose website ...” and they refer to whether the site provides specific types of information, uses particular types of technologies or offers certain facilities to its users.

An enterprise’s website has been defined as the set of pages whose addresses start with the same single URL that characterizes the enterprise. For example, the website of Agilis SA is the set of pages whose addresses start with www.agilis-sa.gr.

The indicators measured in the pilot survey are the following:

1. Percentage of enterprises whose website provides a contact URL: the indicator refers to whether the site gives a contact URL among the contact information presented to users.
2. Percentage of enterprises whose website provides a contact email address.
3. Percentage of enterprises whose website provides a contact telephone number.
4. Percentage of enterprises whose website provides a contact postal address.
5. Percentage of enterprises whose website offer pages in the national language. The national language in the case of the pilot is Greek.
6. Percentage of enterprises whose website offer pages in English.
7. Percentage of enterprises whose website presents the date of its last update. The date does not need to be present on all pages. Presence in at least one page suffices.
8. Percentage of enterprises whose website presents the site’s privacy policy.
9. Percentage of enterprises whose websites provides user registration facility.
10. Percentage of enterprises whose website presents its site map to users.

11. Percentage of enterprises whose website uses web analytic tools. The indicator refers to the deployment in the website of tools that analyse the number, provenance and behaviour of visitors to it. Such tools need not be – and usually are not – visible to the visitors.
12. Percentage of enterprises whose website advertises open positions or provides forms for applying for a job online.
13. Percentage of enterprises whose website provides links to multimedia content.
14. Percentage of enterprises whose website provides links to social networks or blogs.
15. Percentage of enterprises whose website provides links to wikis and wiki-sharing tools.

Sampling procedure

Due to difficulties in obtaining a proper random sample from ELSTAT the project team decided to resort to a non-random sample. It was felt that the actual selection of the sample, carried out in the same manner as it is done in the regular survey, does not offer any input to the testing of the automated data collection method. The novel features of the method are found in the way it measures data; they can be tested on all kinds of samples.

Private business registers, offered at a price by private vendors in Greece, were too costly for the resources of the project and in the end we resorted to a convenience sample. It was drawn from a list of enterprises, which contains contact details of Greek enterprises that have received in the past European funding for research. The total list contains 1777 enterprises. A random sample of 281 enterprises was drawn from this list.

Software tool

The technique used for the automatic collection of data was web crawling. It amounts to visiting web addresses (URLs) and copying their content to a local repository for later processing. Web crawling is commonly used by Web search engines in order to facilitate indexing which is crucial for web searching.

We opted to using Google's Custom Search Engine (CSE), instead of any specific utility. It provides an interface to the user in order to specify a list of sites and a list of keywords to search for in these sites.

Implementation

The collection of the data relies on the use of keywords. Each of the indicators is viewed as resulting from answering "Yes" to a question asking whether the website has / provides / uses / offers the mentioned type of content or facility.

Instead of asking questions we specified a number of keywords relevant to each indicator. Appearance of even one of these in at least one page of a website was considered as a “Yes” to the corresponding fictional question. Therefore we only needed to provide suitable keywords and the addresses of the websites of the sample to the CSE; it would then search among the content that Google has already indexed.

The selection of suitable keywords was not an one-off operation. Initial “trial” sets of keywords were tried and their results were reviewed by human operators and cross-checked versus the findings of manual searches in the websites. Additional keywords and stems of keywords were then proposed and tried again.

The CSE returns a list of URLs (pages, within each website) where any of these keywords has been found. Therefore, if for example site www.agilis-sa.gr contains in four of its pages the keyword “telephone” and in three more (possibly overlapping) it contains the keyword “tel”, the results will list seven URLs with the keyword found in each one attached to them. Post-processing was therefore carried out with a text parser which grouped such findings into a single “hit” per indicator and website.

Conclusions

The automatic collection of data from the web sites of enterprises has merits but the results of the particular approach chosen in the pilot study are not very encouraging. Some of the positive features of this mode of data collection are similar to those of the monitoring software used on individuals.

After an initial set-up period, devoted to the specification of keywords and other site features to be detected, the collection of data is a lot faster than traditional survey data collection. It is also scalable to large sample sizes. This permits the implementation of data collection at higher frequencies and to larger samples than traditional surveys.

Furthermore, the data collection that relies on Google’s search infrastructure and indexing is non-intrusive. Google has already processed the data and the NSI is querying Google’s results and not the sites.

The speed, automation and possible non-intrusiveness of the approach mean that a panel sample of enterprises can be set-up by the producer of statistics. To move things a little further, even a ‘census’ of enterprise sites could be established over the long term, for indicators, which can be measured accurately enough. Financial costs and time requirements of such a census should of course also be taken into account in any decision-making.

The disadvantages of the specific data collection mode used in the pilot outweigh its merits. The most serious is that the data returned by the search engine contain many spurious findings while on the other hand several occurrences of the site characteristics in which we were interested went

un-noticed. The results suffer from lack of both sensitivity and specificity. This is a deficiency of keywords. There seems to exist a limit to how specific keywords can be to the targeted site features: most features are associated with terminology, which also applies to other, unrelated, issues.

An obvious improvement of the approach's sensitivity is to also include keywords in the national language of each country. A second direction for potential improvement is to download the HTML source code of web sites and extract keywords from it as well. This would permit detection of filenames (e.g. 'envelope.gif' for an icon showing a postal envelope and accompanying the display of postal contact information) or reserved words (e.g. 'mailto') indicative of features of the sites. This approach requires the use of additional crawlers besides Google's search engine.

Detection capabilities could possibly improve if linguistic analysis of a site's content identified directly the language it is written in, instead of relying on imprecise keywords. Moreover, key icons (e.g. the logos of Facebook or Twitter) could be detected with some kind of image analysis or image search.

Besides site features that are manifested through keywords that cannot be specific enough there are other features which are not connected to verbal aspects of the sites. For example, video thumbnails may be the links to Youtube videos, without any keywords. Furthermore, web analytics may be deployed on a site invisibly to its visitors. Such features require the utilisation of tools that detect technologies rather than keywords.

Based on the results of the pilot study it can be inferred that the developed methodology for collecting data from enterprise web sites does not produce statistics of high enough quality. A more extended appraisal of the method, which will encompass aspects of multilingualism, extraction of source code and detection of technologies, is needed for a more informed decision about its usefulness.

General conclusions from both pilot surveys

The two pilot surveys gave contrasting results. The one among individuals gave promising results despite its problems. Monitoring of activities online (or offline if required) can give very rich, detailed information, adaptable to changing statistical needs. The reluctance of users to be monitored is a major obstacle. Limits in processing and storage capacity can also emerge in large scale or long-term applications. With suitable sample design for the selection of individuals and devices it seems that statistical issues will not be serious.

On the other hand, the survey among enterprises gave inaccurate results while also missing information that could have been obtained with a questionnaire. The detection of site features

cannot rely only on keywords: linguistic analysis, image search and detection of technologies could be useful additions with considerable impact on the accuracy of results. The type of indicators that can be measured by visiting websites and analysing their content or technologies needs careful consideration and the tools to be used need careful tuning.

The detailed relevant results are presented in deliverable D3 of the project. This deliverable is available in section 12.4 in the annex of this report.

7. ‘Cookbook’ for internet data – based ICT statistics

The ‘cookbook’ is a guide for the application of Internet-data based methods for the production of official statistics. Its audience are the producers of official statistics. The guide borrows its structure and some of its content from Eurostat’s “Methodological manual for statistics on the Information Society”⁸. More specifically, for aspects of the production methods, which will be implemented in the same manner as in the current households and enterprises ICT surveys (e.g. sampling enterprises from the business register of the NSI) the guidelines were copied from the current manual. Even then however, minor changes were made in order to discuss possible difficulties that will be faced by the new methods. A considerable part of the cookbook however consists of original material drafted by the project team.

The cookbook’s structure is the following:

Introduction

Part 1 - Statistics on the use of Internet by individuals

- 1 Statistical product
 - 1.1 Statistical unit
 - 1.2 Target population
 - 1.3 Periodicity
 - 1.4 Observation variables
 - 1.5 Summary measures, aggregated variables, indicators and tabulation
 - 1.6 Explanatory notes
- 2 Production methodology
 - 2.1 Timetable – Survey period
 - 2.2 Frame population
 - 2.3 Sampling design
 - 2.3.1 Stratification
 - 2.3.2 Sample size
 - 2.3.3 Weighting – Grossing up methods
 - 2.4 Survey type
 - 2.4.1 Data collection method

⁸ Eurostat (2013) Methodological manual for statistics on the Information society, v. 3. Luxembourg: Eurostat.

-
- 2.4.2 Independent versus embedded survey
 - 2.4.3 Mandatory versus voluntary survey
 - 2.4.4 Coping with refusals of selected individuals to be included in the sample
 - 2.4.5 Quality control systems
 - 2.5 Data processing
 - 2.5.1 Data validation
 - 2.5.2 Non-response treatment
 - 2.5.3 Unit non-response
 - 2.5.4 Item non-response
 - 2.6 Data analysis
 - 2.6.1 Post-processing
 - 2.6.2 Computation of indicators
 - 2.6.3 Estimation of the accuracy of the indicators
 - 2.7 Confidentiality and privacy issues
 - 3 Annexes
 - 3.1 Software tools
 - 3.2 Model questionnaire
 - 3.3 Transmission format
- Part 2 - Statistics on the facilities of business websites
- 1 Statistical product
 - 1.1 Statistical unit
 - 1.2 Target population
 - 1.3 Periodicity
 - 1.4 Observation variables
 - 1.5 Summary measures, aggregated variables, indicators and tabulation
 - 1.6 Explanatory notes
 - 2 Production methodology
 - 2.1 Timetable – Survey period
 - 2.2 Frame population
 - 2.2.1 Updating the Business Register with website information
 - 2.3 Sampling design
 - 2.3.1 Stratification
 - 2.3.2 Sample size
 - 2.3.3 Weighting – Grossing up methods
 - 2.4 Survey type
 - 2.4.1 Data collection method
 - 2.4.2 Independent versus embedded survey
 - 2.4.3 Mandatory survey versus voluntary survey
 - 2.4.4 Contact person of the survey

- 2.4.5 Coping with refusals of selected enterprises to be included in the sample
 - 2.4.6 Quality control systems
 - 2.5 Data processing
 - 2.5.1 Misclassification treatment
 - 2.5.2 Non-response treatment
 - 2.5.3 Unit non-response
 - 2.6 Data analysis
 - 2.6.1 Post-processing
 - 2.6.2 Computation of indicators
 - 2.6.3 Estimation of the accuracy of the indicators
 - 2.7 Confidentiality and privacy issues
- 3 Annexes
 - 3.1 Software tools
 - 3.1.1 Web crawlers
 - 3.1.2 Google's Custom Search Engine
 - 3.2 Example of mapping between target functionalities and keywords
 - 3.3 Transmission format

The cookbook is deliverable D6 of the project. This deliverable is available in section 12.5 in the annex of this report.

8. Feasibility of big data as a source for the production of official statistics

The potential of big data as a source of official statistics was examined. Of particular interest were the so-called ‘federated open data’ which are (big) data from business or the public sector, generally not accessible by the public, but shared in an agreed and defined way with the producers of official statistics. Five specific ‘use cases’ were examined, all being specific data repositories, most of them currently closed or partly open only, which could possibly be shared with producers of official statistics. Already open big data were also examined:

- Vessel movement data from the Automatic Identification System (AIS)
- Real estate classified advertisements
- Social media message data
- Credit card transaction data (Visa Europe)
- Government financial transparency portal data

Information about the repositories was gathered from metadata and reports that they disseminate, from direct investigation of the data, where possible and from direct communication with their owners, again where possible. The different dimensions of the analysis were:

- to which domains of official statistics are the data of the repository related?
- what current or new statistics in these domains can be produced from the data of the repository?
- what is the feasibility of producing these statistics based on the repository?
- which modes of access are available?
- what are the conditions for opening the repository to a producer of official statistics?

8.1. Vessel movement data from the Automatic Identification System (AIS)

There is a high potential in using AIS data in the production of current statistics:

- Number of vessels, by size and type of vessel
- Gross tonnage of vessels, by size and type of vessel

- Emissions from maritime transport activity sector (currently not compiled by Eurostat but their compilation is under investigation)
- Gross weight of goods

A potential data source for obtaining AIS data is MarineTraffic⁹. Although some data about vessels' characteristics may be missing or may not be readily available, these can either be estimated or obtained from an international database on vessel characteristics.

It is, however, possible to derive statistics on the number of ships almost in a straightforward and simple way from data that can be made available from MarineTraffic. This is possibly the only indicator that could replace official statistics in the very near future.

8.2. Real estate classified advertisements

There is a high potential in using Internet advertisement in the production of current statistics on the housing price index and Purchasing Power Parities (PPPs) related to rental and owner occupied housing. Moreover, there is some potential to using Internet advertisement in production of the owner occupied housing sub index of the Harmonised Index of Consumer Prices (HICP) although there are differences in concepts.

It is unlikely that data from Internet advertisements can replace the rent surveys for the HICP but they can provide helpful new indices and facilitate the survey itself.

8.3. Social media message data

There are a lot of benefits from using social media in the production of subjective indicators, which are used in the current statistics.

It is worth noting that Twitter and Facebook are two potential fascinating sources of sentiment information, however it is important to highlight that those sentiments cannot replace the existing official statistics and its indicators.

The measures of sentiments and their scoring can be used complementary to official statistics and provide us with useful trends over time as well as with comparisons among the different European countries.

⁹ www.marinetraffic.com

8.4. Credit card transaction data (Visa Europe)

There are a lot of benefits from using Visa's data in the production of consumption expenditure statistics. Currently, the Household Budget Survey (HBS), which produces similar data, is carried out at an informal basis every five years.

Visa Europe already compiles an Index, named "EU Consumer Spending Barometer"¹⁰ using real-time card transaction data. Its aim is to provide a robust indicator of total consumer expenditure at a European level.

It is worthwhile using Visa as a source, in a complementary way, for the production of flash estimates about the structure and amount of consumption expenditure. However, it is important to highlight that an index similar to Visa's Barometer, cannot replace the existing official statistics and its indicators.

Although, such a barometer can be used complementary to official statistics, it can only provide a robust indication of real consumer spending trends over time and among the different EU countries.

8.5. Government financial transparency portal data

The Greek government's transparency portal, called "di@vgeia" (the Greek word for transparency) was examined.

A huge amount of data on public expenditure is available through this portal. Main conclusions from analysis of its content and availability include:

- Data can be retrieved and processed for statistical purposes as it is publicly available and contains fields that can be linked to statistical classifications.
- There are several issues affecting data quality, primarily having to do with data entry errors and shortcomings in the current software that was prepared as a pilot. Most of them are expected to be solved with a new version currently under development that is expected to be released on September 2014.
- There are important impediments in terms of coverage; only expenses that require decisions are included. Therefore the source can't become a single source for all government finance data but it can be used as a supplementary source and in that way to:

¹⁰ http://www.visaeurope.com/en/newsroom/all_reports/european.aspx

- Reduce the burden to public administration entities by requiring them to report to the NSI only data that has not being published in “di@vgeia”
- Substantially improve timeliness.
- “di@vgeia” can serve as a primary source for statistics in certain areas where coverage is complete or near complete (e.g. public procurement, R&D spending).

The detailed relevant results are presented in deliverable D4 of the project. This deliverable is available in section 12.6 in the annex of this report.

9. Outline of procedure for the accreditation, by producers of official statistics, of big data sources as input data for official statistics

In this activity we proposed a procedure that NSIs pondering whether to use big data sources as input in the production of official statistics could employ to accredit such sources.

The main issue with big data sources, irrespective of whether these contain transactions, position data, etc., is that they are not compiled for statistical purposes. This is not a new situation for NSIs. In order to reduce cost and burden they routinely use administrative data. In general, data used by NSIs that are not collected for statistical purposes are called secondary data. Our work was based on the analysis of the available recent literature on topics such as quality of statistics in general and quality of administrative data sources in particular.

At first we proposed five **foundational principles**:

Principle 1. Accreditation procedures must be fully compliant with well-established principles of quality frameworks that guide the world of official statistics, and consistent with quality assurance practices embedded deeply in the work of NSIs.

Principle 2. Any accreditation procedure must be flexible in a way that does not unduly prejudice or rule out new opportunities without serious examination.

Principle 3. An accreditation procedure should include sequential decision-making based on a pragmatic step-wise approach, so that we spot early on new data sources that won't work, while we always invest in new sources that will work.

Principle 4. The accreditation procedure must contain an empirical assessment with real data, and it must be carried out by NSIs directly. It cannot be delegated to filling out questionnaires by the source owners.

Principle 5. A systematic accreditation procedure must assess the quality of the statistical outputs, the quality of the statistical inputs (including the source and metadata), as well as the quality of the statistical processes involved.

The actual **accreditation procedure** evolves in a step-wise fashion. It consists of *five stages* with gradual assessments involving indicators measured through scales and hard data, which in turn lead to recommendations associated with *six decision points*:

Stage 1: Initial examination of source, data and metadata. In order for an NSI to even contemplate acquiring and using an external data source some knowledge of it, or at least exposure to it, is surely a necessary condition. At this stage, an early assessment of the data, the metadata and the source is carried out. Anything that can be gauged from the outside or through limited and rather unofficial interaction with the working level at the source organisation should be collected, shared internally, and examined. Such material can come from the media, Web sites, releases, publications or articles and should cover the *raison d'être* of the organisation behind the source and as many aspects of content, data and metadata as possible.

Decision point 1: a Yes/No answer is needed to the question: “Is this data source potentially useful and for what”? This will lead to a recommendation to proceed to the next stage or not.

Stage 2: Acquisition of data and assessment. This stage entails negotiations with the source with a view to acquire a set of files or file extractions adequate for rigorous testing. The primary objective is to clarify whether the source is willing and able to deliver files or extractions at the record level, as well as keep open a communication channel during the testing process. A number of issues must be discussed in a professional manner with the data source, albeit not with the burden of formalizing a legal agreement yet (e.g. MOU) - which is more demanding. These include specifications of files or file extractions, time and method of transmission, as many metadata as possible, and any particular conditions that must be known.

In the process, we can update the results of Stage 1 with more accurate information that becomes available. This is not a repetition of Stage 1. It adds the revised results of that stage to those of stage 2.

Decision point 2: It is decided whether the amount of data that can be obtained are similar to what would be obtained from a survey. This will lead to a recommendation to proceed to the next stage or not.

Stage 3: Forensic investigation. This represents a critical step and requires a fair amount of work by the NSI. It can sub-divided in four distinct phases: i) producing a clean microdata file (halfway through which we meet a decision point); ii) using the file to produce and analyse aggregate statistics iii) producing pilot new outputs or using the file in the production of existing outputs, and; iv) assessing the capacity of the existing statistical tools to handle the new data.

Decision point 3: At the end of phase (i) the quality of the microdata file is assessed. If it is considered as too low it may be recommended to not proceed further.

Decision point 4: An overall assessment of the strengths and the weaknesses of the new data. Recommendation of whether to proceed to the next stage or not.

Stage 4: NSI decision. This stage is dedicated to the assessments necessary for a corporate decision to be made based on as much information and knowledge as possible. It can sub-divided in four distinct phases: i) an itemisation of the exact uses of the new data and their impacts; ii) a top-level cost-benefit analysis, which focuses on the financial picture; iii) assessment of the risks that need to be undertaken and managed by the NSI, iv) assessment of the feasibility of incorporating the new source into the gamut of the NSI's statistical operations from a legislative and socio-political point of view

Decision point 5: A corporate decision of whether to proceed to the next stage or not.

Stage 5: Formal agreement with source. This final stage involves high-level negotiations with the source as an institution to secure cooperation and arrive at a formal and comprehensive agreement.

Decision point 6: Based on the outcome of the negotiations a decision is taken of whether to start using the source.

The detailed relevant results are presented in deliverable D5 of the project. This deliverable is available in section 12.7 in the annex of this report.

10. Conclusions

The project has taken two different views of the potential for producing statistics based on data from the Internet.

The first view is quite narrow in scope: it focuses on Information Society statistics only and on just two possible Internet data-based methods. The first method collects data on the usage of the Internet by individuals and the data collection tool is monitoring software installed on their devices (computers, smartphones, tablets). The second method collects data on the characteristics of business web sites and the data collection tool is a crawler analysing the sites' content for identification of indicative keywords.

Both approaches were analysed at a theoretical level first. They are in line with the current proliferation of data in the Internet and with the necessity of using them for statistical production (at least not ignoring them without having examined them first).

Moreover, the analysis of their feasibility demonstrated that they can produce very relevant statistics, with rich detail and in a much more timely manner than the current ICT surveys can

manage. Their accommodation in the legal context of privacy and personal and corporate data protection requires attention from NSIs but it is feasible. In principle the new methods do not differ, from the legal point of view, from questionnaire-based collections of the same data.

The methods present problems too. They can lead to high refusal rates, especially in the case of individuals. The site-centric method tested in the pilot surveys suffers from problems of accuracy too. The keywords are neither specific enough, i.e. they appear even when the supposed site characteristic is not present, nor sensitive enough, i.e. the characteristics may be evident even when the keywords are absent from a site's content. Moreover, both methods are quite costly. The user-centric one appears more costly than the current ICT surveys but at least it can reduce considerably the response burden and processing time. The site-centric method's costs cannot be offset by the benefits it brings.

The project also produced a "cookbook", i.e. a guide, addressed to NSIs for the implementation of these methods.

The current state of data production and the expected increased rate of data generation shows however, that even more automated methods should be examined in the future. These are methods like identifying the digital footprints of individuals, culling data freely available in the Internet or obtaining data from proprietary servers of private or public organisations. This was the topic of the second, much broader view of the project: the potential of producing official statistics, about any domain imaginable, based on big data repositories.

Five specific use cases were examined, providing data relevant to diverse domains such as: transport, environment, consumer sentiment, government finances, housing prices, consumer expenditure. In all cases it has been demonstrated that large amounts of relevant data are available, at different degrees of openness. The two extremes, among these five cases, were the completely open government expense data from Greece and the "completely closed" VISA transaction data of VISA Europe.

These data can produce, on their own or in combination with statistical data, existing statistical indicators (i.e. they can replace them) or new ones.

The presence of these potential data sources means that NSIs are "suddenly" confronted by a pool of sources much wider than the current one. In order to be able to shift through them and identify those suitable for statistical production the project has proposed an outline of an accreditation procedure.

11. References

Bauwens, M. (2006). The political economy of peer production. *Post-autistic economics review*, 37.

- Berners-Lee, T. (2006). Welcome to the Semantic Web. The Economist - The World in 2007. Retrieved from <http://www.neurophenomics.info/docs/semanticweb.pdf>
- Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1. doi:10.4018/jswis.2009081901
- Capadisli, S., Auer, S., Ngomo, A. (2013). Linked SDMX Data. semantic-web-journal.net. Retrieved from <http://www.semantic-web-journal.net/system/files/swj454.pdf>
- Cyganiak, R., Reynolds, D., & Tennison, J. (2012). The rdf data cube vocabulary. Retrieved July 1, 2013, from <http://publishing-statistical-data.googlecode.com/svn/trunk/specs/>
- Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., & Khan, A. (2012). WHAT DOES “BIG DATA” MEAN FOR OFFICIAL STATISTICS? Retrieved from <http://goo.gl/LfGIM>
- Stephen, A. T., Toubia, O. (2010). Deriving Value from Social Commerce Networks. Journal of Marketing Research, 47(2), 215–228.
- Tapscott, D., Williams, A. D. (2008). Wikinomics: How mass collaboration changes everything. Portfolio Trade.
- Vafopoulos, M. (2011a). The Web economy: goods, users, models and policies. Foundations and Trends® in Web Science, 3(1-2), 1–136. doi:<http://dx.doi.org/10.1561/18000000015>
- Vafopoulos, M. (2011b). A Framework for Linked Data Business Models. *15th Panhellenic Conference on Informatics (PCI)* (pp. 95–99).

12. Annex – Technical deliverables of the project

12.1. D1 - Definition of Internet data-based indicators part I

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of
information society and other statistics’

D1.a Definition of Internet data-based indicators

Part I

November 2013

Document Service Data

Type of Document	Deliverable		
Reference:	D1.a – Definition of Internet data-based indicators (Part I)		
Version:	3	Status:	Draft
Created by:	George Sciadas, Lefteris Angelis, Michalis Vafopoulos, Dimitris Kalogeras	Date:	6/11/2013
Distribution:	European Commission – Eurostat/G6, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	16/7/2013	Initial release
2	23/9/2013	Modifications following Eurostat's comments at the 2 nd progress meeting
3	6/11/2013	Change of comment about q. A4 in page 26.

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

1. Introduction	4
1.1. Proliferation of data in the Web	5
1.1.1. Facets of the Data era	6
2. IW4OS: A novel conceptual framework	7
2.1. Interaction.....	7
2.2. Instantaneousness, Information Overload, Informality & Irregularity	7
3. Comparing IW4OS with existing models	9
3.1. The Web 2.0 economy in a nutshell	10
3.2. The IaD model	12
4. Reflections of the conceptual framework to ICT statistics	13
4.1. ICT Usage.....	13
4.1.1. ICT products	14
4.1.2. ICT infrastructure	14
4.1.3. ICT supply	15
4.2. Content and media	16
5. Weaving the Web economy: goods and users.....	16
6. The Crossroads of Information Society Policies and Indicators	20
6.1. A natural evolution	20
7. Mapping current ICT statistics against the Internet as a data source	23
7.1. Top-level trade-offs	23
7.2. Households/Individuals	25
7.3. Enterprises	28
8. Additional Information Society Indicators and Federated Data	30
8.1. Federated data	31
9. Specification of ICT Indicators from the Internet as data source	33
9.1. Individuals	34
9.2. Enterprises	39
9.3. Attributes of indicators.....	44
10. Methodology and Related Matters.....	46
10.1. Individuals.....	46
10.2. Enterprises	48
10.3. The pilot	49

11. Software Tools which can be used for data collection.....	50
12. References	54
13. Annex.....	56

1. Introduction

Coordinating the collection efforts of National Statistical Offices under a Framework Regulation, Eurostat compiles and disseminates a variety of Information and Communication Technologies (ICT) statistics widely used to monitor the progression of European countries to Information Societies. The main collection approach consists of two questionnaires, one for households/individuals and another for enterprises¹. These instruments were designed in such a way as to provide comparability across countries, as well as across non-European, mainly OECD, countries. At the time of their design, and up to now, the method of collection involves traditional surveys of people and businesses carried out by member states.

With all the recent efforts to tap the increasing trail of digital footprints left behind from the use of ICTs and related transactions, new possibilities open up. While these can lead to significant gains, including timeliness, quality and efficiency, it is still not clear how exactly to exploit them in practice.

Recently, the Directorate-General of the European Commission for Communications Networks, Content and Technology (DG Connect) commissioned a study on how such an evolution can happen. This research project proposed how to use the Internet as a Data source (IaD) to complement or substitute traditional statistical sources. It examined the pros and cons of different Internet-based methods, concluding that the three basic types are:

- User centric measurements that capture changes in behaviour at the client (PC, smartphone) of an individual user;
- Network-centric measurements that focus on measuring properties of the underlying network;
- Site-centric measurements that obtain data from webserver².

In general, network-centric methods are tough to acquire; as well, they were found to be problematic in terms of social acceptance. User-centric and site-centric methods, though, hold much more immediate promise.

This study significantly advanced our collective thinking but now a more specific focus is needed. This project aims to advance the ongoing efforts by targeting the indicators in the current surveys as well as additional indicators not feasible or cost-effective with the survey-based approach. Starting with a conceptual framework, it examines in detail the ICT indicators currently measured, maps them against the Internet-as-a-data-source methods, identifies indicator lists conducive to such measurements, and proposes appropriate methodologies and pilot tests.

¹ European Union survey on ICT usage in households and by individuals 2013, Eurostat Model Questionnaire (version 3.4), and Community Survey on ICT usage and e-commerce in enterprises, 2013, Model Questionnaire version 1.1

² Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering, SMART 2010/030, 2012(p. iv).

Because specific instruments are designed for specific purposes, and contemplating a move to methods of collection that mine digital footprints cannot assume an automatic and wholesale migration of the existing content from the old mode to the new without the need for re-arrangements, the study also discusses important contextual information within which such activities will happen.

1.1. Proliferation of data in the Web

The notion of data, during the last five years, gained huge fashionableness outside the strict borders of statistical methods. Businessmen, technologists and politicians attach their favorite adjective (e.g. open, big, linked) to it in order to describe their point of view for the future.

The European Commission's Open Data Strategy as has been expressed by Commission Vice President Neelie Kroes focus on three reasons why open data is fundamental: promoting the development of new businesses; promoting government transparency, and increased evidence-based policy making³. Lately, Eurostat has started to explore the potential of Linked⁴ and Big data⁵. But, for the time being, existing open data are characterized by *low availability* and *(re)usability*.

Contrastingly to physical and life sciences, where massive amounts of open data revolutionized fields like biology and physics, this is not happening for economic and social research (Lazer et al., 2009). Yet the available data for research are just a tiny fraction of the collected data from search engines, mass merchants, social networks and others in the Web.

The exclusive exploitation of behavioural data in the Web is an issue of primary importance with scientific, economic and social aspects. First, it limits academic research inside the “walled gardens” of companies, excluding open scientific research and official statistics. Second, companies that hold data and afford to analyze them have built comparative advantages against (potential) competitors or simply they are selling them for high profit. Finally, privacy and security risks (e.g. personal data leaks, almost-full profiling practices) create negative externalities in the personal and social level, which are not compensated (Vafopoulos, 2011a).

In addition, most of the available open data are either unstructured or semi-structured and not in machine-processable formats. In this context, the Linked Data initiative promotes the publication of structured and interlinked data in the Web.

³ For details refer to Q&A press release of 12th December 2011.

⁴ <http://eurostat.linked-statistics.org/>

⁵ <http://www.cros-portal.eu/content/big-data>

1.1.1. Facets of the Data era

It looks like that the discussions about the “social media era” surrender their position to the “data era”. For clarity, let us first provide our approach to some basic definitions about the different facets of data.

1.1.1.1. *Inferred data*

As inferred data could be considered the data that are collected through the “traditional” crawling and scrapping processes of unstructured and/or semi-structured webpages. Usually, inferred data are stored in Relational Database Management Systems (RDBMS) and analyzed by “data mining” approaches.

1.1.1.2. *Big data*

Big data is popular term that has various definitions and views. According to Wikipedia big data is considered to be a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

1.1.1.3. *Open data*

Again from Wikipedia: “Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other "Open" movements such as open source, open content, and open access.”

1.1.1.4. *Linked data*

Linked Data enable the creation of better and massive services for data re-usage, driving existing infrastructure in its full potential. For government bodies, Linked Data adoption is focused on open, transparent, collaborative and more efficient governance. For enterprises, the core issue is about effective knowledge management and the implementation of new business models that initiate more energetic involvement and collaboration between producers and consumers (Vafopoulos, 2011a).

Linked Data is an attempt to simplify and spread horizontally throughout the Web the network externalities that exist in Web 3.0. Specifically, two sources of value have been identified for Linked Data technology. First, it enables users to build bidirectional and massively processable interconnections among online data and second, these data are critical enablers for existing infrastructure in the government and business spheres (Vafopoulos, 2011b).

Thus, Big data is more about *scale*, Open data about *access* and Linked data about the *use* of data (small or big, open or closed).

1.1.1.5. Federated open data

The present project is focused on «Federated open data» as have been defined by (Glasson et al., 2012). Federated open data is the counterpart (or supplement) of the so-called “open data” of governments. It refers to a shared sub-set of Big data from private sector entities, which will be “open” for use by National Statistical Institutes (NSIs).

2. IW4OS: A novel conceptual framework

The new sources and forms of data in the Web are raising imperative questions to Official Statistics. The envelope question is which methods should be changed or even introduced to let Official Statistics retain their character, but at the same time exploit the emerging potential of online contexts?

Before starting to form specific proposals and engineer tools for new data sources and indices, a coherent common mindset should be introduced. The proposed conceptual framework for Internet and Web as data sources should facilitate the orchestration of their main characteristics with the approach of Official Statistics (the Internet and Web for Official Statistics framework-IW4OS – is presented in Figure 1).

2.1. Interaction

At the current Web 2.0 era, users are the protagonists of the online ecosystem because they can easily edit, interconnect, aggregate and comment online content as never before. Most of these opportunities can also be engineered in the personal level. The traditional triptych of producers-exchange-consumers has been replaced by the *prosumption* model where consumers contact producers directly or can act, at the same time, as producers. Web 2.0 enables interaction and crowdsourcing through openness, peering, sharing and acting globally (Tapscott & Williams, 2008).

These new modes of human interaction and production could be incorporated in providing more accessible and relevant Official Statistics to the users. For instance, social media can serve both as pools for data collection and data publication in order to get direct feedback from the online users about the usefulness of indices.

2.2. Instantaneousness, Information Overload, Informality & Irregularity

Web 3.0 technologies, such as Semantic Web (Berners-Lee, 2006) and Linked Data (Bizer, Heath, & Berners-Lee, 2009) have been engineered to provide assistance to locate information by human and machine-based tools. Existing *ontologies* and vocabularies have been expanded to handle online statistical information and mainstream statistical standards (e.g. Data Cube

vocabulary (Cyganiak, Reynolds, & Tennison, 2012), Linked SDMX data (Capadisli, Auer, & Ngomo, 2013), etc.).

The most important aspect of the proposed analysis is to identify an effective set of *transformation* and *validation* rules that will enable the timeliness, punctuality, accuracy, comparability, coherence, and eventually, formality of IaD sources.

Based on the past experience in developing Internet and Web standards, these rules should not be all-encompassing from the beginning, but will better follow the “divide-and-conquer” and the procrastination principles. First, the general problem will be demarcated in smaller sub-problems (e.g. IaD for specific indices in ICT statistics) and second, according to the procrastination principle that can be summarized in the phrase “don’t do anything that can be done later by users⁶” most problems confronting the IaD approach can be solved later by other researchers and users of statistics.

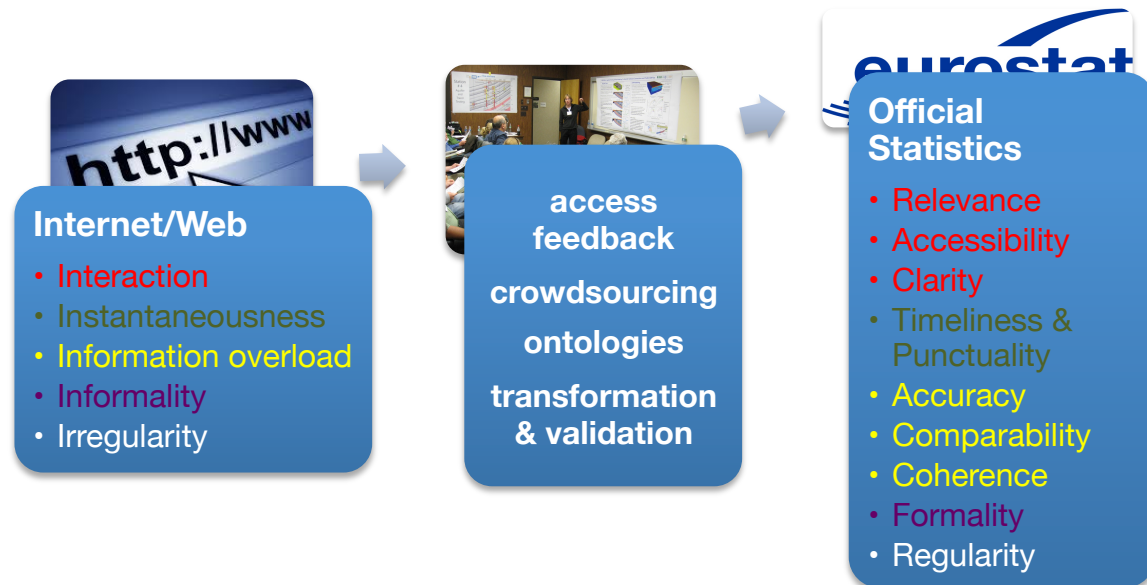


Figure 1: Internet and Web for Official Statistics framework (IW4OS) is designed to orchestrate the main characteristics of the online ecosystem and Official Statistics.

The transition from Official Statistics obtained by real world data through surveys and personal communication with individuals, to a new era of indicators computed complementarily or even solely from Internet and the Web is not easy or obvious. We have to study in depth and understand the universe of Internet and the Web as an extremely complex system in order to fully utilize it for obtaining Official Statistics through the proposed conceptual framework. Next, we discuss this complex nature of the Web and also the problems and challenges in the implementation of the conceptual framework.

⁶ An idea from a 1984 paper by (Saltzer, Reed, & Clark, 1984), that was also used by Zittrain (Zittrain, 2008) to explain Internet’s architecture.

3. Comparing IW4OS with existing models

OECD has proposed a conceptual model which includes the basic elements of ICT supply, demand, infrastructure, products and “content” (OECD, 2009). It is a thorough and detailed description of all the involved parties in the Information Society at the first years of its inception.

During the last five years, the advent of Web 2.0 (e.g. blogs and microblogs, social networks and wikis) and lately Open, Big and Linked data have revolutionalized important aspects of our social and economic life. As Benkler (Benkler, 2007) explains: “What characterizes the networked information economy is that decentralized individual action—specifically, new and important cooperative and coordinate action carried out through radically distributed, nonmarket mechanisms that do not depend on proprietary strategies—plays a much greater role than it did, or could have, in the industrial information economy.”

Section 3.1 describes briefly the main transformative features of the Web 2.0 economy that are not captured by the OECD model and Official Statistics. The specific case of Peer production is further discussed. In accordance to our analysis, the IW4OS model offers a higher level framework that abstracts the aforementioned features in order to directly interrelate them to the core of Official Statistics.

	Users	
	Web 3.0: semantic	
	Web 2.0: social	
	Web 1.0: bulletin board	
	Web	
IaD model (Information flows)	Internet	OECD model (Financial flows)

Figure 2: the IaD model describes the information flows at the technological level of the online ecosystem, whilst the OECD model captures the economic interactions in the Information Society by considering only the first wave of the Web economy.

The “Internet as data source” or IaD model (Dialogic, 2012) depicts only the technological dimensions of Internet and Web operation. In particular, it categorizes three basic types of IaD measurements (User-, Network- and Site-centric). In Section 5 we complement the IaD model by offering a systematic description of the goods and services that are mainly produced and exchanged in the Web (i.e. Web Goods), as well as the different types of online users based on their economic motives and operations. Additionally, the core functions of the Web economy are briefly presented.

Schematically, (M. Vafopoulos, 2011a) describes the emergence of the Web “... as a piece of software code that has rapidly been evolved to an interdependent techno-social system of multi-purpose functionalities. From an interlinked bulletin board with low levels of interaction it has become a construct of multiple interlocking contexts, incorporating a substantial part of financial transactions. Users not only post and link digital content, but also communicate, work, advertise, exchange information in and through it. The social aspects of the Web are fashioned as the ability to create contexts, and an important part of them, economic contexts. Multi-fold social and economic interactions result into a dynamic magma of moral values and code.”

Hendler (Hendler, 2009) defines as Web 3.0, the technology that extends current Web applications using Semantic Web technologies and graph-based, open data. It seems that today, we are witnessing the transition from the social Web (or Web 2.0) to the Web of Data (or Web 3.0).

In accordance to the above definition of the online ecosystem, the IaD model describes the information flows at the technological level of the online ecosystem, whilst the OECD model captures the economic interactions in the Information Society by considering only the first wave of the Web economy (Figure 2).

3.1. The Web 2.0 economy in a nutshell

The OECD model is basically an economic approach to describe the stylized facts of the “traditional” ICT market before the exponential growth of the Web 2.0. Today, during a minute, online users send more than 204 million emails, make 6 million page views in Facebook, watch 1.3 million video clips on YouTube, listen to 61,000 hours of music on Pandora and spend approximately \$83,000 in Amazon⁷. In 2012, only Americans spent 74 billion minutes, or 20 percent of their time, on social networks (Nielsen/Incite's Social Media Report for 2012). This figure could be also interpreted as productivity cost of workplace interruptions that the research firm Basex puts at \$650 billion a year.

These fundamental changes in preferences are supported by new types of consumption and production (e.g. Peer communities), new service sectors (e.g. Software as a Service) and the transformation of existing industries (e.g. mass media). The resulting reconfiguration in the triptych of *production-exchange-consumption* is based on a radical change in the fundamentals of the economy that the Web brings (Figure 3). Basically, the online ecosystem brings a major new source of increasing returns in the economy: *more choices with less transaction costs in production and consumption*.

⁷ <http://abcnews.go.com/blogs/technology/2013/03/what-happens-in-1-minute-on-the-internet/>

This source of value arises from the orchestration of digital and network characteristics of Web goods and services. More choices in *consumption* range from larger variety of available goods, to online consumer reviews and ratings. This updated mode of *connected consumption* allows consumers to make more informed decisions and provides them with stronger incentives to take part in the production and exchange of mainly information-based goods. On the other hand, the provision of more choices with less transaction cost in consumption does not always come without costs. The leading native business model in the Web is the *forced joint consumption* of online information and contextual advertisements in massive scale. Also several cases of users' personal data abuse have been reported.

Consumption in the Web economy becomes more *energetic* and *connected* blurring the borders between production-consumption and (re-) brings in the fore the idea of prosumption. Moreover, the recent emergence of “social commerce” (Stephen & Toubia, 2010) as a consumer-driven online marketplace of personalized, individual-curated shops that are connected in a network, demonstrates the volatile boundaries among production, exchange and consumption in the Web.

Turning to the *production* side, many business operations went online and became less hierarchical, niche online markets and services have emerged and traditional industries revolutionized.

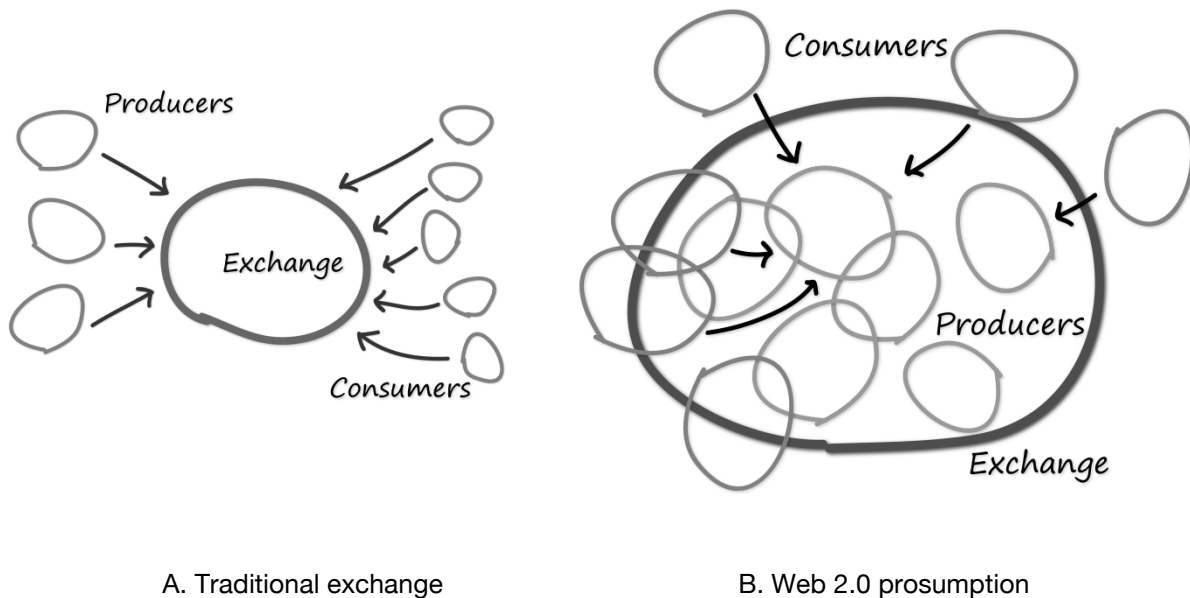


Figure 3: the traditional triptych of producers-exchange-consumers has been updated to the prosumption model where consumers also contact with producers directly in global scale or/and become producers (M. Vafopoulos, 2011a).

The change in user preferences, expectations and behaviour in our networked world is tightly related to the rise of Peer Production communities. Facebook, YouTube, Wikipedia, Twitter and LinkedIn are top in the list of the most popular websites⁸ and worth several billions of dollars.

In particular according to (M. Vafopoulos, 2011a):

“Peer Production is the creative process of user communities, which collaborate, mainly in the Web, to produce sharable goods. These communities enjoy open access to the means of production, share information about inputs and outputs and create pooled knowledge in order to increase the efficiency of future production. In Peer Production communities private information and preferences are revealed and aggregated without frictions, through explicit (e.g. voting, ranking, pricing) and implicit (e.g. tags, reputation) information sharing mechanisms. Because of the fact that information and preferences are public, transparent choice of inputs and outputs is an efficient coordination of rights assignment mechanism. Contrastingly, in traditional business, private hierarchical structures are designed to minimize coordination costs. Peer Production communities could be more efficient than firms or markets if they can operate under less coordination costs in atomizing production. In this context, entrepreneurs have begun to exploit distributed economies of scale in Peer Production on industries with high coordination costs (e.g. social networking, freelancers markets) by providing production platforms.”

If we want to generalize, Peer Production pervades both the private and the public domain and the demand-supply dichotomy by introducing a the third mode of production, a third mode of governance, and a third mode of property (Bauwens, 2006).

In this context, a further investigation is needed on the potential ways of incorporating this new complex and dynamic reality in Official Statistics. IW4OS model offers a fertile ground for evaluating existing tools and methodologies and testing new state-of-the-art approaches.

3.2. The IaD model

The “Internet as data source” or IaD model (Dialogic, 2012) is a recent study that is related to the IW4OS model because it conceptualizes the *technological* aspect of Internet and Web use. It identifies three basic types of IaD methods, namely the User-, Network- and Site-centric measurements.

By just following the bit or the click streams we can only model how the machines are used but not for which purpose by their users. These measurements should be complemented and tested in a conceptual model for Web usage based on the users’ needs and motives. In the next section, following the approach of (M. Vafopoulos, 2011a) we try to identify the nature of goods and services that have been created because of the advent of the Web and are mainly residing in it. It

⁸ http://en.wikipedia.org/wiki/List_of_most_popular_websites

is of equal importance to complete this analysis by modelling users' behaviour based on their main economic incentives and performed functions.

4. Reflections of the conceptual framework to ICT statistics

The OECD guide⁹ for measuring the ICT sector is the commonly adopted way to measure the various facets of ICT activity namely, products, ICT infrastructure, supply, demand by businesses, demand by households and individuals, content and misc. Indicators have also been proposed by the ITU¹⁰. Currently ICT usage surveys cover activities of Internet where individuals interact with web servers, which typically offer services or merchandise.

Social networks tend to capture the major part of human activity and as such constitute a trend (yet though not an indicator) of human activities. For instance Twitter¹¹ messages capture everyday activities. Mining through register IDs reveals significant information of epidemics. Under certain limitations¹² collecting equivalent statistics compared to traditional survey statistics is a challenge because Internet and social media reformulate the form of meaningful queries. Facebook¹³ constitutes another big reservoir of social life and partially ICT usage indicator. All big firms have moved to Facebook to gain from the personal connection to users. In terms of ICT usage, Facebook acts as a social specific web that introduces firm and products to individuals.

M2M communication has started to capture an emerging human need for better management of their facilities (i.e. ports, sewage systems, smart electricity grid, etc.) thus ultimately evolving to smart cities¹⁴. Millions of data fountain from sensors spread around indicating human activities. Those data require efficient manipulation and correlation with the legacy merchandise activities. Something like that is on verge of semantic annotation of linked-data and statistics research.

4.1. ICT Usage

The aforementioned emerging uses of Internet require new types of queries and methodologies in order to be tracked of. This is clearly a revolution when considering Internet as a data source. Adopting an evolutionary approach we propose to track the legacy ICT usage by means of Internet mechanism. Unfortunately the data generation has not yet been restructured in order to recover indicators explicitly from data repositories, thus we still limit ourselves to indirect methods (such as web-crawling, etc.) in order to collect the required data. For instance Facebook

⁹ <http://www.oecd.org/sti/ieconomy/theictsector.htm> , <http://www.oecd-ilibrary.org/deliver/fulltext/3011041ec072.pdf>

¹⁰ <http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>

¹¹ <http://www.twitter.com>

¹² <http://dimacs.rutgers.edu/~graham/pubs/papers/cormodewdsa.pdf>

¹³ <http://www.facebook.com>

¹⁴ <http://www.smartsantander.eu/>

does not provide any gateway (i.e. API) to collect metrics and indicators of references to keywords whether those refer to advertised products and services or to free text.

4.1.1. ICT products

As far the ICT products are concerned data are collected either from surveys or from customs offices. It is expected that this collecting mechanism will deteriorate as more individuals and companies are composing their supplies over electronic stores. It is therefore important to circumvent the existing or immediate foreseeable barriers for this data collection.

As an ever-growing number of ICT products are marketed through on line shops we propose the collection of ICT product data through proper characterization of online store products. As cross border e-market is common practice, we propose additional tagging of ICT product sale data followed by country specification in order to collect country aggregated metrics. Such metrics can be collected through web logs.

4.1.2. ICT infrastructure

As far as infrastructure is concerned, the collection of indicators becomes more feasible given that Internet service providers (ISPs) and regulation authorities provide the technical and legislative means respectively. A typical indirect usage of the ICT infrastructure might be provided by the number of registered domain names (DN). Registered DNs could be supplied through Internet from top level country registrars.

Internet broadband usage has always been in the center of interest of carriers strictly for capacity prediction. Regulation authorities in various countries have started projects with active methods for broadband speed measurements (i.e. www.samknows.eu, www.measurementlab.net/). Those measurements utilize additional components in order to check the quality and the effective speed of broadband connection of consumers. Different metrics such as mean connection time could only be retrieved by using browser-based facilities (i.e. customizable search bar) or carrier based facilities such as the radius accounting database.

An indicator of national ICT infrastructure could be estimated by the total sum of exchanged traffic in national internet exchange points (IXP). As IXPs maintain online volume graphs it is possible to use them as an online data source.

An indication of ICT infrastructure could be given by the number of autonomous systems (AS) in the routing table, while these data are broken out per country. Autonomous Systems are the elementary routing placeholders in the Internet with their own distinctive routing policies that appear in the Internet routing table^{15,16}.

In addition to legacy carriers, content delivery networks (CDN) emerge as competitive Internet rich media (audio, video) transporters. For instance the akamai (www.akamai.com) CDN

¹⁵ <http://bgpmon.netsec.colostate.edu/>

¹⁶ <http://www.ripe.net/data-tools/stats/ris/routing-information-service>

provider delivers indicators for average connection speed, average peak connection speed, high broadband connectivity (>10 Mbps) and normal connectivity (<10 Mbps) through their quarterly edition “State of the Internet”¹⁷.

4.1.3. ICT supply

The overall ICT supply can be estimated by sum of products and services. Internet is used by service oriented companies which declare their presence in terms of domain name. Hence a typical indicator of a national ICT supply might be provided by the number of registered DNS names. Aggregation per country or per subdomain (i.e. .ac, .co) -wherever this fits- is maintained by the country top level domain (cTLD) registrar.

As far the ICT product supply is concerned the only possible means of utilizing the Internet as a data source is to collect metrics from web crawlers or meta search engines for internet shopping. For instance the www.skroutz.gr/ lists the number of products displayed from all shops, the number of individuals and the number of shops. In a relevant way, auction sites (i.e. www.ebay.com) host the number of products on sale by individuals. As auction sites host electronic shops also, it is useful to collect indicators by aggregating data per number of e-shops.

E-government sector

Another fundamental sector of ICT activity is the activity of public sector with respect to the automated internet-ready application for the citizens. The sum of publicly offered services to citizens for purposes of the state or for transactions with the state constitutes the E-government (e-gov) sector. The number of services offered by the central government as well as from the regional and municipal sector constitute the total supply of e-gov sector. The number of services are typically monitored from national ICT observatories or from aggregation portals of e-gov sites.

The demand of e-gov sites could be easily quantified by the number of different registers and the number of submitted and produced objects. Those indicators could be easily accessed by weblogs of portals hosting the respective e-gov services.

ICT demand - transactions

The demand of ICT corresponds to activity from individuals to buy services or products offered in Web sites. Although it is possible for an individual to commit for payment by using a legacy payment method as mail order we will focus our study only to those transactions which are completed electronically. For the merchant sector an indicator of ICT could be retrieved either by the selling web sites or from equivalent web banking activity.

¹⁷ www.akamai.com/stateoftheinternet/

Web banking

Web banking activity corresponds to demand from individuals paying via credit or debit cards. It is assumed that only the banking sector could verify the number and volume of electronic card transactions hence the banking sector should somehow differentiate between legacy electronic credit card machines used in shops and web sites. Indicators for web banking could be traced by web log activity traced indicating the final state of web site visit (i.e. payment or abort to a different site).

Furthermore the banking sector considers internet and mobile banking as means to minimize the operational costs. It is essential to gain access to overall number and volume of transactions conducted by Internet banking sites as opposed to legacy order in front of desks.

State driven ICT demand

Another interesting figure of the ICT demands corresponds to the ICT demand by the state as it is referenced by public request for proposal or even more of contractual agreements with specific CPV codes¹⁸ for ICT.

4.2. Content and media

Media (audio and video) has rapidly captured the interest of broadband users. Media companies all over the world have migrated the majority of their content to Internet allowing users to watch it using new type of end devices such as Internet TVs and smart-phones.

User centric (i.e. youtube) and legacy media owners (newspaper and TV) are quickly migrating to the Internet. Indicators of media content are the average number of viewing time and total number of hits per item selected as a popularity indicator. Unfortunately not all sites provide indicators. It is only through CDN providers where such indicators could be retrieved.

5. Weaving the Web economy: goods and users

First of all, let us define the basic constituents of the online ecosystem. Which are the goods and services in the Web and how users economize the digital information flows. The pre-existing classification of Data, Information and Knowledge seems that is not fully fitting the salient features of digital information in a networked world. As (M. Vafopoulos, 2011a) argues, “Information can be now digitized (if not digital already) and transferred over networks with minimum cost. Data are transformed into information and knowledge in new ways globally. Human networks, and the knowledge which flows through them, become partially observable (e.g. social networking, institutional Web sites) creating new forms of production and consumption.”

¹⁸ http://simap.europa.eu/codes-and-nomenclatures/codes-cpv/codes-cpv_en.htm

Thus, we need new concepts to capture these transformative information life cycles that are relevant to a more self-powered, collaborative and networked economy.

Let us first consider a simple and compact definition about the goods and services that have been emerged because and by the advent of the Web, the so-called “Web Goods” (M. Vafopoulos, 2011a).

Web Goods (WGs) are defined to be sequences of binary digits that (a) are identified and communicated by an exclusively assigned URI and (b) affect the utility of or the payoff to some individual in the economy. Their market value stems from the digital information they are composed from and a specific part of it, the hyperlinks, which connect resources and facilitate navigation and editing over a network of Web Goods with minimum cost¹⁹.

The next step is to model how users are producing and consuming WGs. In this context, a comprehensive categorization of online users is provided in order to systematically present the main behavioral aspects and to analyze the core functions of the Web economy.

The distinction of Users is based on the motivations and the economic impact of their actions in the Web ecosystem.

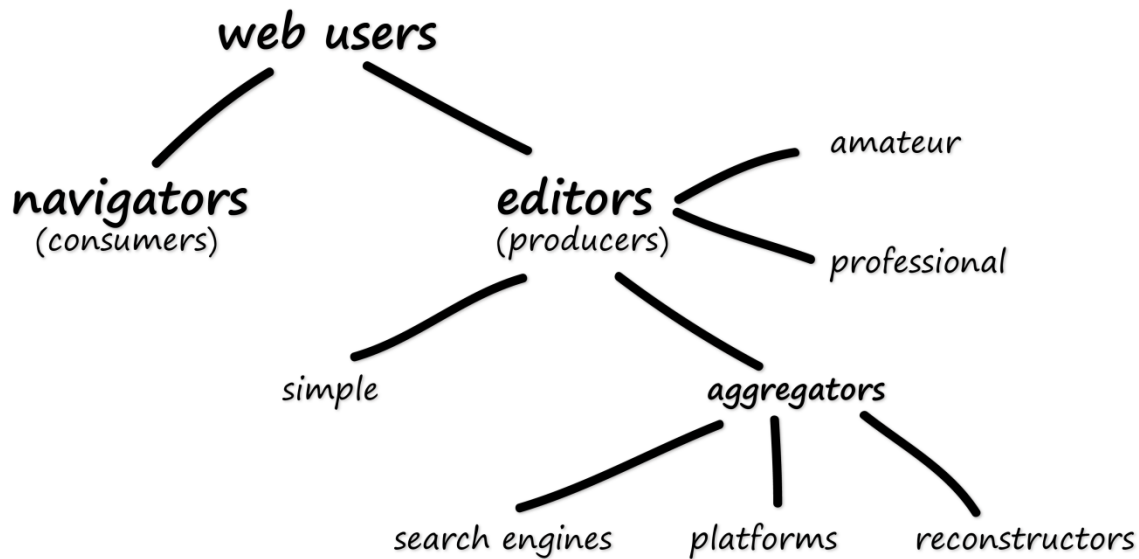


Figure 4: The Web Users are primarily partitioned to Navigators and Editors of WGs (M. Vafopoulos, 2011a).

¹⁹ The notion of “Web Goods” could be conceived as part of the broader definition of “Web beings” (Michalis Vafopoulos, 2012).

First, users are partitioned to Navigators and Editors of WGs (Figure 4). Navigators consume information by browsing, surfing or, in general, accessing the Web network. Editors produce WGs by creating, updating or deleting online content and links in the Web network. Editors could be further categorized to Amateur and Professional based on their economic incentives.

Contradictorily to Amateur Editors (e.g. Wikipedia editors), Professional Editors are profit maximizers and target direct financial compensations in producing WGs (e.g. a blog with paid advertisements). On the other hand, Amateur Editors, in not-for-profit community settings (e.g. Open Source), are motivated by moral reward, self-confidence and reputation-building. This *temporal disengagement* between effort and reward could offer an explanation for the fact that Editors may provide their knowledge, effort and time for free (Quah, 2003).

In the context of social networking, Amateur Editors are stimulated by getting a higher relative contribution status, compared to their peers and future utility from the consumption of connected goods provided by their peers (Kumar, 2009). In such cases, Amateur Editors are the initial producers of WGs that are packaged and commercialized by a Professional Editor acting as a platform (e.g. Facebook).

The service pluralism of Web 2.0 could be also approximated by a *function-based* distinction among Editors. Editors can be elaborated, according to their aggregation capability, to Simple and Aggregators. Aggregators based on their automated mechanisms for selecting and presenting WGs could be distinguished to Search Engines, Platforms and Reconstructors.

Simple Editors create content manually. *Search engines* have been constructed by sophisticated algorithms that can automatically aggregate, index, classify and commercialize WGs. *Platforms* are a set of technologies and incentives that make possible Peer production and aggregation under common infrastructure of WGs (e.g., Flickr). They are an important part of the Web 2.0 corpus because they enable Users to co-create. Commonly, are open-access “walled gardens” since users do not pay financial fees to access them, but they produce online content that is difficult or impossible to be transferred to other platforms (lock-in) and their underlying code is not open source. Most of these Platforms are commercialized by advertisements (e.g. Facebook) and/or subscriptions (e.g. LinkedIn). Nevertheless, there are also not-for-profit platforms that operate as Amateur Editors.

Aggregators based their success on the exploitation of the *multi-sided platforms* (or two-sided network effects) (Evans & Schmalensee, 2007; Evans, 2003), by facilitating three interrelated cost-reducing functions: matchmaking, building communities and providing shared resources.

Reconstructors are advanced technologies (mostly based on Web 3.0) enabling the deconstruction, filtering, modification and reconstruction of structured and personalized WGs. For instance, last.fm unbundles music tracks from albums and playlists in order to reconstruct new playlists based on the collaborative filter that match to personal preferences. Reconstructors

could be considered as the next generation platforms that are based on semantic processing of WGs.

The dominant players in Web economy are fighting to consolidate both horizontally and vertically as Editors (e.g. Google News). Advertisers in the Web are Professional Editors that create online content to promote consumption of specific goods and services.

On the grounds of the previous analysis, we can now answer the question of how navigating and editing online content are interrelated and build economic incentives that are leading to the current gigantic network of online information and interaction.

In few words, Navigators explore the Web because they enjoy utility by consuming WGs (Figure 5). This navigation results traffic streams for Editors. Amateur Editors are concerned to attract traffic for their content, even if they do not actually own it (e.g. personal profile page in Facebook)²⁰. In contrast, Professional Editors, which own WGs, are trying to transform this traffic into income by selling it to third parties, advertising or performing direct sales of both physical and WGs. The resulting income is considered as the basic incentive for Editors to update the already existing and create new WGs, contributing to the new Web network with novel possibilities for Navigators to maximize their utility (Figure 5).

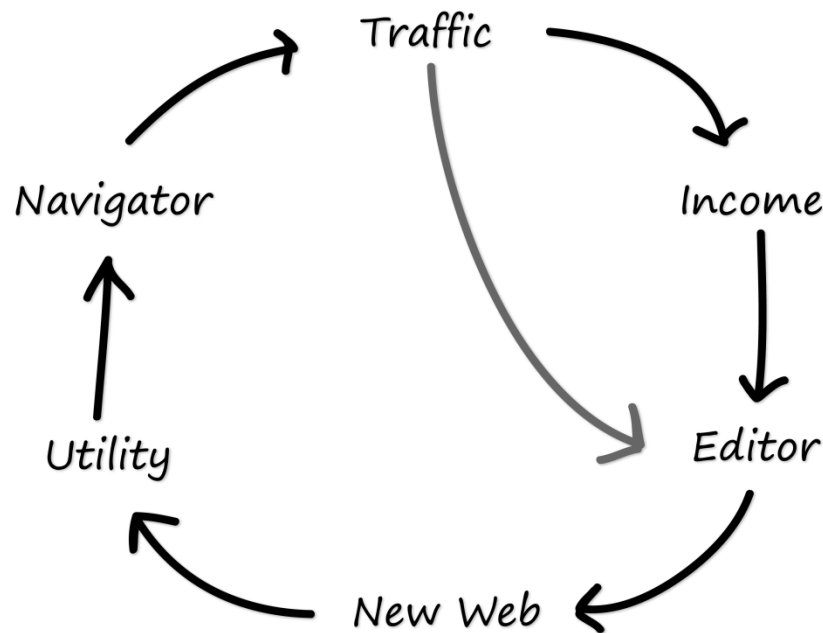


Figure 5: In the Web economy, Navigators explore the Web to acquire utility. This navigation creates exploitable traffic for Editors, which are motivated to update the existing Web (M. Vafopoulos, 2011a).

²⁰ This function is represented in Figure 5 by the line that connects directly traffic to Editors.

6. The Crossroads of Information Society Policies and Indicators

Official ICT statistics must provide information to policy makers on multiple aspects of the economy and society, under conditions of constant evolution. In addition, they can be used by businesses in decision-making based on evidence, and by researchers who endeavour to illuminate the meaning of the non-stop proliferation of new ICTs everywhere and the impacts of their uptake and use in the behaviour of people, businesses, and governments. Any concordance between statistics and policy needs is not static but is inevitably subject to dynamic changes.

At any given time, official statistics must respond in a timely manner to the needs of the day, and do so with an expected level of quality based on trust. For that, the production and harmonization of statistics must rely on the development, adoption, communication and frequent revision of nomenclatures, norms and processes which, in any quality system, imply various lags. As well, all that must be realistically balanced against the use of resources. The eternal interplay between short-term responsiveness and long-term continuity is clearly present.

The NSIs must also provide results that satisfy national needs, within their capacity to implement programming, in a way that will lead to desired comparability in Europe and beyond, respect rather than disrupt existing structural equilibria, and keep costs under control. Undoubtedly, this is a tall order. Today, much promise is assigned to the use of ICTs and other Big Data as data sources to help bridge the gap between evolving data needs and data production. In the very least, it is advisable during planning to look where the ball will be, not where it has already been.

6.1. A natural evolution

From the early days of statistical work on the Information Society, beginning at the OECD, there was a consensus among participants that the overarching framework to guide developments, and match evolving policy needs to indicators, would be one that moves progressively from ICT **access**, to **usage**, and eventually **impacts** (the S-curve)²¹.

In the early days of the introduction of a new technology, access becomes the key policy issue and indicators are critical to monitor the diffusion of the technology and its growth. Such matters relate to the Digital Divide. These policy needs led to the measurement of the penetration of computers, the Internet, mobile phones, including network coverage, and the like. It was always understood, though, that while such indicators were indispensable, they were not the way of the future for two main reasons. First, as penetration increases and saturation is approached the value of these indicators would diminish – and consequently they should not be continued in the future (this has already happened in most European countries for mobile phones). Second, access served as a mere stepping-stone to actual use, which is where the benefits would originate. Therefore, indicators of usage in all their manifestations (frequency, intensity and the like) were terribly important for policy - and indeed business purposes. (Fittingly, many of the existing indicators in the Eurostat questionnaires are devoted to Internet use).

²¹ Guide to measuring the information society, OCED (2005), p. 10

In the end, it is usage that would lead to impacts, which could be examined later under the logic that they represent a higher level of understanding. Social and economic impacts should be sought among people, businesses, and government, as well as from a sectoral perspective, such as education, health, and even phenomena such as e-health, e-education, e-commerce and many more. Impacts are more difficult to quantify directly. Instead, over time, they are painstakingly inferred through accumulated experiences and data, and with linkages to extraneous data sources not collected as part for the new ICT statistics (e.g. correlated to business performance, such as profitability, productivity etc.)²².

The evolution of statistical measurements then naturally moved from access to usage, while the quest for impacts will undoubtedly continue. Evidence of such evolution is still visible in the Eurostat model questionnaires. For one, generally, access was studied at the household level (which is still the entry point of the community survey for basic access), but moving to usage clearly the individual became the unit of observation. Such indicators are now more relevant than ever, but usage too has morphed and assumed a new meaning.

Short lives seem to be the fate of ICTs. As soon as we started to get a handle on the usage of computers and the Internet in their earlier incarnation, we recently witnessed the rapid move to wireless technologies and the concomitant move to a wide variety of portable and mobile devices. In a few short years, we moved from usage of one computer in a household (and not all) to multiple devices per individual. The S-curve got a new lease on life, and a new cycle started. Clearly, this changes the game for all – policies, business decision etc. While this situation does provide new opportunities, it does at the same time complicate the life of official statistics to respond and be relevant.

The fundamentals for measurement though remain the same, especially with respect to the fact that the ever-important component of measuring usage is much better done from the digital footprints rather than direct inquiries on people or enterprises.

We need to move to the new world, and break the links between measurements of interest and dedicated instruments of measurement - or augment such links. It so happens, that when usage is concerned, measuring statistics on the Internet from the Internet is superior to traditional surveys. The same applies to devices like smartphones. Wherever digital footprints originate, data are to be had. How to connect what can be collected to what is needed should become the new skill in official statistics. However, we are still not quite there. This study aims to help in such a transition.

Europe 2020 articulates the EU's strategy until the end of the decade to deliver much needed growth in the economies of its member states, and do so in a way that will be smart, sustainable

²² Information Society: ICT impact assessment by linking data from different sources
http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/documents/Tab/ICT_IMPACTS_FINAL_REPORT_V2.pdf

and inclusive. This is vital under the ongoing geopolitical transformations at a global scale. Not only higher levels of employment have become an imperative, but progress on that front is also intricately linked to social cohesion as manifested by the experience of several countries in recent years. Many of these hopes are tied to current policy objectives in the broad area of ICTs²³. At a high level, seven flagship areas have been identified as key in the efforts to stimulate growth and jobs in Europe: Create a new and stable broadband regulatory environment; New public digital service infrastructures through Connecting Europe Facility loans; Launch Grand Coalition on Digital Skills and Jobs; Propose EU cyber-security strategy and Directive; Update EU's Copyright Framework; Accelerate cloud computing through public sector buying power; Launch new electronics industrial strategy – an "Airbus of Chips".

The digital agenda for Europe is the first of these seven initiatives. Its overarching objective can be realised through increasing investment in ICT, improved e-skills in the labour force, and continuous innovation in the public and private sectors. The digital agenda contains 13 specific goals which encapsulate the change sought to be achieved, and 101 actions, grouped around seven priority areas: Scoreboard, Interoperability & Standards, Trust & Security, Fast and ultra-fast Internet access, Research and innovation, Enhancing digital literacy, skills and inclusion, ICT-enabled benefits for EU society. Many of the policies relate to broadband. Progress against these targets is measured in the annual Digital Agenda Scoreboard. Moreover, a good account of needed indicators for benchmarking is provided²⁴.

To support the Digital Agenda for Europe, Eurostat accommodates as many indicators from there as possible in the ICT surveys. Moreover, the list of indicators is subject to annual review. In addition to the two surveys, additional statistics are provided for telecommunication services and the ICT sector. The aggregated telecommunications data are submitted by NSIs and come from administrative sources. ICT sector statistics are derived from existing source in NSIs, such as Labour Force Surveys, Structural Business Statistics, PRODCOM, R&D statistics, and National Accounts. Timeliness of the data release depends on the timeliness of the source data and thus varies accordingly. The same is true for comparability. Key ICT sector indicators are value-added and employment.

At the international level, these are coordinated under the International Partnership on Measuring ICT for development²⁵ and aim at identifying the state of the evolution of the sector in individual countries, and reveal national strengths and weakness (e.g. ICT manufacturing vs. services).

1. ²³ EU (2013a), Digital Agenda for Europe, <http://ec.europa.eu/digital-agenda/digital-agenda-europe>

2. EU (2013b), Digital Agenda for Europe, <http://ec.europa.eu/digital-agenda/about-our-goals>

24 EC (2009), i2010 High Level Group, Benchmarking Digital Europe 2011-2015: a conceptual framework, European Commission, October, 2009

²⁵ International Partnership on Measuring ICT for development, Core ICT Indicators, http://new.unctad.org/Documents/Core%20Indicators/Core_Indicators_English_2010.pdf

7. Mapping current ICT statistics against the Internet as a data source

The indicators measured by Eurostat over the last several years, both on the household and on the business side, evolved over time to respond to policy needs as described earlier. At the same time, they were designed in such a way as to provide basic comparability across European member states, and OECD countries. With cumulative efforts, the existing arrangements serve their basic purpose.

Contemplating departures from the existing way of compiling and disseminating indicators into the new ways made possible by the digital footprints, particularly the Internet, is a necessary bridge to the future. However, it also needs the build-up of a certain level of comfort based on understanding of what this really entails. For this, a number of dimensions need careful examination. Such work has started, and this report contributes to this direction.

As a minimum, the extent of feasibility must be explicitly addressed.

- Which ICT indicators are possible to measure from the Internet, among the existing measurements?
- Which ICT indicators are not possible to measure from the Internet, and what will happen to them?
- Which ICT indicators are possible to measure from the Internet, among those not currently measured? (This is addressed in section 8).

Then, issues related to quality and/or desirability also assume significance. Just because something is feasible does not mean it is desirable. The quality implications must be understood and the trade-offs identified. As a useful incremental step, this section takes this task on, first through a brief general description and then through a detailed approach specific to the existing questionnaires. Remaining issues, e.g. legal, social acceptability and the like will be addressed in subsequent components of this project.

7.1. Top-level trade-offs

As a rule of thumb, questions related to matters of access cannot be answered from the Internet. To the extent that their measurement is important, the availability of computers, desktop or portable, mobile phones and other ICT devices cannot be known from the Internet. To some extent, this is an oxymoron and reminiscent of the digital divide: knocking at the door of the “haves”, you cannot find the “have-nots”. Due to the hierarchical structure among some ICTs, and in the event of near-saturation in Internet penetration, a case could be made that technologies lower in the structure can be inferred from technologies higher in the structure (e.g. a computer, or old day modem, exists if the Internet is present – akin to not asking if someone with a university degree graduated from high-school). This, however, would presuppose a stagnant technological environment, for which there is no basis.

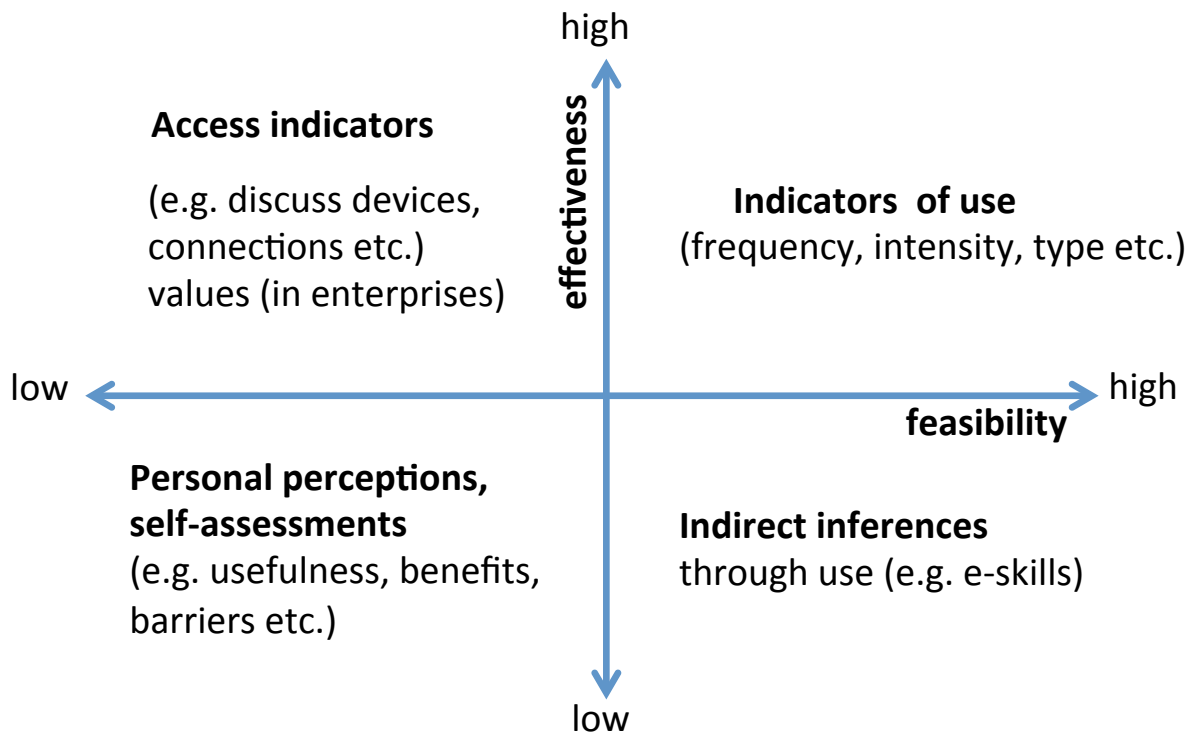
Generally, the Internet as a data source is ideally situated for the measurement of indicators of use. Institutionalizing such an alternative mechanism as a way of the future can provide reams of

data. Not only it avoids the response burden but it can provide more and better data. This is so because questionnaires are restrictive instruments when it comes to the diversity of use. Regardless of how well they may be designed, any questionnaire-based measurements for the frequency, intensity and types of use cannot do justice to the phenomenon of use. Not only the recall issue is a well-known problem for any lengthy time-period, but given the proliferation of devices and the fact that use has become second nature for so many, the granularity involved can be vastly improved. With such data, analysis will be the new challenge - to derive insights.

Lastly, the Internet is not meaningful for content concerning the views by individuals or businesses of their experiences or any subjective assessments and opinions.

The following schematic helps visualize such top-level trade-offs. It shows the degree of feasibility of different categories of measurement against their effectiveness. The latter is defined roughly, as a combination of the desirability of continuing to have such measures and/or their expected fitness for the uses intended vis-à-vis the traditional methods.

The gathering of new data, and their granularity not previously possible, will supersede some of



the current measurements. A specific example would be related to the time period of use, as now the questionnaire asks for the last 3 months – which may well become irrelevant in real-time data. Such analysis can serve as the backdrop for the setting of priorities as to what can be tackled first, as well as situate the detailed analysis that follows.

Before we proceed in the detailed analysis of existing indicators, an important clarification must be made. Migrating from the current norm, where a certain set of indicators is captured through questionnaires, to indicators that can be generated from digital footprints does not render itself to a simple, one-to-one correspondence. For example, as was mentioned earlier, we have moved from one computer per household to multiple connected devices per individual. This complicates tremendously the aspect of use - so long as the individual remains the unit of observation.

This report is concerned primarily with what indicators can be obtained under conditions of access to an individual's desktop or portable computer (or smartphone), as per the project's terms of reference. It must be clearly understood that this will cover a sub-set of an individual's Internet usage, unless we contemplate another paradigm in which, for a number of individuals in a sample, we ask for access to ***all devices*** they use for a period of time. Even then, we would have to deal with several issues such as:

- use from work, as the questionnaire refers to personal use (e.g e-commerce purchases for private use)
- use of shared devices by other individuals at the household
- perhaps different devices added during the period of study
- individuals who still do not have access

Therefore, adhering to the spirit of the project, this report places the emphasis on a gradual migration from the questionnaire to indicators that can be obtained from access to a user's computer or smartphone. The bigger picture is discussed explicitly at the end of the report.

7.2. Households/Individuals

Early policy needs for statistical information were driven by questions regarding access to ICTs at the household level. Very soon, as individuals started to embrace the new technologies and the reach of communication networks expanded, the interest shifted to issues of use. This put the emphasis on the definition and measurement of many indicators of usage. These, logically, changed the nature of the statistical unit from the household to the individual. Moreover, survey questionnaires were subject to constant evolution to align with, feed into, and be relevant to evolving policy needs. While this process continues to this day, it is certainly conceivable that statistical responses to changing policy needs could be greatly aided by indicators collected directly from the Internet.

The existing questionnaire of the 2013 survey used in this study contains 6 modules, the first answered by any household member and the rest by a selected individual. These, then, are followed by a substantial section on socio-demographic background characteristics which make possible the analytical decomposition of the collected data by several groups. In the tables that follow we present the questions and comments about each one's amenability to measurement with automatic Internet-based methods.

HOUSEHOLDS QUESTIONNAIRE		Comments
Module A: Access to Information and Communication Technologies		
A1	Do you or anyone in your household have access to a computer at home?	The availability of computers or other peripheral ICTs, particularly as stand-alone not necessarily connected to the Internet, can only be answered by household members.
A2	Do you or anyone in your household have access to the Internet at home?	
A3	What types of Internet connection are used at home?	Possible to capture this information. The data will be superior to data collected from the existing questionnaire since the technicalities surrounding the type of connection is somewhat problematic for respondents.
A3a	broadband	
A3b	ISDN, dial-up or other narrowband	
A3c	Wired fixed (cable, optical fibre, Ethernet, PLC, etc.)	
A3d	Fixed wireless (satellite, public WiFi)	
A3e	mobile phone network (at least 3G, e.g. UMTS) via a handset	
A3f	mobile phone network (at least 3G, e.g. UMTS) via a card or USB key	
A3g	Dial-up access over normal telephone line or ISDN	
A3h	Mobile narrowband connection (less than 3G, e.g. 2G+GPRS, used by mobile phone or modem in laptop)	
A4	What are the reasons for not having access to the Internet at home?	
A4a	Have access to Internet elsewhere	Information about opinions can only be collected with a questionnaire.
A4b	Don't need Internet (because not useful, not interesting, etc.)	
A4c	Equipment costs too high	
A4d	Access costs too high (telephone, DSL subscription etc.)	
A4e	Lack of skills	
A4f	Privacy or security concerns	
A4g	Broadband Internet is not available in our area	
A4h	Other	
Module B: Use of computers		
B1	When did you last use a computer (at home, at work or any other place)?	These data cannot be obtained from the Internet. Even if it were possible to use federated data, inferring the usage of computers from the Internet is not necessarily optimal. If the intent of the indicator is to capture non-networked use. On the other hand, a good case can be made that such indicators may not serve much purpose in the future.
B2	How often on average have you used a computer in the last 3 months?	
Module C: Use of the Internet		
C1	When did you last use the Internet	Information captured by this question can now become much more granular, in a way that was not possible through the questionnaire. Having access to real-time use information, there is no need to be restricted in knowing the respondent's last Internet use within the last 3 months or a year. Among other uses, analytical profiles of different user groups according to the frequency and intensity of use can also be constructed.
C2	On average how often did you use the Internet	Information captured by this question can now become much more granular, in a way that was not possible through the questionnaire. Having access to real-time use information, there is no need to be restricted in knowing the respondent's frequency as daily, weekly etc. Superior and very detailed information can be obtained, including the exact number of sessions, their duration, their distribution in the course of a day or week, any differences between week days and weekends, and many other angles that can be supported by the collected data on usage. Recall issues among respondents, and the burden imposed on them from the need of detailed answers will be avoided. Among other uses, analytical profiles of different user groups according to the frequency and intensity of use can also be constructed.
C3	Where have you used the Internet in the last 3 months (using a computer or any other means)?	These questions cannot be answered from the Internet.
C4	Do you use any of the following mobile devices to access the Internet away from home or work?	
C4a	Mobile phone (or smart phone)	
C4a1	via mobile phone network	
C4a2	via wireless network (e.g. WiFi)	
C4b	Portable computer (e.g. laptop, tablet)	
C4b1	via mobile phone network, using USB key or (SIM) card or mobile phone as modem	
C4b2	via wireless network (e.g. WiFi)	
C4c	Other devices	
C4d	I don't access the internet via any mobile device away from home or work	
C5	For which of the following activities did you use the Internet	Information captured by this question can now become much more granular, in a way that was not possible through the questionnaire. Generally, taxonomies on the types of use can be as long as their designers desire them, as there is no real end to the amount of detail. Practically, prioritization takes place and choices are made, constrained by scarce resources and concern for response burden. Question C5 contains 6 main categories, which in turn include 17 individual types. With an appropriate design, data on those and many additional types of user activities, not in the current question, can be obtained from the Internet. A clarification is in order: Internet-based use will not be able to capture "use for private purposes", as is currently asked in the questionnaire.
a	Sending / receiving e-mails	
b	Participating in social networks	
c	Reading online news sites / newspapers / news magazines	
d	Seeking health-related information	
e	Looking for information about education, training or course offers	
f	Finding information about goods or services	
g	Downloading software (other than games software)	
h	Posting opinions on civic or political issues via websites	
i	Taking part in on-line consultations or voting to define civic or political issues	
j	Doing an online course (in any subject)	
k	Consulting wikis to obtain knowledge on any subject	
l	Looking for a job or sending a job application	
m	Participating in professional networks	
n	Using services related to travel or travel related accommodation	
o	Selling of goods or services, e.g. via auctions	
p	Telephoning over the internet / video calls (via webcam) over the internet	
q	Internet Banking	

Table 1. 2013 Households ICT survey questionnaire and comments about amenability to measurement through the Internet.

	MODULE D: Use of e-government	
D1	Did you contact or interact with public authorities or public services over the Internet	Capturing user traffic statistics on the Internet can identify such websites, but more specificity and structure would have to be imposed – with the potential of not only obtaining comparable information but improving on its degree of detail and the eventual interpretability. For instance, the open-endedness of what constitutes the public sector can be improved – and it may not be the same across countries. Ministries of national governments can be specified and tracked, as can departments of municipal governments and universities. Moreover, depending on the methodology of the exercise (discussed in Section 10), the period of the last 12 months (used to avoid issues of seasonality) may be possible or may need to change to the period of monitoring, particularly if more than one collection cycles occur during one year. At the same time, the specific transactions asked are not likely to prove possible.
a	Income tax declaration	
b	Downloading official forms	
c	Submitting completed forms	
D2	Did you use websites of public authorities or public services for any of the following	
a	Income tax declaration	
b	Claiming social security benefits	
c	Requesting personal documents (passport, ID card or driver's licence) or certificates	
d	Public libraries (availability of catalogues, search tools)	
e	Enrolment in higher education or university	
f	Notification of change of address	
D3	Have you experienced any of the following problems when using websites of public authorities or public services for private purposes in the last 12 months?	These questions do not render themselves to Internet measurement.
a	Technical failure of website	
b	Insufficient, unclear or outdated information	
c	Support was needed but not found (on-line or off-line)	
d	Other	
D4	Are you satisfied or dissatisfied with the following aspects public authorities or public services in the last 12 months?	
a	Ease of finding information	
b	Usefulness of the information available	
c	The information provided on the progress, follow-up of the request	
d	Ease of using services on the website	
D5	Did you contact public authorities or public services using methods other than websites for private purposes in the last 12 months?	
a	yes, by telephone (excluding SMS)	
b	yes, by e-mail	
c	yes, in person, by visits	
d	yes, by other means (e.g. post, SMS, fax)	
e	no	
D6	What were the reasons for not submitting completed forms to public authorities' websites for private purposes in the last 12 months?	
	MODULE E: Use of e-commerce	For e-Commerce, the information asked in all questions can generally be captured from the Internet. On the other hand, it may not be possible to differentiate between free and bought services/content consumed online.
E1	When did you last buy or order goods or services for private use over the Internet	The time period of the Internet-based data will coincide with the time period of the study.
E2	What types of goods or services did you buy or order over the Internet	This question conceptually fall under the previous discussion regarding the open-endedness of such taxonomies (food, medicine, clothes, hardware, software, hotels etc). Answers to those and transactions for many other products may be obtained – subject to additional back-end work. In other words, the compilation of data will not be automatic. Information will have to be mined not only for the sites visited but specific pages of such sites, and this must be compared against appropriate lists of businesses that must be created. It may well be that we end up with more specific product groupings than in the existing questionnaire, but that would not represent a loss!
a	Food or groceries	
b	Household goods (e.g. furniture, toys, etc)	
c	Medicine	
d	Films, music	
e	Books, magazines, newspapers (including e-books)...	
f	e-learning material	
g	Clothes, sports goods	
h	Video games software and -upgrades	
i	Other computer software and -upgrades	
j	Computer hardware	
k	Electronic equipment (incl. cameras)	
l	Telecommunication services	
m	Share purchases, insurance policies and other financial services	
n	Holiday accommodation (hotel etc.)...	
o	Other travel arrangements (transport tickets, carhire, etc.)	
p	Tickets for events	
q	Other	
E3	Were any of the following products that you bought or ordered over the Internet downloaded or accessed from websites rather than delivered by post etc.	The same more or less applies to this question, although in this case categorization will be more obvious – and can be more specifically itemized than the existing question since films/movies can be separated from music, books from newspapers and the like.
a	Films, music	
b	(Electronic) books, magazines, newspapers, e-learning material	
c	Computer software (incl. computer and video games and software upgrades)	Answers to this question can also be obtained.
E4	From whom did you buy or order goods or services for private purpose over the Internet	
a	National sellers	
b	Sellers from other EU countries	
c	Sellers from the rest of the world...	
d	Country of origin of sellers is not known	

Table 2. 2013 Households ICT survey questionnaire and comments about amenability to measurement through the Internet (continued).

MODULE F: e-skills		
F1	Which of the following Internet related activities have you already carried out	Answers to all the 9 categories under this question can be obtained. As has been already explained, using search engines, sending e-mails, creating a Web page, making Internet telephone calls, uploading games and many more activities will be captured.
a	Using a search engine to find information	
b	Sending e-mails with attached files (documents, pictures, etc.)	
c	Posting messages to chatrooms, newsgroups or an online discussion forum	
d	Using the Internet to make telephone calls	
e	Using peer-to-peer file sharing for exchanging movies, music, etc.	
f	Creating a web page	
g	Uploading text, games, images, films or music to websites	
h	Modifying the security settings of internet browsers	
i	None of the above	
F2	Do you judge your current internet skills to be sufficient?	The self-assessment questions F2 and F3 cannot be answered. It is entirely possible, though, for a skills index to be created based on the profile of usage of individuals if desired.
a	To communicate with relatives, friends, colleagues over the internet	
b	To protect your personal data	
c	To protect your private computer from virus or other computer infection	
F3	Do you judge your current computer skills to be sufficient if you would need to take up a new job on the labour market or change your job within a year?	
MODULE G: Socio-demographic background characteristics		
G1	Age	These personal details can only be collected from the individual.
G2	Sex	
G3	Country of birth	
G4	Country of citizenship	
G5	Legal marital status	
G6	De facto marital status	
G7	Educational level	
G8	Employment situation	
G9	Occupation	
G10, G11	Region of residence (NUTS1, NUTS 2)	
G12	Geographical location	
G13	Degree of urbanisation	
G14	Number of members in the household	
G15	of which, number of children under 16	
G16	Household income	

Table 3. 2013 Households ICT survey questionnaire and comments about amenability to measurement through the Internet (continued).

7.3. Enterprises

There is a fundamental difference between individuals and enterprises. The latter are entities with many individuals, ranging from a few to many thousands. “Usage” of the Internet by enterprises thus has a totally different meaning. More to the point, there is a clear distinction between the Internet and the enterprise’s website. Many Internet functions of interest need not (and frequently do not) go through the website. They are all logged, however, in the enterprise’s servers. The latter are then additional sources of information which should be considered for exploitation in the future. The present report however focuses exclusively on and **investigates statistical indicators that can be obtained from enterprises’ websites.**

The 2013 model questionnaire targets enterprises with 10 or more employees and comprises 5 main modules and an additional module at the end for background information used for groupings of enterprises by economic activity (NACE), turnover and employment. This section will proceed to map the information sought against the possibility of obtaining the same (or almost the same) information from the enterprise’s website – for those enterprises that do have one, and with their consent. The variables in the survey are mainly qualitative in nature. In the

tables that follow we present the questions and comments about each one's amenability to measurement from the enterprise's website with automatic methods.

	ENTERPRISE QUESTIONNAIRE	Comments
	Module A: Use of computers and computer networks	
A1	Did your enterprise use computers?	These questions cannot be answered directly from the website.
A2	How many persons employed used computers at least once a week?	
A3	Did any persons employed have remote access to the enterprise's e-mail system, documents or applications (via fixed, mobile or wireless connection to the Internet)?	
	Module B: Access and use of the Internet	
B1	Did your enterprise have access to the Internet?	This question cannot be answered directly from the website. An enterprise can have a website, e.g. hosted by a third party, without itself having access to the Internet.
B2	Did your enterprise have the following types of external connection to the Internet?	These questions cannot be answered directly from the website.
a	DSL connection	
b	Other fixed broadband Internet connection	
c	ISDN connection or dial-up access over normal telephone line	
d	Mobile broadband connection via a portable device using mobile telephone networks (so called 3G or 4G)	
d1	via portable computer	
d2	via other portable devices like Smartphone, PDA phone	
e	Other mobile connection	
B3	What was the maximum contracted download speed of the fastest Internet connection of your enterprise?	
B4	How many persons employed used computers with access to the World Wide Web at least once a week?	
	Mobile connection to the Internet for business use	
B5	Did any persons employed have portable devices provided by the enterprise, that allowed a mobile connection to the Internet for business use?	These questions cannot be answered directly from the website.
B6	How many persons employed had a portable device provided by the enterprise, that allowed a mobile connection to the Internet for business use?	
B6*	Estimate the percentage of the total number of persons employed which had a portable device provided by the enterprise, that allowed a mobile connection to the Internet for business use.	
	Use of a website or home page	
B7	Did your enterprise have a Website or Home Page	These questions can be answered directly from the website.
B8	Did the Website or Home Page have any of the following:	
B8a	Online ordering or reservation or booking, e.g. shopping cart	
B8b	A privacy policy statement, a privacy seal or certification related to website safety	
B8c	Product catalogues or price lists	
B8d	Order tracking available on line	
B8e	Possibility for visitors to customise or design the products	
B8f	Personalised content in the website for regular/repeated visitors	
B8g	Advertisement of open job positions or online job application	
	Use of the Internet in contact with public authorities	
B9	During 2012, did your enterprise use the Internet for interaction with public authorities to:	This can be obtained by an employee's computer connected to the Internet, and therefore server data, but not through the enterprise's website.
B9a	obtain information from public authorities' websites or home pages	
B9b	obtain forms from public authorities' websites or home page, e.g. tax declaration	
B9c	submit completed forms electronically, e.g. forms for customs or VAT declaration	
B9d	declare VAT completely electronically without the need for paper work (including electronic payment, if required)	These questions cannot be answered directly from the website.
B9e	declare social contributions completely electronically without the need for paper work (including electronic payment, if required)	
B10	During 2012, did your enterprise use the Internet for accessing tender documents and specifications in electronic procurement systems of public authorities	
B11	During 2012, did your enterprise use the Internet for offering goods or services in public authorities' electronic procurement systems (eTendering)	These questions cannot be answered directly from the website.
B11a	in your own country	
B11b	in other EU countries	
	Use of Social Media	
B12	In January 2013, did your enterprise use any of the following social media	All this content can be captured from server data too but unlikely to be had from the website – unless the traffic went through there.
B12a	Social networks	
B12b	Enterprise's blog or microblogs	
B12c	Multimedia content sharing Websites	
B12d	Wiki based knowledge sharing tools	
B12e	Did not use any of the above or used them only for posting paid adverts	
B13	In January 2013, did your enterprise use social media to:	
B13a	Develop the enterprise's image or market products (e.g. advertising or launching products etc.	
B13b	Obtain or respond to customer opinions, reviews, questions	
B13c	Involve customers in development or innovation of goods or services	
B13d	Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non governmental organisations, etc.)	
B13e	Recruit employees	
B13f	Exchange views, opinions or knowledge within the enterprise	
B14	Did your enterprise have a formal policy for using social media?	
	MODULE C: Electronic Invoicing	
C1	In January 2013, did your enterprise send electronic invoices?	It is possible that these indicators can be had from the website – with the exception of the part of C1 that refers to e-invoices not suitable for automatic processing, which can in fact be sent by enterprises that do not have a website. From a technical standpoint, reference to "standard structure" implies the use of technologies whose implementation requires adherence to common standards by developers, and which typically refer to an underlying web infrastructure. However, it is also possible that such applications reside on other enterprise servers and may therefore not be retrievable from the website.
C1a	e-invoices in a standard structure suitable for automatic processing, e.g. EDI, UBL, XML	
C1b	Electronic invoices not suitable for automatic processing, e.g. emails, email attachment in PDF format	
C2	In January 2013, did your enterprise receive e-invoices in a standard structure suitable for automatic processing suppliers or customers	

Table 4. 2013 Enterprise ICT survey questionnaire and comments about amenability to measurement from the enterprise's website with automatic methods.

MODULE D: Automatic share of information within the enterprise		
D1	In January 2013, did your enterprise use an ERP software package	As ERP and CRM software are not typically stored on websites, this information cannot be had.
D2	In January 2013, did your enterprise use CRM software to manage:	
D2a	the collection, storing and making available information about customers to various business functions	
D2b	the analysis of information about customers for marketing purposes.	
MODULE E: e-Commerce		
e-Commerce purchases		
Web sales		
E1	During 2012, did your enterprise receive orders for goods or services placed via a website	This information can be obtained from the website. However, the comments in section C for the availability of such data on the web or other enterprise servers apply here too.
E2	Please state the value of the turnover resulting from orders received that were placed via a website (in monetary terms, excluding VAT)	This information cannot be obtained from the website.
E2*	Please indicate an estimate of the percentage of the total turnover resulting from orders received that were placed via a website	
E3	In 2012, did your enterprise receive orders placed via a website by customers located in the following geographic areas	This information can be obtained from the website. However, the comments in section C for the availability of such data on the web or other enterprise servers apply here too.
E3a	Own country	
E3b	Other EU countries	
E3c	Rest of the world	
E4	Please provide a percentage breakdown of the turnover from orders received that were placed via a website by type of customer	This information cannot be obtained from the website.
E5	Did any of the following obstacles limit or prevent your enterprise from selling via a website?	
E5a	The enterprise's goods or services were not suitable for web sales	
E5b	Problems in web sales related to logistics (shipping of goods or delivery of services)	
E5c	Problems in web sales related to payments	
E5d	Problems in web sales related to ICT security or data protection	
E5e	Problems in web sales related to the legal framework	
E5f	The cost of introducing web sales was, or would have been, too high compared to the benefits	
EDI-type sales		
E6	During 2012, did your enterprise receive orders for goods or services placed via EDI-type messages	This information cannot be obtained, so long as the EDI systems are proprietary and not on the Internet – and thus not on the website either.
E7	Please state the value of the turnover resulting from orders received that were placed via EDI-type messages (in monetary terms, excluding VAT)	
E7*	Please indicate an estimate of the percentage of the total turnover resulting from orders received that were placed via EDI-type messages	
E8	In 2012, did your enterprise receive orders placed via EDI-type messages by customers located in the following geographical areas	
E8a	Own country	
E8b	Other EU countries	
E8c	Rest of the world	
e-Commerce purchases		
E9	During 2012, did your enterprise send orders for goods or services via a website or EDI-type messages?	This information can be obtained from the website.
E10	During 2012, did your enterprise place orders for goods or services via a website	This information cannot be obtained from the website – unless EDI systems happen to be through the website.
E11	During 2012, did your enterprise place orders for goods or services via EDI-type messages	
E12	Please indicate the value of orders that were sent electronically in relation to the total purchases' value (in monetary terms, excluding VAT)	
E12*	Please state the value of the purchases resulted from orders placed electronically (in monetary terms, excluding VAT)	
E12**	Please provide an estimate of the percentage of the total purchases that resulted from orders placed electronically	The information that does not refer to EDI-type messages can be obtained from the website. About EDI-type messages please see the comment to questions E11, E12.
E13	In 2012, did your enterprise place orders via a website or EDI-type messages to suppliers located in the following geographic areas?	
E13a	Own country	
E13b	Other EU countries	
E13c	Rest of the world	
Background information		
X1	Main economic activity of the enterprise	This information cannot be expected to be obtained from the website.
X2	Average number of persons employed	
X3	Total purchases of goods and services (in value terms, excluding VAT)	
X4	Total turnover (in value terms, excluding VAT)	

Table 5. 2013 Enterprise ICT survey questionnaire and comments about amenability to measurement from the enterprise's website with automatic methods (continued).

8. Additional Information Society Indicators and Federated Data

With the seemingly unstoppable evolution of ICTs and related applications, it is incumbent on statistical authorities to intensify their efforts and find new ICT-based ways to measure ICT indicators. As well, in conjunction with the Big Data revolution, a parallel avenue of investigation, widely alleged to hold much promise, concerns the measurement of all kinds of indicators, ICT or not, through the use of ICT methods.

ICT-based ways to measure desired ICT indicators was discussed earlier. Federated data, to be discussed below, can become part of the statistical landscape and among other benefits they can

feed into additional ICT indicators too. Yet, in addition to feeding ICT statistics, ICT-based methods, including federated data, can produce statistics for many other statistical domains – in a way that resembles the discussion here.

As already discussed, moving from traditional surveys as instruments of collecting data from households/individuals and enterprises into the direction of the Internet as a data source contains elements of disruptive innovation, in the sense that other established areas of statistical work are affected – in some ways that can be predicted and others that may not be possible at present.

When an approach that will collect data directly from the websites of enterprises for certain variables of interest, such as Private Policy Statement or other company information, is tested and established, the same approach makes possible the capture of non-ICT data that are parts of different statistical domains. An example would be product information and prices, which could feed into the Consumer Price Index (CPI) or Producer Price Index (PPI) programs. Many retail websites (and not only) include detailed specifications of products and their prices. With familiarity, needed information can be targeted and prove a viable alternative to direct price collection. Some initiatives to that effect have already started.

As well, information on job vacancies can be collected. Although not all enterprise websites contain such information, data can be incorporated into existing methodologies in a useful way. In the very least, they can help validate movements in survey data such as labour force surveys.

8.1. Federated data

The view of data as a productive resource, effectively equivalent to intangible capital, has gained momentum with the arrival of Big Data. Such developments are the direct off-spring of ICTs and digitization. Moreover, they are now extending to qualitative data too, but we shall remain within the realm of quantitative data.

While the progression to where we are today has been long in the making and gradual, it appears that we have left behind the inflexion point before which there was relative indifference towards data and after which we encounter changed attitudes – complete with new practices, policies and eventually culture.

Just-in-time inventories started humbly but by now these logistics operations have become almost a science, and it is aided enormously by powerful electronic systems. Big retailers, manufacturers and wholesalers are already there and, by all accounts, derive significant benefits. All that is data driven and with massive amounts of accumulated data analytics are coming of age. They are expanding everywhere, partially fuelled by the Big Data talk. Even the Obama

campaign used such data and techniques to gain a competitive edge during the last presidential campaign²⁶.

In truth, we must acknowledge that there have always been centres where lots of data flowed, either passing through and destroyed or archived. This includes both public administration and private entities. Easy examples to relate are the customs data, with a long history of feeding international trade statistics and telecom companies' data with the detailed billing they had, including details of every long distance call (duration, time, units etc.). However, in the pre-digital age only targeted and limited uses of such data were contemplated. Transcribing paper-based data would have been daunting and hardly efficient. The new thing now is their electronic capturing. Today, such companies and the newer players in that industry, ISPs, have more data than ever before but they are notoriously reluctant to share. . Indeed, the study for DG Connect identified those as good sources but concluded that, for now, these network-centric nodes should not be considered.

Looking ahead, however, this should not be a foregone conclusion. We live in a period in need of new models, where many things happen for the first time and test many of our boundaries. A major one among them concerns the property rights of such data. Already, due to some social media and their policies with personal data debates are raging of how best to handle such matters in the future – even to the extent that the individual is a stakeholder in a stream of revenues. In other words, since we don't know what the future holds, as we are not ready to decide “societally”, we have to be open to what will happen – and perhaps influence it.

In any event, it should not be taken for granted that network-centric data cannot be put to the use of society. After all, many parts of such industries are highly regulated and licensing is involved. Perhaps, for now, of greater interest is not to force an premature Yes or No answer but to ask instead what parts of such accumulated data can become “federated data”, what form they can assume, what tools would be needed and what would be the conditions, privilege and obligations of access.

A public-private-societal dialogue would have to be part of future developments, for what can become a valuable shared resource while safeguarding privacy and confidentiality. After all, many breaches occur daily and societal attitudes are changing. As well, surveys are becoming more problematic by the year. The data generated in the course of our digital interactions with businesses, governments and among ourselves, appropriately re-packaged, can help us all.

One model would be that the statistical value of such data gains prominence and in collaboration with impartial statistical authorities, efforts start to define appropriate levels of aggregation. There would be many uses for such data (including benchmarking of the Internet data or help identify biases in sample survey data). Then, many ICT and non-ICT indicators can be had. For instance, exports and imports of telecom services, business spending on telecoms, the Internet

²⁶ The Economist, Scientists are already helping to shape the Obama campaign, Feb. 11, 2012, <http://www.economist.com/node/21547279>

and the like that are needed for Input-Output tables. Such data, at the level described here, could become federated data.

Other obvious sources for potentially federated data would be utilities, for many similar reasons as before. Granted monopoly or quasi-monopoly rights, they have huge data holdings that can be exceptionally useful. Countries with surveys on energy consumption among houses and buildings would have all the data they need from the detailed billing information systems. Many surveys or parts of them would become redundant and the flow of data would really become a productive resource. Again, at some negotiated level of aggregation, these sources have a high potential to become federated sources.

Major retailers with sophisticated point-of-sale (POS) systems also collect hoards of data. So do social media sites and many other organizations. As the banks share volumes of data, there is social responsibility everywhere. It may well be that citizens, as individual or business respondents, would support such federated data, particularly if it were done in a coordinated way with the involvement, if not the custody, of statistical authorities under an umbrella of confidentiality.

Federated data can support both ICT indicators and other statistical programs. The list that one can think of would be truly enormous. A step-wise approach is needed for the management of the new approaches at the appropriate time. At the outset of such efforts, some thinking that would lead to a taxonomy of such data would be helpful. Identifying the sources of data and arranging them accordingly would not present particular difficulties but its usefulness would be limited. A more functional classification of information holdings would be necessary, through which data sets are used to create thematic categories which, perhaps, would then be populated by data from more than one source. Not only this would lead to more meaningful data, capable of addressing questions for which data are needed rather than questions that can be answered by the available data, but it would also enable confrontation of data from diverse sources leading to improved quality. An illustrative example would be having the records of a smartphone user from the wireless company and then linking it to sales that might have been paid for via the smartphone. Together, they would shed light on general use and e-commerce.

9. Specification of ICT Indicators from the Internet as data source

Two sets of indicators will be compiled and presented. The first set will consist of the combination of those indicators from the questionnaires deemed to be collectable from the Internet for individuals augmented by additional indicators not currently collected. The second set will be the indicators for enterprises found to be collectable from their websites, also including additional indicators that have not been part of the existing survey. Neither set will include Information Society indicators from potentially federated data or other sources. The latter will be assessed in great depth in task 2 of the project.

The indicators presented below can be used to partially replace parts of the questionnaires, as well as extend the data gathering to new areas or provide a more detailed account of the phenomena sought to measure. As per the discussion in section 3, replacement of indicators from the survey through the alternative source will not happen in the literal sense but will be subjected to some adjustments necessitated by the new context. Examples are the time periods in the surveys, the reference to personal rather than business use, etc.

9.1. Individuals

A plethora of indicators can be measured through access to individuals' desktop computers. For the most part, the same can be obtained from those mobile users with smartphones enabled with an Internet-enabled OS. The approach below accommodates both, with appropriate explanations as we move along for warranted adjustments.

A very important point concerns a crucial difference between the survey method and the new Internet-based approach. In the former, questions are specified as clearly as possible ahead of time and the respondents' answers are sought to populate them with data. In the new approach, all kinds of data will be collected and the indicators of interest will be "measured" afterwards. This will also have implications in the programming of the data collecting application that will be used. (These matters are taken on and explained in section 7 of the report). Any number of indicators can be constructed, so long as the collected data can support them. However, as we shall do below, it is advisable to have specified ahead of time a number of key indicators – at least for comparability purposes. For that reason, we stay as close as possible to the familiar content of both the thematic groupings and the actual indicators of the existing surveys and any extensions are easily followed. It must be borne in mind, though, that the indicators that will be presented represent the minimum of what can be constructed with the data obtained.

The discussion will be synoptic, and will be kept at the level of thematic categories of indicators – akin to those in the existing questionnaires. Then, all detailed indicators will be presented in tabular form at the end. (No new numbering is assigned to the indicators in the table, but the detail added is dealt with through the use of extra letters. New indicators for smartphones are denoted by S).

Access to the Internet: As in the outset of the survey, an early indicator is *"Do you have access to the Internet - on your desktop computer/smartphone"*. This would effectively represent a variant of the existing A2 question, avoiding the home in the case of the smartphones, as it is immaterial in mobile connectivity. This question may be redundant, and by design it will amount to 100% among the chosen "respondents". It is included, though, for completeness.

In the case of smartphones, additional detail can be added, such as the telephone calls that go through wireless networks, including WiFi, or use of specific apps which may not be routed through the Internet.

Then, we can continue with indicators on the type of Internet connection used, as in question A3 of the survey – again dropping the home in the case of mobiles.

Use of the Internet: Numerous indicators can be obtained here, shedding light on the frequency, intensity and types of use. This serves as a prime example of the possibilities to answer new questions that were not even asked up to now. The complete data set that will be collected will be capable of answering detailed questions of use by the number of times a day, for every day of the week, and much more. Considering the continuous monitoring of the devices, and perhaps the tracked history, analytical thinking is of the essence to pick useful indicators from all those that will be possible. What is certain is that we cannot be exhaustive in the fine detail that can be had.

Examples of such indicators, along the lines of those in the survey are:

- How many times did you use the Internet
- What was the maximum duration of a session
- The minimum duration
- The average duration
- Total daily duration by weekday day
- Total daily duration during the week-end
- What is the distribution of duration by hour, by day

With all the detailed data on actual usage, questions such as C1 (When did you last use the Internet) become rather unintelligent.

Types of use: Following on the footsteps of C5 (*For which activities did you use the Internet*) we can proceed to include all the survey content and add some (again, with the proviso that we cannot isolate private-purpose use).

Not only we can arrive at indicators for communication but do so in a finer way. For instance, we can separate incoming from outgoing traffic and therefore create individual indicators for sending and receiving e-mails. Moreover, we need not stop there: we can measure the number of e-mails and other communication sessions, and their characteristics (e.g. size, with or without attachments).

All other categories of use in the survey can be had. As one additional example, we can compare again how much more can be obtained that the designers of the questionnaires could not possibly dare contemplate, as the magnitude would be overwhelming. While lists are always curtailed to be manageable, the fact remains that the numbers of activities users carry out are so many and they can be enumerated, sliced and diced in so many numerous ways, are endless. Now, we have the opportunity to zero-in on whatever level of detail the captured data can support.

Rather than being content by asking if the user has downloaded videos or music, we can ask details such as number, size, and time of such downloads.

Analogous comments hold true for the modules on *Use of e-government*, *Use of e-commerce*, and *e-Skills*. They too contain lists of goods, services or activities in the questionnaire, which now can be answered in excruciating detail. Specifically for e-skills, enough detail will be captured to enable the construction of indexes based on the sophistication of use, if needed. Although not a true, direct measurement of e-skills, it will still be superior to the self-assessment now captured through a couple of questions. For example, having a detailed trail of activities carried out by individuals online, such as searching the web, downloading content or applications, participation in social networks, uploading photos or videos, participating in the creation of open source software, manipulating online databases, and any number of activities ranging from the simple to sophisticated would enable the conceptualization and creation of indexes for individual users. It will be possible, for instance, to group usage patterns according to their degree of needed skills and categorize users as novice, experienced or advanced, all the way up a continuum of skills such as creators or wizards.

A host of additional indicators can be added to the above list. As well, more itemization can be had in the existing categories, such as measuring the transmission of pictures, audio, and videos separately. The creation of a typology would be advisable here, perhaps in conjunction with the discussion of skills above. For instance, what is the relationship between skills needed to download songs and music and those involved in sending pictures and video? In the case of smartphones, additional variables that may be of interest include:

- GPS positioning (data that can be used to map out the movements of the individual, inside and outside the country)
- Ringtone downloads
- Specific apps used, such as Google maps, calendars etc.

INDIVIDUAL INDICATORS - Internet based (desktop or smartphone)	
	Access to the Internet
A2	Do you or anyone in your household have access to the Internet
A3	Type of Internet connection
A3a	broadband
A3b	ISDN, dial-up or other narrowband
S1	Do you use a smartphone
S2	Type of Internet connection
S2a	Mobile phone network via a handset
S2b	Mobile phone network via a card or USB key
S3	GPS positioning
	Use of the Internet
C1	How many times during the reference period
C2	Maximum session duration
C5	Minimum session duration
a	Average session duration
b	Total daily duration, weekdays
c	Total daily duration, weekends
d	Distribution of duration by hour, by day
C5	Types of use
a1	Number of e-mails sent
a2	with attachments
a3	Number of e-mails received
a4	with attachments
b	Social networks
c	Reading news
d	Seeking health-related information
e	Information about education, training
f	Finding information about goods or services
g	Downloading software (other than games software)
h	Posting opinions on civic or political issues via websites
i	Taking part in on-line consultations or voting to define civic or political issues

Table 6: ICT data about households / individuals that can be collected from the Internet.

j	Doing an online course (in any subject)
k	Consulting wikis to obtain knowledge on any subject
l	Looking for a job or sending a job application
m	Participating in professional networks
n	Using services related to travel or travel related accommodation
o	Selling of goods or services, e.g. via auctions
p	Telephoning over the internet / video calls (via webcam) over the internet
q	Internet Banking
	Use of e-government
D1	Did you contact or interact with public authorities or public services over the internet
a	Income tax declaration
b	Downloading official forms
c	Submitting completed forms
D2	Did you use websites of public authorities or public services for any of the following
a	Income tax declaration
b	Claiming social security benefits
c	Requesting personal documents (passport, ID card or driver's licence) or certificates
d	Public libraries (availability of catalogues, search tools)
e	Enrolment in higher education or university
f	Notification of change of address
	Use of e-commerce
E1	When did you last buy or order goods or services for private use over the Internet
E2	What types of goods or services did you buy or order over the Internet
a	Food or groceries
b	Household goods (e.g. furniture, toys, etc)
c	Medicine
d	Films, music
e	Books, magazines, newspapers (including e-books)...
f	e-learning material
g	Clothes, sports goods
h	Video games software and -upgrades
i	Other computer software and -upgrades
j	Computer hardware
k	Electronic equipment (incl. cameras)
l	Telecommunication services
m	Share purchases, insurance policies and other financial services
n	Holiday accommodation (hotel etc.)...
o	Other travel arrangements (transport tickets, carhire, etc.)
p	Tickets for events
q	Other

E3	Were any of the following products that you bought or ordered over the Internet downloaded or accessed from websites rather than delivered by post etc
a1	Films
a2	size
a3	Music
a4	size
b	(Electronic) books, magazines, newspapers, e-learning material
c	Computer software (incl. computer and video games and software upgrades)
E4	From whom did you buy or order goods or services for private purpose over the Internet
a	National sellers
b	Sellers from other EU countries
c	Sellers from the rest of the world...
d	Country of origin of sellers is not known
	e-skills
F1	Which of the following Internet related activities have you already carried out
a1	Using a search engine to find information
a2	number of times
c	Posting messages to chatrooms, newsgroups or an online discussion forum
d	Using the Internet to make telephone calls
e	Using peer-to-peer file sharing for exchanging movies, music, etc.
f	Creating a web page
g	Uploading text, games, images, films or music to websites
h	Modifying the security settings of internet browsers
i	None of the above

Table 7: ICT data about households / individuals that can be collected from the Internet (continued).

9.2. Enterprises

A two-stage approach for indicators is recommended here. First, websites of enterprises are identified and captured, and then they are mined for specific questions.

Stage 1: Websites or Home Pages

Currently, the Eurostat survey asks companies whether they have a website, and these data are very useful as a baseline indicator of progress at that high level. Eventually, they are used as the springboard to explore the sophistication and functionalities offered by businesses, particularly when it comes to e-commerce. However, these are data based on the relatively small sample of each national survey. They serve their intended purpose, but they do not contribute to the enlargement of existing statistical infrastructure. This becomes more important now, if websites become part of statistical operations, and will be explained next.

Visiting websites of enterprises for ICT or non-ICT data collection will have to be a methodically-organized and executed effort. Like all statistical efforts, it must be guided by a

population frame. Today, under the specifications of the survey, we know that X% of enterprises in a country have a website, but we do not really know who they are and how to make use of this information for additional statistical purposes. This is so because these key pieces of information are not captured as additional fields in the national business registers (BRs) in a way that would enhance them, and be used for more benefits later. Their ready availability could assist data quality efforts for existing surveys, as they can provide an additional source for verification of survey responses. They can also reduce follow-up time and costs for non-respondents or for edits/imputation in incomplete questionnaires. Moreover, they can be potentially used as a frame of their own for surveys that may target that very population alone. And, of course, they will be indispensable in the event of data collection through web scraping or of the type discussed in this report.

The message is not merely that the URLs for the enterprises found in the survey to have a website should be collected and captured – neither of which is the case. Transitioning such a key indicator to automated ICT collection, there is no reason to restrict it to the survey samples. They can be extended to the whole BR, and carried out either on an industry-by-industry approach (NACE) or by firm size bands or any other means (alphabetically) etc.

Moreover, rather than being tied to the annual survey, such data can be produced with any frequency desired. Yet, upfront efforts will progressively become smaller. For example, after the first cut, the exercise may be repeated only for those who did not have (or were not found to have) a website in the first instance. (Occasionally, a sample of those who were found to have one can also be taken, to check for continuity). With experience, maintenance and update of this new BR field will be subject to similar procedures and operations as any other field – say, in the event of a firm's birth or death through bankruptcy or otherwise. In the end, a new BR field will eventually exist for all those enterprises with a website, complete with their URL/s.

Then, not only this key indicator can be disseminated and serve the policy and business needs of today, but additionally it can be used for many other purposes sooner or later, including website-based surveys as those suggested in this report.

This approach, with appropriate modifications, can be extended to include company contact information, including names etc. This could help verify existing BR information, and potentially add yet more fields to national BRs, which could be valuable in all kinds of future surveys, especially as existing surveys increasingly migrate to electronic questionnaires, and contact e-mails are expensive and time-consuming to obtain. In fact, various contacts can be obtained – from management to financial or sales, which may prove more appropriate depending on the focus of future surveys.

Stage 2 – Website information

Content that would replace existing questions, with the earlier provisos, can be re-grouped, as well as new content can be added, as follows (new questions, not in the questionnaire are denoted by N):

Tombstone information: At first a few indicators will be collected that support the identification and key vitals of the enterprise, as per the discussion in Stage 1. The URL of the enterprise's website will be collected, as well as contact information, such as e-mail. (In the future, it may be interesting to distinguish between enterprises that have invested in their own domain name and those that have websites hosted by third parties, e.g. by aggregator sites. Typically, these can be differentiated from the URLs and they may be indicative of how smaller enterprises usher in the Information Society). In this category, we can include an additional indicator capturing the languages offered by the website.

In-depth information: Next we can look at the websites last update and traffic, and create categorical indicators along the lines of the questionnaire. We can capture whether or not the enterprise has a Privacy Policy Statement, a registration facility for frequent/repeat visitors, a sitemap to facilitate navigation, or a "last updated" indicator – which, usually, denotes fresh rather than stale content. Each of them may have their own significance for certain users.

Product and price information: Another category can capture indicators based on information that websites typically do a very good job in conveying to visitors. That is, information about their goods and/or services, including detailed specifications. These, then, can be complemented with the prices, if available. The latter is expected to be a sub-set of the products, as experience has shown that many companies, especially in services, while they have detailed descriptions of their offerings they do not always list their prices. As well, if the extent that websites permit visitors to design or customize their products, that too can be turned into an indicator.

E-commerce: Indicators can be constructed that would illuminate e-commerce, as per the questionnaire. Starting with whether or not the enterprise has the capability of receiving orders or reservations online (e.g. shopping cart), indicators can progressively move to whether orders for goods or services are actually received, their volume, from which geography of interest, as well as whether the online capability to track the status of orders is offered. (As discussed in section 7, depending on the technical set-up of the enterprise, it is likely that detailed indicators of online orders can only be had through access to other servers than those hosting the website).

Employment: We can find out if the enterprise lists job openings, and how many – to the extent that such information is contained on the website.

Social media: Finally, we can have some indicators related to audio-visual content and social media. These are not of the type in the existing questionnaire, which asks for the enterprise's own use of such social media but concern instead linkages offered from the website. They can

include the availability of audio or video content on the website, links to social networks or blogs, content linked to multimedia content sharing sites, Wikis and wiki-based sharing tools. Moreover, these can be complemented by information on marketing strategies and the extent to which such strategies integrate the Internet with other efforts. The sets of content in the Eurostat business survey that refers to the benefits from the use of ICTs or impediments/barriers to use are not conducive to online collection for obvious reasons. In fact, the benefits which are part of the quest for impacts are much better dealt with analytically through linkages to extraneous data sets rather than asking the subjective opinion of business.

ENTERPRISE INDICATORS - Website based	
	Tobstone information
B1	Did your enterprise have access to the Internet
B7	Did your enterprise have a Website or Home Page
N1	Contact information
N2	Language options
N2a	National
N2b	Other (specify)
	In-depth information
B8	Did the Website or Home Page have any of the following:
N3	Last updated date
B8b	A privacy policy statement, a privacy seal or certification related to website safety
N4	Registration facility
B8f	Personalised content in the website for regular/repeated visitors
N5	Site map
N6	Number of visitors
	Product and price information
B8b	Product catalogues
B8c	Price lists
B8e	Possibility for visitors to customise or design the products
	e-Commerce
B8a	Online ordering or reservation or booking, e.g. shopping cart
E1	Did your enterprise <i>receive</i> orders for goods or services placed via a website
N7	How many
E3	In which geographic areas
E3a	Own country
E3b	Other EU countries
E3c	Rest of the world
B8d	Order tracking available on line
	Employment
B8g	Advertisement of open job positions or online job application
N8	How many
	Social networks
N9	Does the enterprise's website have links to:
N9	Multimedia content (audio, videos etc)
N11	Links to social networks or blogs (Facebook, LinkedIn, Yammer, Twitter, etc)
N12	Content linked to multimedia sharing sites (You Tube, Flickr, etc)
N13	Wikis and wiki-based sharing tools

Table 8: ICT data about enterprises that can be collected from their website.

9.3. Attributes of indicators

Indicators that are collected objectively from digital footprints and relate to usage are in principle expected to produce higher quality data than asking people or businesses. Issues with recall are well documented in such surveys, and they become particularly relevant when detailed traffic patterns and numerous transactions are performed – especially in situations when the novelty has worn off and people go about the business of navigating and transacting on the Internet in a rather instinctive than calculated manner. (Even impulse buying is happening).

In any event, considering the novelty of the proposed approach, a discussion of desired quality properties of indicators according to the ESS framework follows.

Relevance: As far as meeting user needs, and to the extent that the Internet-based indicators will complement those from the surveys, more satisfaction should be expected. Users will have access to more detailed data unavailable up to now.

Accuracy: Many of the indicators produced with the new approach will be more accurate than survey-based indicators. This is particularly true for usage indicators, where respondent recall is a known concern. The objective nature of the indicators based on data recorded as digital footprints makes them quite accurate.

At the same time, new biases may creep in related to the pattern of users' refusals to participate or perhaps modified behaviour during the reference period. This would represent non-response and will be quantified. How such biases would compare with the existing biases from non-response in the surveys now cannot be assessed a priori, as the non-respondent populations in the two exercises cannot be assumed to be similar in their characteristics. For instance, the smartphone non-respondents may live busier lives whereas the paper questionnaire ones may be poorer. (This can be dealt with to a certain extent as up to now, that is, with analysis of non-response in the best way possible, say, through the profiles available in the frame). Over time, experience and more data accumulated will help. These concerns are not independent of the exact technical intervention. To that effect, the pilot tests will help assess such biases better.

Coherence and comparability: With respect to the underlying concepts, the reference population, and the coverage of activities, there is no reason to expect the Internet-based indicators to differ from those derived from the surveys. For the most part, they represent standalone outputs and they are not integrated in composite or aggregate statistics.

As for comparability, much will depend on whether enough member states adapt the new approach. As well, it may be affected by differences in participation across different countries perhaps due to cultural attitudes. For now, all that can be said is that this should improve over time. Another aspect of comparability that may be affected concerns the reference periods of the new indicators, both across countries and against the existing indicators.

Accessibility and clarity: It is expected that the new indicators will be as accessible as the existing ones through the dissemination channels of Eurostat and the NSIs. Clarity will be high as users will be able to relate to such indicators since for many they represent parts of their daily reality. Moreover, clear definitions and metadata will be made available.

Timeliness and punctuality: Overall, timeliness is expected to be superior to that of surveys. Online collection is subject to much smaller time lags. The analytical and data-crunching activities that must follow for the construction of the indicators will absorb time to set up, but they are expected to become largely automated following the early iterations – with the tasks and procedures eventually becoming routine.

The indicator fiches are in Annex 1.

10. Methodology and Related Matters

We now turn to a discussion of how the Internet-based approaches for data collection discussed above can be implemented in practice. This will be presented as an experimental methodology, and emphasis will be placed on the first time. With accumulated experience, improvements can then be introduced.

It must be emphasized upfront, that several possible methods exist for implementation of the new approach. Many aspects, including financial, will have to be factored in to make a choice. One method is to run the Internet-based collection as a partial replacement of questionnaire content, and run the surveys with reduced content. Another approach, with much merit, is to run the new approach in parallel with, and in addition to, the existing surveys at least once in a cross-section of countries. The overlap through such a parallel run will provide both valuable data and experiences to help identify viable substitutions, and therefore replacements for survey content, and intelligence on the quality and the impact of the change.

10.1. Individuals

The methodology discussed below follows the previous analysis so far in this report, and covers desktops and tablets but also extends to smartphones – with additional explanations as warranted.

To begin with, the methodology is not a complete and radical departure from the existing statistical survey practices. A phased transition is instead proposed. This can be accomplished by starting with traditional representative sampling of the type NSIs are accustomed too for the annual community survey. A random sample is drawn from the frame of households, which will be used for the surveys as is the case every year.

Then we proceed as usually. In a CATI (or telephone survey), a call is made for the early part of the questionnaire that concerns the household. In the process, households without Internet at home are excluded, whereas for the rest an individual is chosen to respond to the remainder of the questionnaire. Whether that happens on the spot or through a subsequent call, contact is established with the individual respondent. There is no reason why we should opt to recruit less than all those would-be respondents – unless, perhaps, in the case when the pilot new approach is carried out in parallel with the full survey, in which case a sub-sample may suffice.

It is at that time that a well-prepared script is read to the respondent, explaining the exercise and seeking their consent to participate by giving access to their desktop (or tablet) for a week (or another specified time period). In the case of a parallel run with the traditional survey, this will offer valuable benchmarking data to gauge the comparability of the two approaches. In the case of running the new approach as a partial replacement of the existing questionnaire, no such

benchmarking would be possible. The same method is applicable to smartphones.²⁷ As explained in earlier sections, this exercise focuses on securing access to an individual's desktop/tablet computer or smartphone. For individuals using multiple devices, this will not capture their entire use. So, in the early phases of such migration it is advisable that we ask them to indicate which device they mainly use, perhaps with a proportional allocation, and this is factored in the results. At later stages, it is conceivable that we ask for access to all their devices as usage patterns may differ across them. (This too is subject to caveats but as has already been discussed, and will be discussed again in the next section, is beyond the purview of this report).

The same basic method can be applied in the event of a mail-out/mail-back national survey. An introductory letter can explain the approach and seek consent, as well as provide a telephone number or a website. However, a telephone call becomes inevitable at some point. Although we do not really have frames of individuals, household frames typically do contain telephone numbers. At the time of the contact, individuals may also be asked for their own smartphone number and e-mail.

Clearly, ahead of the survey, a program has been created that the consenting user/respondent can download from a specified website. This app will be programmed to monitor Web navigation and Internet usage on the desktop (tablet). It can also be programmed to track other traffic too. The same applies to smartphones – but in addition to the OS it can be extended to other function, such as calls and GPS.

Then, the application can be downloaded (or other technical arrangement can be made). Perhaps, depending on the modalities of collection in individual countries, it can be accompanied by the questionnaire in electronic format with only one link to the same website. (The default option remains to complete the rest of the questionnaire the traditional way, e.g. telephone). Rather than have the individual transmit the stored data, the app will transmit through a router to a back-end server for storage, and later analysis, at desired set intervals – in real-time, daily or at the end of the week.

It is important that the individual has the power to turn it off, either temporarily or permanently. In such cases, these intervals will be known and decisions similar to those made in traditional surveys with regards to partial non-response will have to be made. The entire set of data is either discarded or, if it's so deemed, imputation may be made for missing data.

Back-end work can assume different forms, from investing in an integrating application upfront, to processing the two datasets (Internet-based and survey) separately, with an analytical merger later. Experience will be needed to accumulate in order to understand the pattern of usage in a

²⁷ However, it is important for the existing methodology that even if everyone has a smartphone we only ask those who would have responded to the survey. That is, if the individual selected to respond to the Internet usage part of the questionnaire would have been ineligible to continue because he/she was a non-user (e.g. no home computer and no usage elsewhere) he/she should also be excluded here (e.g. the smartphone may be used only for calls. This way, we will not put the results of the survey in jeopardy by “contaminating” the sample.

way that will point to an optimal collection period and frequency. In any event, the back-end now becomes enormously important. From the time the data arrive at the designated server what happens? An approach must be in place to analyze the data in the desired way. This is discussed more in the next section.

Additional and important aspects that must also be examined in time:

It is technically possible, always with the consent of the individual, to gather information not of the “flow” type during the period of observation but also of the “stock” variety, that is, historical information, say, for the whole year. Such issues should be further discussed in the future.

Additional user demands for information can be satisfied, particularly through smartphones. It will be easier than up to now to track usage by individuals under the age of 16. In that case, when it arrives, the consent of the individual and the parent/guardian will be needed.

10.2. Enterprises

The two-stage approach outlined earlier can be implemented as follows:

For the community survey, a national sample is drawn and stratified to represent the targeted population of enterprises (NACE, size etc.). In the first stage, the entire sample will be subject to automated searches for websites, and the collection of URLs. This can be surely helped by the creation of a specific application that can accommodate the national samples. Depending on the nature and sophistication of such an application, it may accommodate the entire sample at once or it may be necessary to break it in manageable pieces (say, by NACE). Other than that, this exercise is straightforward. The collected URLs should then be inputted and maintained in the national BRs.

The first time, additional spot checks will be advisable as quality assurance measures. First, any outside knowledge must be brought to bear to avoid and/or complement the automated procedure – for type 1 and type 2 errors. For example, any available information from previous surveys or from subject-matter knowledge should be used to verify that enterprises for which URLs were not found, indeed do not have any – this can be done even manually, particularly if this happens to be the case for some large enterprises which may be suspect. Second, for those enterprise for whom a website was found, a small sample check could verify that they actually exist. (While no such experience exists, validation rules can be developed. For instance, additional searches can be performed through search engines using the enterprise name available in national BRs or with certain instructions, such as ‘name’ .com or .org etc.).

Up to this point, this is one indicator for which no prior arrangements with the business may be needed – e.g. out of concern for privacy, confidentiality or legal or social acceptance. In other words, business with a website is akin to having an ad or a sign in the street – they want to be found, and in fact they advertise to be found and pay for it rather than hide.

Then, the surveys proceed in the known way for those without a website, and with a modified shorter questionnaires for those with – with content equal to the original minus what will be collected from the websites (unless a parallel run of both approaches takes place with the questionnaire intact). As in the case of individuals, a parallel run of the two approaches will make benchmarking possible.

During stage one, e-mail contact information can be obtained too. This will be useful for stage two, but almost certainly telephone calls will be needed too (during which a better e-mail contact can be asked for).

For stage two, several options exist for collecting and processing the data. For collection, one option is to develop an application which can be installed on the website of the enterprise, with prior consent. The data will be accumulated there and either at the end of the period or periodically they can be routed automatically to a specified server. A good option - even if not needed technically - is to obtain permission from the enterprise to send a crawler or web harvester that can scrape all available information over a pre-specified period.

It must be realized that some of the content of websites does not change much for long periods of time. For instance, the presence of a Privacy Policy Statement is not expected to change often, and therefore it may not be desirable to continue collecting such information. Price information, on the other hand, may well be subject to frequent revisions. With experience, we can calibrate the software applications to focus in areas of interest, minimizing the effort.

In the event of a future exercise to also go after some of the servers of the enterprise, similar techniques can be used – simply have the enterprise give access to the server or even transmit old captured log files. Statistics for orders received, for example, will likely be stored in separate servers than the website.

10.3. The pilot

Obviously, this potential approach to surveying and data compilation does not need to be part of the pilots for smartphones and websites that will be undertaken in this project. There is no reason to add a methodological layer to complicate or confuse the real testing sought at this stage, which is the collection of some of the indicators outlined above – in the feasibility sense, and learning from the experience. What works, what needs to be improved, as well as what the back-end entails, will offer feedback to the critical issues of analysis, quality control and dissemination. A reasonably selected cross-sectional group of individuals and enterprises would be fine, without the need to be linked to the actual surveys. If, of course, it becomes possible to draw the samples of individuals and enterprises from the actual recent samples used in national surveys, it would generate valuable benchmarking information and would be ideal.

Inevitably, there will be unforeseen snags during the early phases of implementing the new approach. Whether related to technical peculiarities, particular customization of the IT

infrastructure or other reason, issues will be encountered. These should be dealt with as they arise.

11. Software Tools which can be used for data collection

Measurement for ICT indicators have been presented in the previous “*Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering*”. In this study the collection of tools had been classified in three separate categories namely, i) user centric, ii) site centric and iii) network centric. This classification is still valid in light of the moving target of Internet based statistics with respect to indicators. The moving target of indicators relates to advances of current technology and to the emerging technology of social media and linked data.

Network Centric

With respect to network centric techniques in order to cope with the ever growing traffic size, it is advisable to exploit sampling techniques such as netflow²⁸ from Cisco or the sFlow²⁹. Both of these techniques are supported from network-gear vendors in their mid-to high end products and as a consequence they can scale to tens of gigabits per second. A typical characteristic of the aforementioned techniques is that they produced a stream of information corresponding to a sampled version of traffic. The stream is collected by appropriate components^{30,31}. In that sense, the stream of information can be processed in order to produce useful statistics such as entropy, average, min and max per IP address. This stream of information can be safely considered as Bid Data stream.

In terms of public statistics, Netflow or/and sFlow methods can be pipelined to anonymization techniques^{32,33} in order to produce public anonymized trends of indicators under interest.

Another interesting emerging technique for traffic flow estimation already tested in high speed backbones related directly with Big Data techniques using probabilistic cardinality counting³⁴. This work shows that usable statistics can be obtained in almost real-time (once every 10 seconds) with low average relative error (less than 5%). with low processing, storage and communication overhead for a OC-192 (10 Gbps) line speed. There is no need for changes the current routing infrastructure of most ISPs.

²⁸ http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html

²⁹ <http://www.sflow.com>

³⁰ <http://www.ntop.org>

³¹ www.inmon.com/technology/sflowTools.php

³² <http://pages.uoregon.edu/joe/ipv6-mask.html>

³³ <http://www.cs.jhu.edu/~coulls/USENIX07.pdf>

³⁴ <http://gridsec.usc.edu/files/tr/tr-2005-12.pdf>

A possibly alternative method for obtaining indicators directly from ISPs is to feed gain data from the radius³⁵ [8] accounting system. The feed should correspond to aggregated traffic of adjacent time window intervals indicating number of users, incoming and outgoing traffic.

User centric

With respect to user centric method the previous study has indicated the pros and cons of those method. New advances in the browser extensions and add-ons indicate new potentials. For instance the netusage³⁶ allows the user to retrieve online statistics of his connection from his broadband provider indicating the total time and MB used. It seems that this effort has mobilized service providers towards a common xml based usage retrieval scheme. Similar efforts such as Datafox³⁷ are also popular in India where limited broadband usage is a common practice.

With respect to user based epidemics for internet usage there are half of dozen of solutions.

8AWeek.com³⁸ - This is a downloadable toolbar that monitors a user's surfing habits and shows the sites the a user visitw most. It also let the user control the time he spend on different sites.

GetMeeTimer.com³⁹ - MeeTimer logs a user's web activity with outputs like the following one.

³⁵ <http://tools.ietf.org/html/rfc2865>

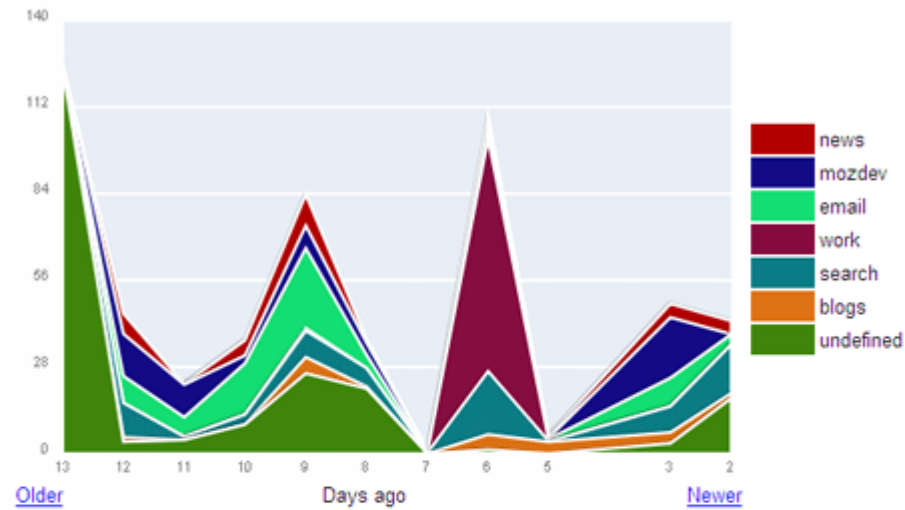
³⁶ <http://netusage.iau5.com/>

³⁷ <http://thegoan.com/datafox/>

³⁸ <http://8aweek.com/>

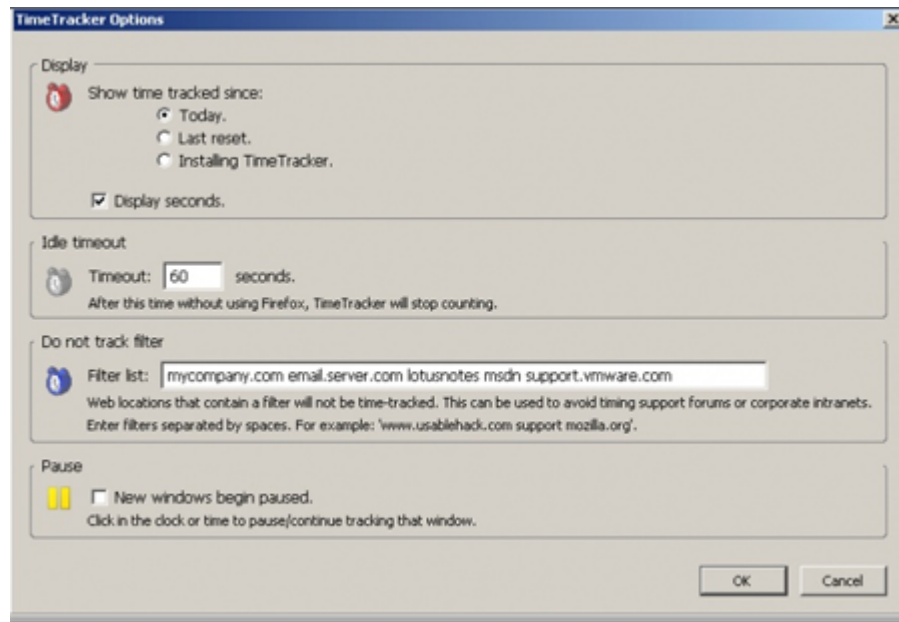
³⁹ <http://getmeetimer.com/>

Minutes you have wasted in the past:



Timer⁴⁰ - Is a very simple extension that counts the time a users starts until he signals it to stop it.

⁴⁰ <https://addons.mozilla.org/en-US/firefox/addon/4354>



TimeTracker⁴¹ - Is an extension that will allow a user to monitor his browsing history..


⁴¹ <https://addons.mozilla.org/en-US/firefox/addon/1887>

12. References


- Bauwens, M. (2006). The political economy of peer production. *Post-autistic economics review*, 37.
- Benkler, Y. (2007). *The wealth of networks: How social production transforms markets and freedom. Information Economics and Policy* (Vol. 19, pp. 278–282). doi:10.1016/j.infoecopol.2007.03.001
- Berners-Lee, T. (2006). Welcome to the Semantic Web. *The Economist - The World in 2007*. Retrieved from <http://www.neurophenomics.info/docs/semanticweb.pdf>
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1. doi:10.4018/jswis.2009081901
- Capadisli, S., Auer, S., & Ngomo, A. (2013). Linked SDMX Data. *semantic-web-journal.net*. Retrieved from <http://www.semantic-web-journal.net/system/files/swj454.pdf>
- Cyganiak, R., Reynolds, D., & Tennison, J. (2012). The rdf data cube vocabulary. Retrieved July 1, 2013, from <http://publishing-statistical-data.googlecode.com/svn/trunk/specs/>
- Dialogic. (2012). *Internet as data source* (p. 243). Luxembourg.
- Daily Mail (2013), <http://www.dailymail.co.uk/sciencetech/article-2340714/The-Secret-Life-Cat-What-mischievous-moggies-gets-owners-backs.html>
- EC (2009), i2010 High Level Group, Benchmarking Digital Europe 2011-2015: a conceptual framework, European Commission, October, 2009
- The Economist (2013), <http://www.economist.com/news/business/21567403-techniques-presidents-election-campaigns-have-spawned-one-lot-young-firms-obama>
- EU (2012), Internet as data source, Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering, FINAL REPORT, A study prepared for the European Commission DG Communications Networks, Content & Technology SMART 2010/30
- EU (2013a), Digital Agenda for Europe, <http://ec.europa.eu/digital-agenda/digital-agenda-europe>
- EU (2013b), Digital Agenda for Europe, <http://ec.europa.eu/digital-agenda/about-our-goals>
- Eurostat (2012a), European Union survey on ICT usage in households and by individuals, 2013 Eurostat Model Questionnaire (version 3.4)
- Eurostat (2012b), COMMUNITY SURVEY ON ICT USAGE AND E-COMMERCE IN ENTERPRISES 2013 Model Questionnaire version 1.1
- Eurostat (2012c), Methodological Manual For statistics on the Information Society, Survey year 2012, v1.1
- Evans, D. (2003). Some empirical aspects of multi-sided platform industries. *Review of Network Economics*, 2(3), 1. doi:10.2202/1446-9022.1026
- Evans, D., & Schmalensee, R. (2007). Two-Sided Platforms The Industrial Organization of Markets with Two-Sided Platforms. *Competition Policy International*, 3(1).


- Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., & Khan, A. (2012). WHAT DOES “BIG DATA” MEAN FOR OFFICIAL STATISTICS? Retrieved from <http://goo.gl/LfGIM>
- Hendler, J. (2009). Web 3.0 Emerging. *Computer*, 42(1), 111–113. doi:10.1523/JNEUROSCI.0382-11.2011
- Information Society: ICT impact assessment by linking data from different sources http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/documents/Tab/ICT_IMPACTS_FINAL_REPORT_V2.pdf
- International Partnership on Measuring ICT for development, Core ICT Indicators, http://new.unctad.org/Documents/Core%20Indicators/Core_Indicators_English_2010.pdf
- Kumar, V. (2009). Why do Consumers Contribute to Connected Goods ? A Dynamic Game of Competition and Cooperation in Social Networks. *Social Networks*, 1–53.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., et al. (2009). Social science. Computational social science. *Science (New York, N.Y.)*, 323(5915), 721–3.
- LAWA (2012) Seventh Research Framework Program "Longitudinal Analytics of Web Archive data", <http://www.lawa-project.eu>
- OECD (2005), Guide to measuring the information society, p. 10
- OECD (2009), Guide to measuring the information society.
- Quah, D. (2003). *Digital goods and the new economy. New Economy*. Centre for Economic Policy Research.
- Saltzer, J., Reed, D., & Clark, D. (1984). End-to-end arguments in system design. *ACM Transactions on Computer* Retrieved from <http://dl.acm.org/citation.cfm?id=357402>
- Stephen, A. T., & Toubia, O. (2010). Deriving Value from Social Commerce Networks. *Journal of Marketing Research*, 47(2), 215–228.
- Tapscott, D., & Williams, A. D. (2008). *Wikinomics: How mass collaboration changes everything*. Portfolio Trade.
- UNECE (2013), 'What does big data mean for official statistics?', www1.unece.org/.../Big+Data+HLG+Final+Published+Version.docx
- Vafopoulos, M. (2011a). The Web economy: goods, users, models and policies. *Foundations and Trends® in Web Science*, 3(1-2), 1–136. doi:<http://dx.doi.org/10.1561/18000000015>
- Vafopoulos, M. (2011b). A Framework for Linked Data Business Models. *15th Panhellenic Conference on Informatics (PCI)* (pp. 95–99).
- Vafopoulos, Michalis. (2012). Being, Space, and Time on the Web. *Metaphilosophy*, 43(4), 405–425. doi:10.1111/j.1467-9973.2012.01762.x
- Zittrain, J. (2008). *The future of Internet and how to stop it*. Yale University Press.


13. Annex



<p>Dataset: Type of Internet connection</p> <p>Data source: Possible new ICT-based data collection based on users' home computers or smartphones</p>
<p>Description: Measurement of the speed of the connection of devices connected to the Internet. For home computers it differentiates between broadband (always-on) access and dial-up (including ISDN). For mobiles smartphones it differentiates by phone network, (3G, e.g. UMTS, card or USB key)</p>
<p>Key indicator(s) included in the dataset:</p> <ul style="list-style-type: none"> Percent of home connections using broadband Percent of home connections using narrowband Percent of smartphones connected via broadband (at least 3G) Percent of smartphones connected via narrowband (less than 3G, ISDN, normal lines)
<p>Policy relevance: In addition to overall connectivity measures, the quality of the connection is closely related to the type of value added services that can be consumed by users, particularly applications optimized for broadband (e.g. audiovisual streaming). This has also business relevance as broadband connections enable the production of higher value-added applications which can be accessible by users. In addition, the speed of connections relates to the deployment of advanced networks and has price implications</p>
<p>Statistical population: Those individuals with home access to the Internet in the case of computers or those individuals with smartphones. The population of those individual refers to those that accept to participate to this potential data collection</p> <p>Unit of measurement: While broadband can be measured using specific speeds of connection involved (including differentiation between uploads and downloads), the unit in this indicator will be the number of individuals by each connection type expressed as a percentage of all those participating</p>
<p>Main concepts used: Speed of Internet traffic</p>
<p>Other concepts: Particular technologies used, such as 3G or 4G networks, ISDN, etc.</p>
<p>Calculation method: Counts of individuals with devices connected to the Internet through a specific type (broadband or narrowband) divided by all participating individuals and multiplied by 100</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the household/individual survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Many desirable breakdowns can be constructed based on the socio-demographic profile of users, e.g. by age, gender, and income</p>


Reference period: The observation week during which individuals agree to participate (or any period chosen for the new collection method)



<p>Dataset: Frequency and intensity of Internet use</p> <p>Data source: Possible new ICT-based data collection based on users' home computers or smartphones</p>
<p>Description: Number of individual Internet sessions by users in the course of a specified time period (e.g. a week), and time spent on the Internet (or other uses in the case of smartphones)</p>
<p>Key indicator(s) included in the dataset:</p> <ul style="list-style-type: none"> Number of times using the Internet within period Total time of usage within period Maximum session duration Minimum session duration Average session duration Distribution of usage (sessions and time) by time of day and day of the week
<p>Policy relevance: Captures the integration of the Internet (or the use of smartphones) in people's daily lives. Frequency and intensity of use denotes changes in people's behavior towards the technologies and is a precursor to impacts. Moreover, they are related to the effort invested to achieve certain outcomes depending on the speeds of networks</p>
<p>Statistical population: Those individuals with home access to the Internet in the case of computers or those individuals with smartphones. The population of those individual refers to those that accept to participate to this potential data collection</p>
<p>Unit of measurement: Counts of times Internet sessions occurred per time period, and cumulative time of use per time period</p>
<p>Main concepts used: Frequency and intensity of Internet (or smartphone) use</p>
<p>Other concepts: Use of non-Internet apps on smartphones (possibly)</p>
<p>Calculation method: Counts of sessions and cumulative time will be arrived at through straight addition of the observations in the collected data. Indicators of maximum and minimum duration of sessions will also be derived in a straightforward manner from the data, after ranking. Distributions of sessions and time used by time of day or day of the week will be arrived at through grouping by the relevant time period</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the household/individual survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Many desirable breakdowns can be constructed based on the socio-demographic profile of users, e.g. by age, gender, and income</p>
<p>Reference period: The observation week during which individuals agree to participate (or any period chosen for the new collection method)</p>



<p>Dataset: Types of Internet use</p> <p>Data source: Possible new ICT-based data collection based on users' home computers or smartphones</p>
<p>Description: What activities individuals undertake on the Internet. Usage of the Internet involves an endless list of tasks that can be carried out, some of a repetitive or routine nature and others infrequently or occasionally. Many of those activities are captured by this dataset</p>
<p>Key indicator(s) included in the dataset:</p> <p>Number of e-mails (sent and received)</p> <p>Proportion of individuals:</p> <ul style="list-style-type: none"> Reading news Participating in social networks Seeking health-related information (or education-related) Downloading software Looking for a job Posting opinions Etc...
<p>Policy relevance: The types of Internet use, and their evolution over time, are the defining features of the new technologies. They matter enormously in understanding the demand side of the Information Society, the uptake and the possible developments of new applications both from the private and the public sectors. Moreover, types of use are critical to the understanding of impacts.</p>
<p>Statistical population: Those individuals with home access to the Internet in the case of computers or those individuals with smartphones in that case</p> <p>Unit of measurement: Percent of individual undertaking each task on the Internet</p>
<p>Main concepts used: Various types of Internet use</p> <p>Other concepts:</p>
<p>Calculation method: Number of individuals using one of the listed categories of Internet use divided by the total number of participating individuals and multiplied by 100.</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the household/individual survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Any breakdown desirable based on socio-demographic profile of users, e.g. by age, gender, and income</p>
<p>Reference period: The week of the observation (or any chosen period)</p>


Dataset: Use of e-government
Data source: Possible new ICT-based data collection based on users' home computers or smartphones
Description: This dataset addresses specifically the interactions of individuals with their governments through the Internet
<p>Key indicator(s) included in the dataset:</p> <p>Contact or interaction with public authorities for:</p> <ul style="list-style-type: none"> Income tax declarations Downloading official forms Submitting completed forms <p>Using websites of public authorities or public services for:</p> <ul style="list-style-type: none"> Income tax declarations Claiming of social security benefits Requesting personal documents (e.g. passport, ID etc.) Etc...
<p>Policy relevance: Captures the interactions that citizens have with their governments, and informs both the level of e-government uptake and the availability of services offered. Such indicators help assess over time the progress made in facilitating the provision of public services, and continuing to design user-friendly applications</p>
<p>Statistical population: Those individuals with home access to the Internet in the case of computers or those individuals with smartphones in that case</p>
<p>Unit of measurement: Percent of individual interacting with governments through the Internet for each service listed</p>
<p>Main concepts used: Availability of government services on the Internet and citizen uptake</p>
<p>Other concepts: Specific government services, such as income taxes, social security benefits, specific form, personal documents etc.</p>
<p>Calculation method: Number of individuals using one of the listed services divided by the total number of participating individuals and multiplied by 100</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the household/individual survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Any breakdown desirable based on socio-demographic profile of users, e.g. by age, gender, and income</p>
<p>Reference period: The week of the observation (or any chosen period)</p>


<p>Dataset: Use of e-commerce</p> <p>Data source: Possible new ICT-based data collection based on users' home computers or smartphones</p>
<p>Description: This dataset investigates whether or not individuals ordered goods or services over the Internet, what types of goods or services (from a list), which ones were downloaded directly from the Internet, and from where (own country, other EU, elsewhere)</p>
<p>Key indicator(s) included in the dataset: Did you buy goods or services on the Internet What types: Food or groceries, Medicine, Books, magazines, Clothes, sports goods, Computer hardware, electronic equipment etc... Which of the following were downloaded or accessed from websites: Films, Music, Computer software etc... Origin of seller (own country, other EU, rest of the world)</p>
<p>Policy relevance: Other than using the Internet for information, knowledge or entertainment, carrying out commercial transactions on the Internet is a key policy issue. It relates directly to issues of technological sophistication and competitiveness.</p>
<p>Statistical population: Those individuals with home access to the Internet in the case of computers or those individuals with smartphones in that case</p>
<p>Unit of measurement: Percent of individual engaging in e-commerce for each activity or category listed</p>
<p>Main concepts used: Ordering goods or services on the Internet</p>
<p>Other concepts: Downloading goods or services directly from the Internet; origin of sellers</p>
<p>Calculation method: Number of individuals ordering a listed good or service divided by the total number of participating individuals and multiplied by 100 – whether ordered on or downloaded from the Internet. Number of transactions by origin of seller (own country, other EU, rest of the world) divided by total transactions and multiplied by 100</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the household/individual survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Any breakdown desirable based on socio-demographic profile of users, e.g. by age, gender, and income. As well, breakdowns by origin of seller.</p>
<p>Reference period: The week of the observation (or any chosen period)</p>


<p>Dataset: e-skills</p> <p>Data source: Possible new ICT-based data collection based on users' home computers or smartphones</p>
<p>Description: Indirect gauging of Internet users' e-skills through activities they carry out</p>
<p>Key indicator(s) included in the dataset:</p> <p>Internet activities already carried out:</p> <ul style="list-style-type: none"> Using a search engine Posting messages to chat rooms etc. Make Internet phone calls Creating a web page Modifying security settings in browsers etc...
<p>Policy relevance: Skills necessary to use the new technologies adequately come in a wide range, from the expert IT user to the novice with very basic skills. Policies for training at various levels, such as schools and workplaces, especially as the technologies become more complex, need planning and resources.</p>
<p>Statistical population: Those individuals with home access to the Internet in the case of computers or those individuals with smartphones in that case</p>
<p>Unit of measurement: Percent of individual engaging in e-commerce for each activity or category listed</p>
<p>Main concepts used:</p> <p>e-skills as assessed through activities carried out by individuals</p> <p>Other concepts:</p>
<p>Calculation method: Number of individuals that carried out tasks among those listed divided by the total number of participating individuals and multiplied by 100</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the household/individual survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Any breakdown desirable based on socio-demographic profile of users, e.g. by age, gender, and income</p>
<p>Reference period: The week of the observation (or any chosen period)</p>


<p>Dataset: Website enterprise information</p> <p>Data source: Possible new ICT-based data collection based on enterprises' websites</p>
<p>Description: A host of indicators that can be collected from an enterprise's website. This ranges from basic tombstone information, such as language/s in which the website is available and contact information, to in-depth information regarding the availability of a Privacy Policy Statement or a registration facility, to information about product and prices, job vacancies and/or linkages to social media sites.</p>
<p>Key indicator(s) included in the dataset:</p> <ul style="list-style-type: none"> Web site or Home Page Language Privacy Policy Statement, Registration facility Product lists and information, Pricing by product Job postings Links to social media, etc...
<p>Policy relevance: The availability of websites is monitored to assess hierarchical progress towards technological sophistication among European businesses. Policies for sustainable growth, for example, relate both to the existence and diffusion of skills necessary to build the requisite technical infrastructure as well as the ability of enterprises to expand their reach and be internationally competitive.</p>
<p>Statistical population: Those enterprises with a website who agree to participate in the web-based data collection method</p> <p>Unit of measurement: Enterprises with a website and counts of individual pieces of information, such as language/s available, registration facilities, product/price information, job postings and links to social media</p>
<p>Main concepts used: Availability of website and key characteristics</p> <p>Other concepts: Sophistication of website, information with potential to substitute for existing surveys or useful for future surveys</p>
<p>Calculation method: Number of enterprises with a website offering each of the categories listed as a percentage of all participating enterprises with a website</p>
<p>Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the enterprise ICT survey, the data may be used by themselves or may have to be combined for the production of indicators</p>
<p>Breakdowns: Any breakdown desirable based the profile of enterprises, e.g. NACE and size</p>
<p>Reference period: The week of the observation (or any chosen period)</p>


Dataset: e-commerce
Data source: Possible new ICT-based data collection based on enterprises' websites
Description: Examines whether an enterprise website is equipped to offer online ordering, reservations or bookings, if it actually received such orders, how many, and from which geographic areas (own country, other EU, rest of the world)
Key indicator(s) included in the dataset: Website capable of online ordering Orders received online (Y/N) How many Geographic origin of online orders Own country Other EU Rest of the world
Policy relevance: e-commerce has become a key preoccupation of policy makers for many years because of the technological sophistication it entails, and therefore the associated need to ensure that there are enough skilled human resources to build the technical infrastructure needed, as well as because of the growth agenda and the future competitiveness of Europe. Moreover, e-commerce needs monitoring for many related issues, such as cross-border consumer protection in the online world, customer redress, logistics etc.
Statistical population: Those enterprises with a website who agree to participate in the web-based data collection method
Unit of measurement: Enterprises with a website capable of online orders or reservation
Main concepts used: Online ordering, orders received, geographical location
Other concepts: Capability to track orders online
Calculation method: Number of enterprises offering each of the functionalities listed as a percentage of all participating enterprises with a website
Methodological issues: Depending on whether or not this potential data collection method takes place independently or as a component of the enterprise ICT survey, the data may be used by themselves or may have to be combined for the production of indicators
Breakdowns: Any breakdown desirable based the profile of enterprises, e.g. NACE and size
Reference period: The week of the observation (or any chosen period)

12.2. D1 - Definition of Internet data-based indicators part II

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’

D1.b Definition of internet data-based indicators – Part II (vision doc)

September 2013

Document Service Data

Type of Document	Deliverable		
Reference:	D1.b – Definition of Internet data-based indicators Part II (vision doc)		
Version:	2	Status:	Draft
Created by:	Lefteris Angelis, Michalis Vafopoulos, Dimitris Kalogeras, George Sciadas	Date:	23/9/2013
Distribution:	European Commission – Eurostat/G6, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	16/7/2013	Initial release
2	23/9/2013	Modifications following Eurostat's comments at the 2 nd progress meeting

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

Executive summary	3
1. The Official Statistics approach.....	3
2. Internet, the Web and the 5Is.....	4
3. IW4OS: A novel conceptual framework	5
3.1. Interaction.....	6
3.2. Instantaneousness, Information Overload, Informality & Irregularity	6
4. Issues in implementing IW4OS due to the complex nature of the Web	7
4.1. Issues related to the Web as a self-organizing network.....	8
4.2. Issues related to the nature of Web entities	10
4.3. Statistical issues related to Internet & Web as data sources	11
4.4. Sampling issues related to Internet and the Web as data sources	12
5. Facets of the Data era.....	14
5.1. Inferred data	14
5.2. Big data	14
5.3. Open data.....	14
5.4. Linked data.....	14
5.5. Federated open data	15
6. Reflections of the conceptual framework to ICT statistics.....	15
6.1. ICT Usage.....	16
6.1.1. ICT products	16
6.1.2. ICT infrastructure	16
6.1.3. ICT supply.....	17
6.2. Content and media	18
7. To the future: the Internet of Things.....	19
8. Additional forward-looking remarks.....	20
8.1. Dis-assembling and re-assembling	20
8.2. Some additional matters	22
9. References	25

Executive summary

The recent NSA scandal could be also viewed as a reassurance that data collection and analysis from the Internet and the Web is the Holy Grail of the political and business games in global scale. This fast evolving online environment permeates almost every aspect of reality and influences the behavior of physical and legal entities. But, still, has not fallen in the range of the Official Statistics radar.

In this study, we initiate a fresh view in investigating the potential transformations that the Web induces to Official Statistics. We argue that the Internet and the Web and Official Statistics should be analyzed under a common framework that puts equal attention to both worlds and enables efficient feedback loops between them.

The study is organized as follows. The first two sections highlight the stylized facts about the Official Statistics and the Internet and the Web, respectively. The proposed conceptual framework is analyzed in the third section. The fourth section investigates the main issues that arise in implementing the aforementioned framework due the complexity of the online ecosystem. These issues are related to the nature of Web entities and their characteristics as statistical objects of study. The fifth section briefly discusses the data approach to Internet and the Web as data sources for Official Statistics. The next section initiates the application of the proposed conceptual framework to the ICT statistics and the final sections refer to a future and promising online data source: the Internet of Things and to future challenges for Official Statistics.

1. The Official Statistics approach

Almost every country has designated a public sector (or public-funded) agency, which is responsible of the production and dissemination of official statistics under local and global standards. According to the Principle 1 of the Fundamental Principles of Official Statistics¹, *«Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation»*.

Official statistics must be characterized by:

- Relevance to the users
- Accessibility
- Clarity

¹ <http://unstats.un.org/unsd/methods/statorg/FP-English.htm>

- Timeliness & Punctuality
- Accuracy
- Comparability
- Coherence
- Formality
- Regularity

The assessments of *relevance* should consider (a) who are the users of the statistics, (b) what are their needs, and (c) how well does the output meet these needs.

The main fields where *accessibility* and *clarity* should be examined include (a) needs of analysts, (b) assistance to locate information, (c) clarity and dissemination.

An assessment of *timeliness* and *punctuality* should incorporate the production time, frequency and punctuality of release.

Accuracy is divided into sampling error and non-sampling error, where non-sampling error includes coverage error, non-response error, measurement error, processing error and model assumption error.

The issue of *comparability* should be investigated in terms of comparability over time, spatial domains (e.g. sub-national, national, international), and domain or sub-population (e.g. industrial sector, household type).

Coherence should be considered in terms of coherence between data produced at different frequencies, other statistics in the same socio-economic domain and sources and outputs.

2. Internet, the Web and the 5Is

Just twenty years from its inception and the World Wide Web (or simply Web) has a transformative impact on almost every facet of our society. While the Internet had been introduced twenty years earlier, the Web has been its most popular application with more than two billion Users worldwide accessing some trillion Web pages. Searching, social networking, video broadcasting, photo sharing, blogging and micro-blogging have become part of everyday life whilst the majority of software and business applications have migrated to the Web.

The concept of Internet or to be more accurate, the “Web ecosystem” (or simply “Web”), in the context of this report, includes three interconnected parts: (1) Internet infrastructure, (2) Web technologies and online content and (3) Users. Users navigate, create and edit existing content in the Web.

The Web has been originated as a software program of interlinked hypertext documents accessed via the Internet. Using a browser, Users access Web pages that may contain text, images, videos,

or other multimedia and navigate among them using hyperlinks. The Web constitutes an information space in which the items of interest, referred to as resources, are marked up by a set of rules (i.e. HTML²), identified by global identifiers (URI³) and transferred by communication protocol (HTTP⁴).

Web has become the most successful and popular piece of software in history because it is based on a technical architecture, which is simple, free or inexpensive, networked, based on open standards, extensible, tolerant to errors, universal (regardless of the hardware and software platform, application software, network access, public, group, or personal scope, language and culture operating system and ability), powerful and enjoyable.

Web has been evolved from a piece of software code to a dynamical and multi-purpose system. In its early stages addressed mainly technological needs, such as an interlinked bulleting board with low levels of interaction. In subsequent years, though, has become a decentralized construct of diverse interlocking contexts. Users not only post and link digital content, but also interact and exchange information in and through it.

Today, the Web is described as a techno-social artifact characterized by:

- Interaction
- Instantaneousness
- Information overload
- Informality
- Irregularity

Web enables new forms of asynchronous and synchronous interaction and it is instantaneous in the sense that many things are happening nonstop at every point of time (approaching the notion of continuous time), causing an unprecedented information overload. Data produced in and through the Web are characterized by their high *volume*, *velocity* and *variety*⁵.

Central or formal authorities do not regulate its basic functions. Its enormous impact, scale and dynamism in time and space have been resulted a series of novel and irregular social phenomena.

How these phenomena can be quantified and analyzed in order to become strategic knowledge in the personal and social level? Are existing methods and institutions ready to explore and exploit this new field of human interaction?

3. IW4OS: A novel conceptual framework

² HyperText Markup Language.

³ Uniform Resource Identifier.

⁴ HyperText Transfer Protocol.

⁵ For more details on the data – big, linked and open – approach refer to Section 5.

The new sources and forms of data in the Web are raising imperative questions to Official Statistics. The envelope question is which methods should be changed or even introduced to retain the aforementioned characteristics of Official Statistics, but at the same time will exploit the emerging potential of online contexts?

Before starting to form specific proposals and engineer tools for new data sources and indicators, a coherent common mindset should be introduced. The proposed conceptual framework for Internet and Web as data sources should facilitate the orchestration of their main characteristics with the approach of Official Statistics (the Internet and Web for Official Statistics framework-IW4OS – is presented in Diagram 1).

3.1. Interaction

At the current Web 2.0 era, users are the protagonists of the online ecosystem because they can easily edit, interconnect, aggregate and comment online content as never before. Most of these opportunities can also be engineered in the personal level. The traditional triptych of producers-exchange-consumers has been replaced by the *prosumption* model where consumers contact producers directly or can act, at the same time, as producers. Web 2.0 enables interaction and crowdsourcing through openness, peering, sharing and acting globally (Tapscott & Williams, 2008).

These new modes of human interaction and production could be incorporated in providing more accessible and relevant Official Statistics to the users. For instance, social media can serve both as pools for data collection and data publication in order to get direct feedback from the online users about the usefulness of indicators.

3.2. Instantaneousness, Information Overload, Informality & Irregularity

Web 3.0 technologies, such as Semantic Web (Berners-Lee, 2006) and Linked Data (Bizer, Heath, & Berners-Lee, 2009) have been engineered to provide assistance to locate information by human and machine-based tools. Existing *ontologies* and vocabularies have been expanded to handle online statistical information and mainstream statistical standards (e.g. Data Cube vocabulary (Cyganiak, Reynolds, & Tennison, 2012), Linked SDMX data (Capadisli, Auer, & Ngomo, 2013), etc.).

The most important aspect of the proposed analysis is to identify an effective set of *transformation* and *validation* rules that will enable the timeliness, punctuality, accuracy, comparability, coherence, and eventually, formality of IaD sources.

Based on the past experience in developing Internet and Web standards, these rules should not be all-encompassing from the beginning, but will better follow the “divide-and-conquer” and the procrastination principles. First, the general problem will be demarcated in smaller sub-problems (e.g. IaD for specific indicators in ICT statistics) and second, according to the procrastination

principle that can be summarized in the phrase “don’t do anything that can be done later by users⁶” most problems confronting the IaD approach can be solved later by other researchers and users of statistics.

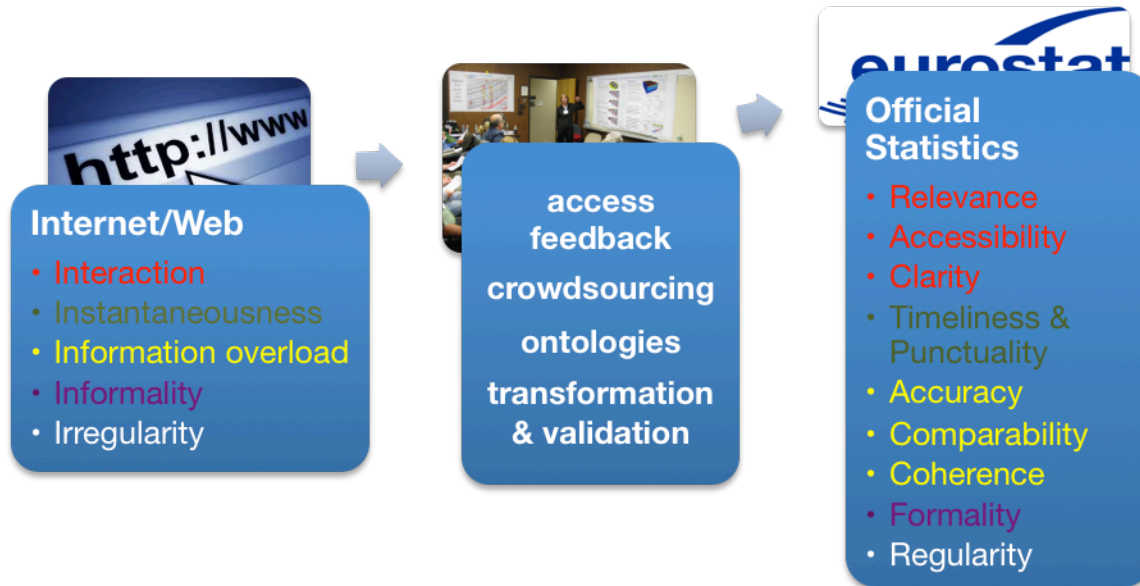


Figure 1: Internet and Web for Official Statistics framework (IW4OS) is designed to orchestrate the main characteristics of the online ecosystem and Official Statistics.

The transition from Official Statistics obtained by real world data through surveys and personal communication with individuals, to a new era of indicators computed complementarily or even solely from Internet and the Web is not easy or obvious. We have to study in depth and understand the universe of Internet and the Web as an extremely complex system in order to fully utilize it for obtaining Official Statistics through the proposed conceptual framework. Next, we discuss this complex nature of the Web and also the problems and challenges in the implementation of the conceptual framework.

4. Issues in implementing IW4OS due to the complex nature of the Web

The aforementioned evolution and the impact of the Web in every area of life have led to new ways of communication, interaction and socialization. Millions of people today live their everyday lives in two worlds: the real world of physical entities and the virtual world of Web, consisting of ideas, information, abstract concepts and relations between them. The impact of Web in any human activity is so strong that these worlds are not perceived as separate any more, especially by the younger ones. The existence of this new world emerged the need for a new

⁶ An idea from a 1984 paper by (Saltzer, Reed, & Clark, 1984), that was also used by Zittrain (Zittrain, 2008) to explain Internet’s architecture.

science, known as Web Science, trying to study, model and systematize the entities and the activities of the virtual world which is strongly depended on the physical world.

Web is a potential source of data for indicators and official statistics regarding individuals and organizations, their activities and their interactions. However, in the collection of such data the special characteristics of the Web should be taken into account. The nature of the Web as a new universe has not been fully explored and it is a subject of ongoing research. The structure and the peculiarities of this gigantic network connecting physical and conceptual entities should be taken into account when searching for information. The fact is that the new sources of data can radically change the ways we define and measure official statistics and indicators in society.

The transformation and validation of data from Internet and the Web to indicators and official statistics with the characteristics given in the IW4OS conceptual framework is certainly a complicated task and several research issues are expected to be raised in the adoption and the implementation of the proposed approach. In the following sections we provide an account of related issues, problems and challenges along with potential approaches.

4.1. Issues related to the Web as a self-organizing network

An example of how Web can change the perspective of the world is the notion of *communities*. In the physical world communities are well defined and strictly founded on clear relations or contract agreements among entities that are officially registered in archives and they are recognized by the authorities (e.g. communities of people living in a certain area or commercial/business communities). However, Web communities (Flake, Lawrence, Giles, & Coetzee, 2002) are not well defined although they exist and can be identified through algorithms and data mining. This is a result of the unique, self-organizing, without central authority, nature of the Web. Of course mathematical notions can be used to define such communities but this is totally different from what is happening in the real world.

Communities can be characterized by the effort towards a common goal or a common interest, like the communities of open source software where people are linked and interact in a decentralized way. Furthermore, communities can be more abstract; i.e. a number of web pages (personal, business or scientific) that are seemingly irrelevant can evolve strong connections between them for an unanticipated reason (for example reaction against a governmental decision or as support to a research in a specific health issue). These communities, although not legally defined, should be taken into account when searching for indicators, especially in ad hoc studies (health, learning, etc.).

Even the *sampling* procedures that are followed in traditional surveys should be reconsidered and revised according to the self-organizing network nature of the Web while the relevant communities that exist but are not known to the researchers should be identified and addressed.

Although, ad hoc studies and relevant indicators can be greatly benefited by getting information by specific communities, there are concerns for issues related to representation and coverage of the target population.

The complex nature of the Web and the wealth of information that can be drawn from it, may lead several of the old sampling methods and survey tools to restriction or even to extinction in the coming years. Telephone calls and traditional questionnaires were of course useful for many decades for capturing instances of the real world, but they are too inadequate in a world where information flow is limitless. In this regard, it is expected that new, more sophisticated, algorithmic-based and Web mining-based approaches will be implemented or even devised.

Apart from the old methods of data collection that are expected to change in the new Web era, it is our belief that the types of questions and even the indicators will be changed. Indeed, in *common questionnaire surveys*, questions should be expressed in simple, easy-to-understand natural language so that even people with minor education can understand them. The effect of this limitation was that a questionnaire had to contain a large amount of indirect simple items in order to capture a latent and partially abstract notion like “customer satisfaction” from all of them. And even then, difficult-to-measure notions like “skills” were subject to measurement error. However, in the Web era, since the “Big data” originating from the Internet and the Web are multidimensional and they are produced continuously in huge amounts, there is no point to try to answer questions like “how many times did you used a browser last month?”. The old data were comprised of answers like “Yes” or “No” or even numbers but the new data contain clickstreams, sessions, connections, etc.

Let us consider for example an ad hoc study regarding the *skills* of experts in enterprises, which develop software products. A possible question like “In teams of developers how many of the team members provide knowledge to others and how many seek knowledge?” would be of interest since it is related to skills and experience. However, this is not a simple question that can be answered by asking questionnaires, since no employee or manager has a clear overall picture of the entire network of information flow within the company. Email message exchanges between employees asking each other questions about problems encountered in their work could be analyzed by graph/network mining algorithms in order to understand and measure how knowledge is being shared within companies and in order to discover and enumerate whether and how many knowledge providers and seekers exist. Network analysis (M. Newman, 2008) is able to recognize people that are “hubs” in a network and can be considered leaders or experts. Of course, such methodologies are not easily accepted by companies but fortunately the communities of open source provide the field for developing tools for measuring similar indicators.

Therefore, the main point of all the aforementioned discussion is that we have to realize that by taking into account the underlying network nature of the Web and by applying the vast research results in the areas of statistics, data mining and web mining, we can change radically the way

we collect data from targeted individuals, organizations or communities which can be located by appropriate algorithms.

4.2. Issues related to the nature of Web entities

Another special characteristic of the Web as a human creation determining the availability, dissemination and access to a major range of information, is its triadic hypostasis. Indeed, there are three main dimensions - components of the Web entities (Zaiane, 1999), (Paparrizos, Koutsonikola, Angelis, & Vakali, 2010): the *structure*, the *content* and the *use* which constitute the main aspects that have been recorded in the literature and are in accordance to the mathematical concepts of structure, function and evolution of dynamical systems (MEJ Newman, 2003) .

This perception of the Web is justified by the fact that each entity in it is created in a certain way (structure), includes a plethora of (often heterogeneous) information (content), while each user perceives, interprets and uses the content according to individual opinion and requirements (use). At the same time, the structure, the behavior of users and the availability of the content in the Web represent opinions, attitudes, trends, culture and communities' information. Therefore, any method trying to extract data for indicators and official statistics should consider this triad and focus on one of the three dimensions or on a weighted combination of any of them.

The three-dimensional concept of Web entities is meaningful when trying to orchestrate the Internet/Web data sources characteristics with the Official Statistics requirements under the introduced conceptual framework and depending on the type of study. For example "Informality" vs "Formality" may be a subject related more to content than to structure or use.

In any case, such a triangulation of Web can aid the validation procedure of the quality of data that will be used for specific indicators. Especially for site-centric methods monitoring and using data, it is essential to have a quality rating for Web sites. This quality rating can be assessed from the structural characteristics of sites, their content and their usability. These quality standards can be used for the user-centric approaches especially when an approach is based on the monitoring of the users' browsing behavior.

In general, systematic study and analysis of the Web is needed in order to be used as a source of data for official statistics. This study involves practical difficulties and challenges and therefore considerable research. These difficulties are continuously increasing due to the fact that the Web is driven by constantly evolving human ideas and tactics in accordance to the rapidly evolving technology. It is our strong belief that the use of the Web as data source for indicators and official statistics should follow a stage of well-understanding and systematization.

In conclusion, the research connecting Web data with official statistic through the conceptual framework, has to be conducted in a hierarchical manner, using the methodologies and tools (taxonomies, ontologies etc) that have been developed for the study of the Web. This research will facilitate the monitoring and the extraction of data related to a variety of applications, decisions and activities of scientific, business and social interest.

4.3. Statistical issues related to Internet & Web as data sources

In this section we discuss statistical issues and challenges raised in the procedures of transformation and validation of data from Internet and the Web to indicators and official statistics according to the conceptual framework.

Internet sources are expected to provide data for appropriate information related to specific indicators. The task of substituting the traditional surveys by Internet data is surely difficult and complicated; therefore it is reasonable to assume that at the first stages of such a project there will be a phase of co-existence and collaboration with the traditional methods. In fact, the data collection and usage from Internet can be revised on the basis of data from surveys or administrative records.

Another possibility is that new types of data collected from the Internet are not in direct alignment with any of the questionnaire surveys conducted so far. In such a circumstance, the new data can provide the basis for a completely new indicator that was not otherwise available. In general, we cannot always expect to have a 100% accordance of the traditional surveys to the new Internet data sources and therefore we may need to adapt the indicators to the new data. In any case we always have to be careful and evaluate scholarly new sources of data before deciding to use them. Initial experiments or pilot studies are necessary in order to assess the benefits and the limitations of the new sources. The transition from traditional methods for collecting data to totally new methodologies and types of data is an evolutionary process which is expected to open new perspectives to official statistics.

The new Internet-based, nontraditional data sources raise challenges related to the evaluation of their accuracy and the measurement of their error. As an example (Martin, Straf, & Citro, 2005) we can mention the harvesting of website data to develop current consumer price indexes. The benefit in comparison to traditional methods is significant timeliness and cost savings, however it is not clear how to adjust these data for consumer expenditures that occur off-line so that they accurately represent the universe of purchases.

Another problem and at the same time a challenge is that Internet-based data should be continuously under control for their consistency. Otherwise, results relying on uncontrolled data are subject to error and risk due to unobserved and unregistered changes in the content or structure of the sources.

The use of textual data in the formation of indicators is another issue, especially for ad hoc studies concerning for example public opinion or consumer preferences. The problem is quite old due to the “free-text questions” used in traditional questionnaires, but when it comes to Web data it becomes a problem of much larger scale. Textual data can be found everywhere in the Web; web sites, blogs, social media, news, scientific databases are examples of sources containing text that can be exploited by applying text-mining methodologies and converting it to numerical data that can be subsequently used for the construction of new indicators. The wealth of the Web in text makes the task challenging.

Bias is major problem with Internet data. A significant part of any population does not have access to Internet by any circumstances. So there is always a significant coverage error in any sampling scheme based only on individuals who have access to the Internet. This can lead to biased estimates when generalizing to a larger scale (e.g. national) population. Even if in future, in ideal situations, the advances of technology permit access to the entire population, sampling procedures will still be necessary since the cost problems of maintaining enormous databases in contrast to the limited funding resources will always be present.

4.4. Sampling issues related to Internet and the Web as data sources

Sampling designs are of fundamental importance in any process of collecting data. As we already mentioned, the Web is a gigantic self-organizing network so its structure cannot be ignored when applying sampling schemes. The usual probability-based sampling schemes (samples selected so that each sample member has some known nonzero probability of being selected into the sample) are not applicable when we want to draw a sample of sites or users. The main problem here is that *randomness* is not easily defined or achieved, for example the generation of random addresses of sites or emails is considered a difficult problem.

Therefore in Internet-based sampling schemes we should think of using non-probabilistic methods, i.e. methods for selecting a sample where units are not chosen so that each one has some known nonzero probability of being selected into the sample.

Common examples of nonprobability samples are convenience sample, quota samples, and expert choice samples. In any case, it is important to know the limitations of each method.

Non-probabilistic sampling schemes (for websites or users) (*Questions and answers when designing surveys for information collections*, 2006):

- Convenience samples: The sample comprises of units/individuals of the population under study that are available or willing to participate voluntarily or by some reward. There is no meaning to ensure that the samples are representative of the population or to try to generalize the results to a population. However they are relatively inexpensive, easy to plan, and easily monitored. They are especially useful for pilot research or assessment studies, for example in order to align the results of Internet data to results of a survey

with questionnaires. In such a scenario, a readily available, easily accessible and willing to participate group of units (individuals or companies) could be used for collecting data for the same indicator by both methodologies: the traditional questionnaires and the Internet. Then, a preliminary statistical analysis could be used for detecting the correlation between these methods. Although this sampling technique suffers from lack of representation and systematic bias, the results could provide valuable indications of associations and trends which need to be investigated in future extensive and systematic studies.

- Quota samples: These are samples where units are selected non-randomly based on a percentage which is defined in such way that the final numbers of participating units (websites or users) with given characteristics have the same proportion as corresponding units have in the population. This method gives a sense of representativeness; however there are still characteristics of convenience.
- Expert choice: An expert chooses specific sample units (websites or users) with certain attributes in order to simulate representative members of the population. This method can also produce inconsistent, entirely different types of samples, depending on the different opinions of the experts used in the procedure.
- Snowball samples: These schemes are for rare populations or populations that are hard to define or locate. A sample for the rare population is created or identified by starting with a set of individuals belonging to the target population, and asking this initial set to provide information for other members of this population. These units are then contacted for information that they may have on others in the population. This method of sampling is ideal for creating a sample based on informal social networks. For the Internet-based data this procedure can be applied to websites automatically by following consecutive and relative links from website-to-website but also to users by personal contact. This scheme seems to be very promising for sampling in targeted Web communities, especially by exploiting the power of social media.
- Cut-off samples: The potential units (Websites or users) are first ordered with respect to some important characteristic (for example a quality ranking according to the triad structure-content-use or to a reliability assessment) and then the units with the greatest amount of the characteristics (most qualified) are selected until some pre-specified percentage is included in the sample. This is also a promising sampling scheme for Internet-based data since it is able to involve quality criteria.

Although non-probability sampling provides convenient methods for collecting data, overcoming the problem of randomness, all methods have drawbacks which should be taken into account. Their main problem is the accuracy of indicators calculated from the collected data. For example Yeager et al (2011) showed that the accuracy of non-probability samples in telephone and internet surveys is consistently lower than the accuracy of probability surveys, even after post-stratification adjustments for improvement of non-probability methods. Especially the sampling methods using volunteers are prone to bias not only due to the lack of representation but also due

to the strong correlation between the inclusion probability and the targeted variable. Finally, in statistical literature, and also in literature of other disciplines using statistics, there are recommendations and correcting methods for non-probability sampling techniques like the quota samples (see for example Berinsky, 2006). Due to the novel nature of problems related to Internet data, it is expected that research on improving the accuracy of non-probability sampling methods will be launched.

In general, more complicated and sophisticated sampling methodologies used in various research areas can be proved beneficial for drawing data from the Web. Such an example is the measurement of auxiliary variables that are highly correlated to the actual variable under study. This methodology is used in certain sampling designs (for example double sampling) when the measurement of the actual variable (indicator) is too expensive or sensitive to confidentiality. Then by designing an appropriate sampling scheme and by measuring inexpensive auxiliary variables, we can estimate the indicators.

5. Facets of the Data era

It looks like that the discussions about the “social media era” surrender their position to the “data era”. For clarity, let us first provide our approach to some basic definitions about the different facets of data.

5.1. Inferred data

As inferred data could be considered the data that are collected through the “traditional” crawling and scrapping processes of unstructured and/or semi-structured of webpages. Usually, inferred data are stored in RDBMS and analyzed by “data mining” approaches.

5.2. Big data

Big data is popular term that has various definitions and views. According to Wikipedia big data is considered to be a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

5.3. Open data

Again from Wikipedia: “Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other "Open" movements such as open source, open content, and open access.”

5.4. Linked data

Linked Data enable the creation of better and massive services for data re-usage, driving existing infrastructure in its full potential. For government bodies, Linked Data adoption is focused on

open, transparent, collaborative and more efficient governance. For enterprises, the core issue is about effective knowledge management and the implementation of new business models that initiate more energetic involvement and collaboration between producers and consumers (Vafopoulos, 2011a).

Linked Data is an attempt to simplify and spread horizontally throughout the Web the network externalities that exist in Web 3.0. Specifically, two sources of value have been identified for Linked Data technology. First, it enables users to build bidirectional and massively processable interconnections among online data and second, these data are critical enablers for existing infrastructure in the government and business spheres (Vafopoulos, 2011b).

Thus, Big data is more about *scale*, Open data about *access* and Linked data about the *use* of data (small or big, open or closed).

5.5. Federated open data

The present project is focused in «Federated open data» as have been defined by (Glasson et al., 2012). Federated open data is the counterpart (or supplement) of the so-called “open data” of governments. It refers to a shared sub-set of Big data from private sector entities, which will be “open” for use by NSOs.

As we indicated earlier, the proposed framework is meant to be applied to specific statistical indicators. The first implementation concerns data collection and analysis for ICT statistics.

6. Reflections of the conceptual framework to ICT statistics

The OECD guide⁷ for measuring the ICT sector is the commonly adopted way to measure the various facets of ICT activity namely, products, ICT infrastructure, supply, demand by businesses, demand by households and individuals, content and misc. Indicators have also been proposed by the ITU⁸. Currently ICT usage surveys cover activities of Internet where individuals interact with web servers, which typically offer services or merchandise.

Social networks tend to capture the major part of human activity and as such constitute a trend (yet though not an indicator) of human activities. For instance twitter⁹ messages capture everyday activities. Mining through register IDs reveals significant information of epidemics. Under certain limitations¹⁰ collecting equivalent statistics compared from traditional surveys is a

⁷ <http://www.oecd.org/sti/ieconomy/theictsector.htm> , <http://www.oecd-ilibrary.org/deliver/fulltext/3011041ec072.pdf>

⁸ <http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>

⁹ <http://www.twitter.com>

¹⁰ <http://dimacs.rutgers.edu/~graham/pubs/papers/cormodewdsa.pdf>

challenge because Internet and social media reformulate the form of meaningful queries. Facebook¹¹ constitutes another big reservoir of social life and partially ICT usage indicator. All big firms have moved to Facebook to gain from the personal connection to users. In terms of ICT usage, Facebook acts as a social specific web that introduces firm and products to individuals.

M2M communication has started to capture an emerging human need for better management of their facilities (i.e. ports, sewage systems, smart electricity grid, etc.) thus ultimately evolving to smart cities¹². Millions of data fountain from sensors spread around indicating human activities. Those data require efficient manipulation and correlation with the legacy merchandise activities. Something like that is on verge of semantic annotation of linked-data and statistics research.

6.1. ICT Usage

The aforementioned emerging uses of Internet require new type of queries and methodologies in order to be tracked of. This is clearly a revolution when considering Internet as a data source. Adopting an evolutionary approach we propose to track the legacy ICT usage by means of Internet mechanism. Unfortunately the data generation has not yet been restructured in order to recover indicators explicitly by data sources, thus we still limit ourselves to indirect methods (such as web-crawling, etc.) in order to collect the required data. For instance Facebook does not provide any gateway (i.e. API) to collect metrics and indicators of references to keywords whether those refer to advertised products and services or to free text.

6.1.1. ICT products

As far the ICT products data are collected either from surveys or from customs offices. It is expected that this collecting mechanism will deteriorate as more individuals and companies are composing their supplies over electronic stores. It is therefore important to circumvent the existing or immediate foreseeable barriers for this data collection.

As an ever-growing number of ICT products are marketed through on line shop we propose the collection of ICT product data through proper characterization of online store products. As cross border e-market is common practice, we propose additional tagging of ICT product sale indicators followed by country specification in order to collect country aggregated metrics. Such metrics can be collected through web logs.

6.1.2. ICT infrastructure

As far infrastructure is concerned, the collection of indicators becomes more feasible given that Internet service providers (ISPs) and regulation authorities provide the technical and legislative means respectively. A typical indirect usage of the ICT infrastructure might be provided by the number of registered domain names (DN). Registered DNs could be supplied through Internet from top level country registrars.

¹¹ <http://www.facebook.com>

¹² <http://www.smartsantander.eu/>

Internet broadband usage has always been in the center of interest of carriers strictly for capacity prediction. Regulation authorities in various countries have started projects with active methods for broadband speed measurements (i.e. www.samknows.eu, www.measurementlab.net/). Those measurements utilize additional components in order to check the quality and the effective speed of broadband connection of consumers. Different metrics such as mean connection time could only be retrieved by using based facilities (i.e. customizable search bar) or carrier based facilities such as the radius accounting database.

An indicator of national ICT infrastructure could be estimated by the total sum of exchanged traffic in national internet exchange points (IXP). As IXP maintain online volume graphs it is possible to use them as an online data source.

An indication of ICT infrastructure could be associated by the number of autonomous systems (AS) in the routing table, while these data are broken out per country. Autonomous Systems are the elementary routing placeholders in the Internet with their own distinctive routing policies that appear in the Internet routing table^{13,14}.

In addition to legacy carriers, content delivery networks (CDN) emerge as competitive Internet rich media (audio, video) transporters. For instance the akamai (www.akamai.com) CDN provider delivers indicators for average connection speed, average peak connection speed, high broadband connectivity (>10 Mbps) and normal connectivity (<10 Mbps) through their quarterly edition “State of the Internet”¹⁵.

6.1.3. ICT supply

The overall ICT supply can be estimated by sum of products and services. Internet is used by service oriented companies which declare their presence in terms of domain name. Hence a typical indicator of a national ICT supply might be provided by the number of registered DNS names. Aggregation per country or per subdomain (i.e. .ac, .co) -wherever this fit- is maintained by the country top level domain (cTLD) registrar.

As far the ICT product supply is concerned the only possible means of utilizing the Internet as a data source is to collect metrics from web crawlers or meta search engines for internet shopping. For instance the www.skroutz.gr/ lists the number of products displayed from all shops, the number of individuals and the number of shops. In a relevant way, auction sites (i.e. www.ebay.com) host the number of products on sale by individuals. As action sites host electronic shops also, it is useful to collect indicators by aggregating data per number of e-shops.

¹³ <http://bgpmon.netsec.colostate.edu/>

¹⁴ <http://www.ripe.net/data-tools/stats/ris/routing-information-service>

¹⁵ www.akamai.com/stateoftheinternet/

E-government sector

Another fundamental sector of ICT activity is the activity of public sector with respect to the automated internet ready application for the citizens. The sum of publicly offered services to citizens for the sake of state constitutes the E-government (e-gov.) sector. The number of services offered by the central government as well as from the regional and municipal sector constitute the total supply of e-gov sector. The number of services are typically monitored from national ICT observatories or from aggregation portals of e-gov. sites.

ICT demand - transactions

The demand of ICT corresponds to activity from individuals to buy services or products offered in Web sites. Although it is possible for an individual to commit for payment by using a legacy payment method as mail order we will focus our study only to those transactions which are completed electronically. For the retail sector an indicator of ICT could be retrieved either by the selling web sites or from equivalent web banking activity.

Web banking

Web banking activity corresponds to demand from individuals paying via credit or debit cards. It is assumed that only bank sector could verify the number and volume of electronic card transaction hence the banking sector should somehow differentiate between legacy electronic credit card machine used in shops and web sites. Indicators for web based banking could be traced by web log activity traced indicating the final state of web site visit (i.e. payment or abort to a different site).

Furthermore the banking sector considers internet and mobile banking as a means to minimize the operational costs. It is essential to gain access to overall number and volume of transactions conducted by Internet banking sites as opposed to legacy order in front of desks.

E-gov sites

The demand of e-gov sites could be easily quantified by the number of different register and the number of submitted and produced objects. Those indicators could be easily accessed by weblogs of portals hosting the respective e-gov services.

State driven ICT demand

Another interesting figure of the ICT demands corresponds to the ICT demand by the state as it is referenced by public request for proposal or even more of contractual agreements with specific CPV codes¹⁶ for ICT.

6.2. Content and media

Media (audio and video) has rapidly captured the interest of broadband users. Media companies all over the world have migrated the majority of their content to Internet allowing users to watch it using new type of end devices such as Internet TVs and smart-phones.

¹⁶ http://simap.europa.eu/codes-and-nomenclatures/codes-cpv/codes-cpv_en.htm

User centric (i.e. youtube) and legacy media owners (newspaper and TV) are quickly migrating to the Internet. Indicators of media content are the average number of viewing time and total number of hits per item selected as a popularity indicator. Unfortunately not all sites provide indicators. It is only through CDN provider where such indicators could be retrieved.

7. To the future: the Internet of Things

The Internet today provides access to continuously increasing amount of information universally, at any time and from any device. In the evolving Internet of Things (IoT) landscape, any device equipped with sensors is essentially an information warehouse, capable of collecting and transmitting real-time data originating from and interacting with the surrounding environment (people, places and things). These types of data are invaluable for official statistics since they contain information about the everyday life of individuals and communities and environment.

There is a growing need and interest in this regard by the Commission highlighted in its report “Internet of Things in 2020: A roadmap for the future”, where the key topics identified were the “smart living” and “mastered continuum of people, computers and things”. There are a growing number of innovative social and human-centric application areas, including social networking, smart metering, smart data collection, environmental models and so on. It is clear that with the growth of Web 2.0 and the social media, a wide sharing of information and know-how is held and such social networking activities can be properly harvested for the benefit of official statistics.

However, data streams generated from sensors are not readily usable for computation of indicators. Applications which are able to exploit IoT data streams and at the same time capture social pulse are necessary. Social pulse can be captured from the Web 2.0 new generation of applications and particularly Location-based Social Networks (LBSNs), which enable users to publish their actual “real time” geographic location online. Recent advances in mobile and sensor technologies provide new possibilities for supporting services and users supporting activities that can be distributed and incorporate different physical and environmental sensory data.

Therefore sensor devices and social interactions along with powerful applications can provide data for calculating various indicators related not only to ICT use and their social impact but also to other financial and social indicators related to either individuals or enterprises. Sensor data can be used for official statistics related to agriculture, forestry, environment, urban traffic and accidents, travels, health services, tourism, natural disasters, etc. Interaction of sensors with humans through applications converting sensor data to natural language expressions and social media is a potentially interesting perspective for validating the quality of data. In any case, this potential source of official statistics requires powerful technological infrastructure.

8. Additional forward-looking remarks

For some time now, there is talk of the Internet as data source, of Big Data, of Open Data, now of Federated Data, organic data, and several other notions – overlapping to various extents¹⁷. Truth is, addressing such matters is exactly what we ought to be doing. We are all confronted with a paradigm shift for some time now, and struggle to really comprehend the implications, fight inertia, make choices, and plot specific courses of action. In the process, we are inevitably hampered by existing pre-conceptions and try to justify any transition to the new with familiar apparatus from the old. Key among this set, for our purposes here, are the issues of quality and efficiency (mostly costs). Avoiding the pitfalls of linear thinking (e-mail kills the post office, e-commerce kills retail and the like) and up-front realism will be definite assets.

Opportunities for data collection, and eventually the construction of indicators afforded by digital footprints, should not be examined under the lens of substitutability vis-à-vis the established norms. Even though it may be an inevitable, instinct-driven early step, such linear thinking does not do justice to what is becoming increasingly possible, and it may well lead to missed opportunities – at best (at worst it may lead to irrelevance of the existing official statistics system). It must be fully understood that the momentum cannot be stopped, and a willing transition does not need to be justified on cost savings, however convenient. While “moving to Internet-based collection for ICT statistics will save costs” would make a terrific sound bite, to use a very specific example, it would be missing the point. Bluntly put, not moving along is a price official statistics cannot afford to pay.

8.1. Dis-assembling and re-assembling

While it is out of scope of this particular report to take on such issues, a few words are in order. Right now we are in a transition and some crucial aspects appear blurred, inertia is still strong, we still do linear thinking, and are driven by cost efficiencies. All this is understandable, but we must develop a new mindset. It is perhaps honest to say and admit upfront that many of the gains will be in the volume and quality of future outputs – with processes different from the ones we know (what is more difficult to come to terms with is that all those too will be evolving)! It is akin to the desktop computer replacing the typewriter many years ago – the gains did not come from replacing the typewriter with a word-processor alone, but from the fact that the desktop computer brought with it numerous new applications too transforming many processes.

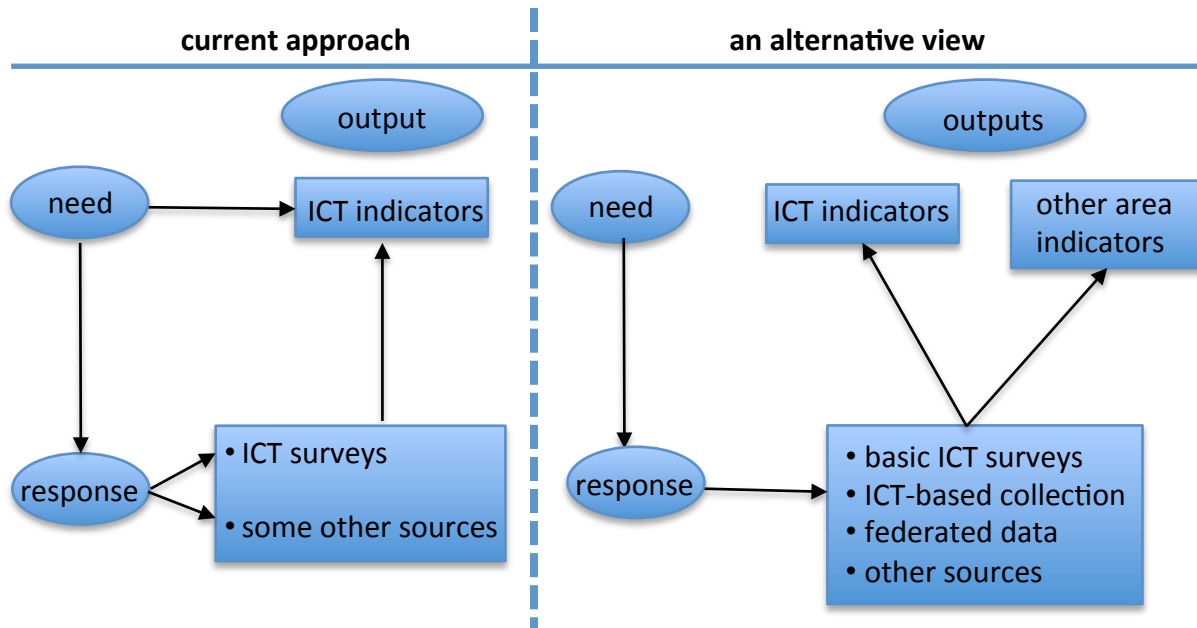
The new situation therefore calls for new models. Starting with some questionnaire of the traditional type, which responded to policy, business and general societal needs and attempting to fill it through digital footprints is not the correct approach. Surely, in the interim at least, there will be questions that will render themselves to such substitution, and others that cannot. For the most part, these can be known and articulated and part (a) of this report has done so. What is

¹⁷ Robert M. Groves, Keynote address on Big Data, <http://www.cros-portal.eu/content/ntts-2013-robert-m-groves-speech>

more important, though, are the data and indicators that can be had, and which did not make the cut in the early wish list – either because they were not thought of at the time or because the designers of the instrument thought they cannot be had.

We need to realize at a deeper level that the “whole” questionnaire/approach we had - and went after filling it in its entirety with one process - will have to be broken down to pieces that fit the new reality. The classic “you can’t simplify the real world to fit your model” applies. These pieces then will feed not only the old “whole” but also many more different “wholes”. This is fundamentally different from the habitual, and perhaps more painstaking, but the sooner we start to develop a degree of comfort, the better.

The following schematic displays simply what all this means. We can draw lessons, albeit imperfect, from familiar examples of integrating activities, such as the SNA satellite accounts. Under the habitual approach, data needs (typically advocated by policy makers) were met through a survey (entirely new or addition of modules to an existing one) – unless an administrative source existed (highly unlikely given that the new demands were associated with information concerning new and emerging phenomena). This essence of the approach connecting new needs to eventual statistical answers is depicted on the left-hand side of the schematic. Today, if we are to capitalize on the advent of ICT-based collection and/or federated data, more options for responding become available (right-hand side). Moreover, additional possibilities open up, which may turn orthodox processes upside-down. It is possible that in the process of tapping the new resources to answer a defined set of questions, answers to totally different questions can be fetched. As well, something much more intriguing is in the horizon: identifying the data that can be collected from where they exist (with the qualifications discussed earlier) and communicating such information to the demand side (e.g. policy makers), their thinking may be influenced in a way that they modify the questions asked. Thus, the interplay between data needs and responses elevates to another level. There may well be a link between perceived statistical needs and statistical outputs (not shown in the schematic).



At the same time, such transition will afford a prime opportunity to upgrade our core statistical infrastructure and prepare it for entrance to the new reality. A good example would be our registers – of businesses and of populations.

8.2. Some additional matters

We must be cognizant of the real underlying reasons why we are thinking along such lines, and indeed why this particular project is undertaken. We sense, if not know, that our digital times represent a structural change in the old order of things – which includes the way we go about measuring things. We are saying, simply and logically, that a lot of the data we need to understand the behaviour of individuals (or businesses) can come not from the individuals themselves but from their footprints in cyberspace. Then, we must explore and map the new world.

Quality and trade-offs: Collection of data, as we have historically known it, has been expensive. Rigorous methodological techniques for statistical sampling provided part of the answer, by making it possible to “measure” the needed characteristics of a population by questioning a fraction of its size. At the same time, the onus was placed on up-front decisions to design questions that would respond to the sought-after answers, complete with capturing, processing and analytical apparatus. Recently, several comments have been made to the effect that the new data differ from traditional collection methods in numerous ways, but one of them specifically refers to that they are more like “censuses” and somehow they may not be as representative as sample surveys. This is an interesting argument, albeit somewhat bizarre.

The beginning in which this argument derives its ostensible validity, seems to be the fragmentation of the population frame. When the interest is placed on producing a number of indicators, the universe on which these indicators apply must be well specified. In the case of the

enterprise survey, for instance, the universe includes specific NACE and businesses with 10 or more employees. In the survey of individuals, the line was drawn to individuals 16 years of age and older. Then, representative samples are drawn, with appropriate stratification, from those population frames and, through the use of individual weights, the survey estimates are blown up to speak for the entire population. Statistical theory is sufficiently advanced to be able to do that with relatively small samples, as well as produce indicators of quality for all the estimates produced with this method (for sampling errors). This method is the same regardless of the mode of collection, that is, paper, CATI, electronic questionnaire, CAPI or else. Their key distinguishing feature is that the answers come from the respondent – and that the questions were decided and specified precisely, after considerable thinking, by some “committee”. In the new data, the “answers” are all captive, and wait for questions that are useful to ask.

In a world where we move from this standard approach and exploit the power of digital footprints, big data, federated data and other sources, a number of issues must be understood that would precipitate adjustments in our comprehension. Take for instance the existence and use of smartphones among a part of the population of interest (for simplicity, but without loss, assume it remains at 16 years and over). Tapping into their digital footprints as a preferred substitute to get answers to questions outside the traditional survey method in which the answers come from the respondent, will clearly not speak for the population at large. This is not dependant on sample size and the like, as even if a census of all smartphone users was taken, and that size was bigger than the entire sample traditionally drawn, it does not represent the population – so long as there is a difference between the underlying population and that part that owns smartphones (and the bigger the difference, the bigger the bias). However, as explained in the previous section of this report, there is no reason to proceed with such an approach –particularly if not combined with other approaches that nearly exhaust the use sought, e.g. desktops, portables, tablets and even work and other places of use in a way that they can eventually be aggregated.

First, as has been stated before that given the proliferation of digital devices used by a single individual, it is not possible to use only one such device (e.g. a computer, a tablet, a smartphone) and get our answers. Therefore, tapping the digital signatures of smartphones will only provide part of the answers for one individual. This issue must be dealt with by factoring in other approaches, including resorting to the traditional method to the extent necessary (again, as done in this report).

Second, even if/when our object is only smartphone users, and because the technology allows us to track them all, it does not mean we have to do so. The reason is at least twofold: the digital age does not require throwing out statistical theory as we know it; stemming from the first reason, there will always be additional “outlets” to factor in our work and they can provide useful tests and benchmarking. (The latter of course, properly used, can improve the practical application of statistical theory as we know it. For instance, after tapping the data of smartphone users we may want to check them periodically against the same that are captured at the aggregate level by the service providers (perhaps federated data). In other words, there is another level

with the “truth”, something that has not been the case in traditional surveys). This helps sampling too, as it needs some estimate of variability – nothing better than having the true moments!

Back-end, tools and skills: Things become more complicated when the nature of the new approach is considered, as that too is not tantamount to a singular intervention. For instance, it may involve the insertion of a generic cookie, the design and installation of a specialized app, gaining control of a user’s computer (akin to a network administrator) etc. – all of which represent different degrees of “intrusiveness”. There is an imperative to think of impacts on quality as well, as there are trade-offs.

However, even addressing such technical issues in the most satisfactory way possible at any given time, the big issue is the back-end. What to do, how to do, and who will do?

We start to see a proliferation of analytics. Visualisation and tools for non-quantitative data also vie for attention. Part of the new skill sets that will matter as we move forward would be those that will enable us to manipulate, synthesise and decipher what part of the new possible is useful and why. Put differently, simply because something can be had does not make it desirable. This, points to the area of intercept of policy, business decisions and indeed societal evolution.

Caveats: In all the approaches discussed above, as in every new effort, there are not only advantages but we must be mindful of potential drawbacks as well that may well affect the quality and /or impartiality of the data. One of them concerns the need for consent in the case of either scraping the Website of an enterprise or having access to the traffic information of a smartphone. As this is not tapping digital footprints by stealth, particularly in the case of individuals, something must be done at some point to ensure that the data are not tainted because of modified behaviour of the individual for the duration of the test. While beyond the scope of this study, there is ample literature that the behaviour of individuals changes when we’re monitored. Typically, we try to show off our good selves avoiding habits and/or behaviours that we do not want others to see. While, as explained earlier, part of the new paradigm can take care of that at some level (benchmarking to aggregate data) another part of that such biases may not be inherently different from those obtained through questionnaires, anyways...

Other biases can come from intentional non-response depending on the mechanism. Part of the experience would be to start understanding the changed biases from refusals and so on with useful metadata – which, in all likelihood, will be subject to a different pattern than up to now.

Legal framework: Internet-based methods for data collection will surely bring about new implications in terms of Intellectual Property Rights, Data Protection and Privacy regulations. Consent and/licencing agreements will be necessary to ensure compliance especially during likely hand-overs related to value-added transformations of the data in the processing, aggregation and dissemination phases. Certainly, the persons involved will have to adhere to the conditions that will be set out, with regard to who has access to what and for how long. Moreover, the issue of confidentiality will have to be re-thought to arrive at an explicit

understanding of what data, at what level of aggregation, and under which conditions will be made available. At this point, we do not feel that sufficient progress has been made on these fronts. Thorny issues related to data confidentiality and the protection of privacy conflict seriously with information demands for information, especially microdata, whose manipulation has been made possible by technological advancements. Much more research will be needed to guide us through this labyrinth area, which is well beyond the scope of this project. In any event, we may have to re-think anew, and rationally, many of the established norms.

Social acceptance: The possibilities afforded by the digital era for statistics comes with a good deal of apprehension at present. Partly fueled by mishaps in violation of privacy, disclosure breaches or theft of personal data, a certain amount of confusion prevails among people, businesses and governments about what the future holds. At the same time, through millions of individual decisions daily, people make choices and effectively vote to expand their use of digital media. Social acceptance of using those very decisions to produce data is definitely conceivable, provided a governing framework starts to emerge and starts to be understood. At present most of the examples to which people can relate come from private firms with short lives, and it do not inspire much confidence. Frequent news stories of possible surveillance under “big brotherhood” from metadata left on choke points of the digital infrastructure and related to trade-offs between security and privacy also instill fears. Social acceptance will require more time, and demonstration of examples complete with societal benefits. The lead of impartial organisations, such as statistical authorities, could be welcome.

Realistically, we do not have immediately in front of us a case of replacing the existing questionnaires. There are opportunities though, for incremental and appropriate substitutions. This seems to be a desired path for the foreseeable future. To be clearer, more often than not, progress is not achieved by leaps and bounds or by radical departure from the familiar but follows an incremental trajectory. By exploring the data that can be obtained from the home computers/tablets or smartphones of individuals, progress is made – even if the data are not a complete replacement of existing collections. Realizing that users are prolific in their use of devices and focusing our subsequent efforts on the individual as the unit of observation would constitute a logical next step. What is more important is that we learn every step of the way, factor in the new knowledge in the modification of existing approaches, and maintain an open mind set.

9. References

- Berinsky, A. J. (2006). American Public Opinion in the 1930s and 1940s The Analysis of Quota-Controlled Sample Survey Data. *Public Opinion Quarterly*, 70(4), 499-529.
- Berners-Lee, T. (2006). Welcome to the Semantic Web. *The Economist - The World in 2007*. Retrieved from <http://www.neurophenomics.info/docs/semanticweb.pdf>

- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1. doi:10.4018/jswis.2009081901
- Capadisli, S., Auer, S., & Ngomo, A. (2013). Linked SDMX Data. *semantic-web-journal.net*. Retrieved from <http://www.semantic-web-journal.net/system/files/swj454.pdf>
- Cyganiak, R., Reynolds, D., & Tennison, J. (2012). The rdf data cube vocabulary. Retrieved July 1, 2013, from <http://publishing-statistical-data.googlecode.com/svn/trunk/specs/>
- Flake, G., Lawrence, S., Giles, C., & Coetzee, F. (2002). Self-organization and identification of web communities. *Computer*. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=989932
- Glasson, M., Trepanier, J., Patruno, V., Daas, P., Skaliotis, M., & Khan, A. (2012). WHAT DOES “BIG DATA” MEAN FOR OFFICIAL STATISTICS? Retrieved from <http://goo.gl/LfGIM>
- Martin, M., Straf, M., & Citro, C. (2005). *Principles and practices for a federal statistical agency*. Retrieved from http://books.google.com/books?hl=en&lr=&id=U3DiaH7IDCgC&oi=fnd&pg=PR1&dq=Principles+and+Practices+for+a+Federal+Statistical+Agency&ots=TW9o-7Nf_o&sig=BQdAPam7kAetmGqy-4qxiZEQxX0
- Newman, M. (2008). *The mathematics of networks*. (L. Blume, Ed.) *The New Palgrave Encyclopedia of Economics*. Palgrave Macmillan.
- Newman, MEJ. (2003). The structure and function of complex networks. *SIAM review*. Retrieved from <http://epubs.siam.org/doi/abs/10.1137/S003614450342480?journalCode=siread>
- Paparrizos, I., Koutsonikola, V., Angelis, V., & Vakali, A. (2010). Automatic extraction of structure, content and usage data statistics of web sites. *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (pp. 301–302). Retrieved from <http://dl.acm.org/citation.cfm?id=1810685>
- Questions and answers when designing surveys for information collections*. (2006) Office of Information and Regulatory Affairs Office of Management and Budget, January 2006.
- Saltzer, J., Reed, D., & Clark, D. (1984). End-to-end arguments in system design. *ACM Transactions on Computer ...*. Retrieved from <http://dl.acm.org/citation.cfm?id=357402>
- Tapscott, D., & Williams, A. D. (2008). *Wikinomics: How mass collaboration changes everything*. Portfolio Trade.
- Vafopoulos, M. (2011a). The Web economy: goods, users, models and policies. *Foundations and Trends® in Web Science*, 3(1-2), 1–136. doi:<http://dx.doi.org/10.1561/18000000015>
- Vafopoulos, M. (2011b). A Framework for Linked Data Business Models. *15th Panhellenic Conference on Informatics (PCI)* (pp. 95–99).
- Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709-747.

Zaiane, O. (1999). Resource and knowledge discovery from the internet and multimedia repositories. Retrieved from
<ftp://142.58.111.31/ftp/pub/cs/theses/1999/OsmarZaianePhD.pdf>

Zittrain, J. (2008). *The future of Internet and how to stop it*. Yale University Press.

12.3. D2 – Results of the feasibility analysis

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of
information society and other statistics’

D2. Results of the feasibility analysis

March 2014

Document Service Data

Type of Document	D2. Results of the feasibility analysis		
Version:	3	Status:	Draft
Created by:	Lefteris Angelis, Dimitris Kalogeras, Michalis Petrakos, Thanasis Priftis, Vasilis Sotiropoulos, Photis Stavropoulos, Michalis Vafopoulos	Date:	20/3/2014
Distribution:	European Commission – Eurostat/G4, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	6/12/2013	Initial release
2	31/12/2013	Revised version
3	20/3/2014	Revised version based on Eurostat's comments received on 15/1/2014

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

1. Introduction	4
2. Assessment of technical feasibility	4
2.1. Introduction	4
2.2. Network-centric methods	4
2.3. Web site-centric methods	6
2.4. User-centric methods	12
3. Feasibility within the conditions of the ESS.....	16
4. Methodological approach.....	20
4.1. Production of statistics on the characteristics of business web sites	22
4.1.1. Relevance	23
4.1.2. Accuracy	26
4.1.3. Coherence and comparability.....	28
4.1.4. Clarity.....	29
4.1.5. Timeliness	29
4.1.6. Conclusions about the statistics on the characteristics of business web sites	30
4.2. Production of statistics on the use of Internet by individuals	30
4.2.1. Relevance	31
4.2.2. Accuracy	32
4.2.3. Coherence and comparability.....	34
4.2.4. Clarity.....	35
4.2.5. Timeliness	35
4.2.6. Conclusions about the statistics on the use of Internet by individuals	35
5. Cost-benefit balance.....	35
5.1. Web site-centric methods.....	35
5.1.1. The site search market	36
5.1.2. Costs	37
5.1.3. Benefits and conclusion.....	40
5.1.4. To the future.....	40
5.2. User-centric methods	42
5.2.1. Costs	43
5.2.2. Benefits and conclusion.....	45
6. Legal feasibility.....	46
6.1. Introduction	46
6.2. Legal compatibility analysis	46

6.2.1.	Data protection terms and conditions.....	46
6.2.2.	Course of action for NSIs	50
6.3.	Data protection legal framework	50
6.4.	The <i>sui generis</i> Database Right	64
6.5.	Conclusion	68
7.	Socio-political acceptance	68
8.	Conclusions	73
9.	References	75
10.	Annex	76
10.1.	Appendix 1 - Synonym XML definition.....	76
10.2.	Appendix 2.....	79
10.3.	Appendix 3 – Topics for discussion with the NSIs for the assessment of feasibility in the ESS	80

1. Introduction

The first deliverable of project ‘Internet as a Data Source’, namely deliverable D1 ‘Definition of Internet data-based indicators’, proposed a number of Information Society-related statistical indicators on a) the use of Internet by individuals and b) on the characteristics of the web sites of enterprises. The aim of the present report is to examine whether the proposed indicators and methods for their compilation are feasible from the methodological and the practical point of view.

The feasibility analysis consists of the following elements:

- Technical feasibility (chapter 2)
- Feasibility within the conditions of the European Statistical System (ESS – chapter 3)
- Methodological feasibility (chapter 4)
- Cost-benefit balance (chapter 5)
- Legal feasibility (chapter 6)
- Assessment of the socio-political acceptance (chapter 7)

It must be noted that each aspect of feasibility is examined in isolation from the others. For example, when assessing the methodological feasibility of the methods, no concern is raised about their legal implications. Cross-references to the different chapters of the report are given when appropriate. Moreover, there are references to two additional deliverables of the project, deliverable D3 which presents the results of two pilot studies and deliverable D5 which discusses the evaluation of the potential of existing data sources to be used as input for official statistics.

The present report closes with the presentation of conclusions in chapter 8.

2. Assessment of technical feasibility

2.1. Introduction

Various paradigm shifts are shaping Internet usage in terms of the omnipresent social networking and the emerging WEB 3.0 with the linked data annotated with semantic information.

Previous work¹ has indicated the classification of Internet data-based methods in three categories, namely: user-centric, web-centric and network-centric. That study concluded that network centric methods had reached a plateau of technical feasibility while web site-centric and user-centric methods are open ended. In the following text we are going to re-establish the technical feasibility of the aforementioned methods as new techniques and usage scenarios emerge.

2.2. Network-centric methods

The previous study indicated that the metrics achievable with network-centric methods could estimate aggregate traffic, especially in InternetExchange Points (IXP), port-based statistics and deep packet inspection (DPI). Our experience on network-centric methods indicated the usage or access-based

¹ European Commission (2012) Internet as a data source. Luxembourg: Publications Office of the European Union.

mechanisms such as the introduction of sampling methods, transparent proxy/DPI and big data measurement.

General practical feasibility

Average throughput statistics and sampled flow mechanisms have proven their feasibility with an upper bound error of 10%-15% for specific aggregate and application-specific measurements. In general those mechanisms generate volume-based indicators and fail to address specific types of applications, which have transient port usage such as the p2p application.

Technical feasibility of network-centric method

Broadly speaking, large-scale network performance based on a limited subset of the nodes was initially addressed by Y. Vardi²³ who was one of the first to rigorously study this sort of problem and coined the term network tomography due to the similarity between network inference and medical tomography. This technology does not rely on the direct measurements on all physical transmission lines (links). It collects information at each end of the network (i.e. Origin and Destination, thus usually coined as OD estimation), eliminates the cost of deploying measurement equipment inside the network and reduces the volume of analysis data by up to 90%. Network partitioning extension can be used on large-scale networks to improve the speed of analysis even further. With this technology, the calculation speed can be reduced by one-half to one-twentieth.

Such type of techniques can be used in p2p traffic estimation in order to address the problem mentioned in the previous European Commission study, in the **Quality of data** section of network-centric methods, as the “Achilles’ heal” of the of network centric measurements: “the measurements would miss all the small branches of the Internet, that is, all the traffic flows that stay locally within other autonomous systems or local private networks”. The rationale of this method is based on the facts that:

- much of the broadband termination occurs in DSLAM (DSL Access Multiplexers)
- local IP cross-connect occurs near the broadband access concentrator, which is centrally located in the ISP.

hence it is possible to estimate the traffic for a certain type of applications without enormous capital expenditure.

In a similar context a Big Data technique based on probabilistic cardinality counting coupled with OD accounting⁴ is a technically feasible technique which can be applied on a large scale. This technique is scalable on a large scale and has been proved on a smaller scale on the Internet ²⁵.

However there is a trend in the dynamics of Internet usage. Following the differences between 2010 and 2013 according to the Global Internet Phenomena report^{6,7} bit-torrent submerges to online film viewing

² Y. Vardi. Network tomography: estimating source-destination traffic intensities from linkdata. J. Amer. Stat. Assoc., pages 365–377, 1996.

³ Network Tomography: recent developments, <http://ftp.stat.berkeley.edu/~binyu/ps/cny.pdf>

⁴ M. Cai, et. al., *Fast and Accurate Trafic Matrix Measurement Using Adaptive Cardinality Counting*,

⁴<http://gridsec.usc.edu/files/tr/tr-2005-12.pdf>

⁵www.internet2.edu/network/

⁶<http://www.sandvine.com>

rental such as Netflix. Youtube seems to capture the majority of http-based traffic. In order to follow this type of trend it is beneficial to measure central points (IXP) as it has been pointed out in the previous study along with targeted volume-based collection of specific sites as they are indicated by top viewing sites of popular visiting collection sites such as *alexa.com*⁸.

In terms of time/duration-based statistics the radius subsystem of fixed providers may provide useful data. Aggregate usage can be obtained in terms of daily, weekly, and even monthly basis. Aggregate data can be directed to a local statistics collection agency or to central European one. For mobile users, *always-on* is a typical case and hence duration based indicators have limited if negligible significance.

Quality of data and financial cost

Network-centric methods enjoy the best performance in terms of aggregate indicator-based metrics with solid theoretical background and have been verified in academic conferences and journals. Time-based indicators can be obtained relatively more easily from fixed providers and marginally from mobile operators.

However, there is a capital expenditure cost hidden in those methods. DPI has a floor of around 30keuros with no ceiling, while tomography and Big Data-based methods require custom development for the interested providers, which indicated a relative high price tag.

2.3. Web site-centric methods

Introduction

Deliverable D1 has progressed up to the point of defining indicators for quantifying aspects of the Information Society using the Internet as a Data (IaD) source for enterprises and individuals separately. In the proposed lists of indicators for enterprises a column describing the possible technical means points to crawlers as a respective tool.

Crawlers

Web crawlers, or crawlers or web robots are software systems which visit web addresses (i.e. URLs Universal Resource Locators in the terminology of W3 (WWW) Consortium) and copy their content to a local repository for later processing. A typical use of a Web crawler is common in Web search engines in order to facilitate indexing which is crucial for web searching. In our scope, a web crawler can be utilized in what is commonly called as web scraping a data mining process, which focuses on collecting specific parts of information of a web site and not the whole web site.

⁶ The Global Internet Phenomena Report (2013), <https://www.sandvine.com/downloads/general/global-internet-phenomena/2013/sandvine-global-internet-phenomena-report-1h-2013.pdf>

⁷ http://www.sandvine.com/downloads/documents/Phenomena_1H_2013/Sandvine_Global_Internet_Phenomena_Report_1H_2013.pdf

⁸ <http://www.alexa.com>

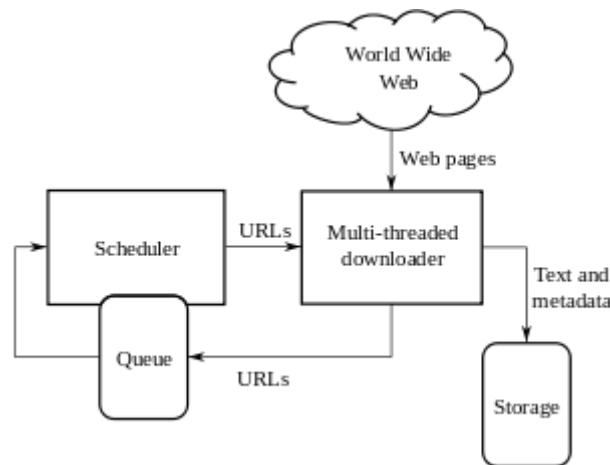


Figure 1. A typical block diagram of web crawler [from http://en.wikipedia.org/wiki/Web_crawler]

There is a multitude of free web crawlers available⁹. The most popular are the following:

- Wget¹⁰: It is one of the oldest web crawlers in the Internet (since January 1996). It is implemented in C, is available in most operating systems, among them MS Windows and Unix/Linux, and supports FTP and HTTPS in addition to the standard HTTP protocol.
- cURL¹¹: It is a command line utility for getting and sending file URL syntax. It utilizes the libCURL library (i.e. an aggregate software implementation artifact) for implementing numerous Internet protocols (Http, https, ftp, sftpimap, pop etc). cURL is implemented in almost every operating system.
- Heritrix¹²: It is currently the main crawler and indexer of the Internet archive¹³. Heritrix was developed jointly by the Internet Archive and the Nordic national libraries on specifications written in early 2003. It is implemented in java.
- scrapy¹⁴: It is a fast, high-level screen scraping and web crawling framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.
- DataparkSearch¹⁵: It is a search engine designed to organize search within a website, group of websites, intranet or local system.
- Norconex¹⁶: It is a web spider, or crawler, initially created for Enterprise Search integrators and developers. It began as a closed source project developed by Norconex. It was released as open source under GPL3 on June 2013.

⁹ http://en.wikipedia.org/wiki/Web_crawler#Open-source_crawlers

¹⁰ <http://www.gnu.org/software/wget>

¹¹ curl.haxx.se

¹² <http://crawler.archive.org/>

¹³ <http://www.archive.org>

¹⁴ <http://scrapy.org/>

¹⁵ <http://www.dataparksearch.org/>

¹⁶ <http://www.norconex.com/product/collector-http/>

- PHP-Crawler¹⁷: It is an open source crawling script based on PHP and MySQL. Created to implement as simple as possible local website search it became popular for small websites on shared hosting.
- Httrack¹⁸: It is a free and open source Web and offline browser implemented in MS Windows, Mac OS X and various Linux alternatives.
- iMacros¹⁹: It is a browser-based macro recorder which records and can repeat in the future the activities a human does on a website. Therefore, the software can record the data compilation activities carried out in the past by humans and carry them out itself. It is used by ISTAT, amongst other NSIs, for scrapping price data from the price lists of online stores (see Box 4, in chapter 3).

Static versus Dynamic Web Sites

The aforementioned utilities can handle web sites of the WEB 1.0 era with static content. Dynamic Web page creation via the Asynchronous JavaScript and XML or AJAX has revolutionized the Web, but it has also hidden its content. For instance, if you have a Twitter account it is not possible to view the source of your profile page. There are no tweets there—just JavaScript code! Almost everything on a Twitter page is built dynamically through JavaScript, and the crawlers cannot see any of it.

Although that might be the desired result for some web site it is seriously affecting its visibility as it is not involved in web search results. For that reason sitemaps provide important aid in web crawling and scraping activities.

A sitemap informs search engines and crawlers of website content. In other words, it provides to the crawler detailed information about the content of a website. Hence accessing and analyzing the sitemap of a web site, it is possible to mine the desired content.

Web pages that do not reside on sitemaps require more advanced techniques, e.g. in case everything in the page is built via JavaScript with hash tags²⁰. This situation appears to users as a fixed URL in their browser followed by a new hash tag for every different web page. In order to be able to crawl such dynamic content the AJAX web crawling technique should be adopted. This technique²¹ is based on the fact that when a crawler finds an AJAX URL (that is, a URL containing a #! hash fragment) it will request its content from the remote site in a slightly modified form. The remote server will return the content in the form of an HTML snapshot, which is then processed by the crawler.

It is also possible to use custom crawlers, which can cope with JavaScript. Although such effort requires manual customization for every site, which could easily consume projects resources, we mention the following capabilities:

- The Selenium²² regression web project with the WWW::Selenium module

¹⁷ <http://astellar.com/php-crawler/>

¹⁸ <http://www.httrack.com/>

¹⁹ http://wiki.imacros.net/Main_Page

²⁰ <http://en.wikipedia.org/wiki/Hashtag#Hashtags>

²¹ <http://coding.smashingmagazine.com/2011/09/27/searchable-dynamic-content-with-ajax-crawling/>

²² http://search.cpan.org/~lukec/Test-WWW-Selenium-1.23/util/create_www_selenium.pl

- Ruby's Capybara²³: an integration test library, which can also be used to write stand-alone web-crawlers. Given that it uses back ends like Selenium or headless WebKit, it interprets JavaScript out-of-the-box:
- Spider²⁴: programmable spidering of web sites with node.js and jQuery
- The Mechanize²⁵ library: used for automating interaction with websites. It automatically stores and sends cookies, follows redirects, and can follow links and submit forms. Form fields can be populated and submitted. With WWW::Mechanize::Firefox it is possible to let Firefox handle the complex JavaScript issues and then extract the resultant html.

Utilities vs. search engines

The various utilities must be customized around the specific content of each particular web site, even when hundreds or thousands of web sites must be scraped. This simulates, in many ways, a human web search, mostly in the sense of its logic of execution and the expected results. Therefore, the use of utilities in massive scraping demands excessive amounts of customization with respect to the acquired benefit.

Furthermore every indicator requires multiple lookups in order to cope with differences among sites. So instead of using a fixed search pattern scheme, we considered the use of probabilistic matching. This matching provides results when a portion of keyword is found. In our constrained timeframe we decided to use such techniques. Typically such techniques are found implemented in search engines.

Probabilistic search engines rank items based on measures of similarity between them and the search query, typically on a scale from 0 to 1, the latter score denoting perfect match, and sometimes popularity. The use of web search engines implicitly involves scraping, which was a requirement for indicator drilling, since the engines index web sites based on their scraped content.

Many web sites offer a site search option powered by one of the major web search engines. In our case though we need a search operation, which is neither limited in a specific web site nor does it expand to the whole web. We need it to search only among the web sites of a selected sample of enterprises.

Furthermore, a simplistic web UI interface with web search in the background is not sufficient; a programmatic interface is needed. The search engines which offer a programmatic interface are the following:

- [The google CSE \(Custom Search Engine\) from Google](https://www.google.com/cse/all)<https://www.google.com/cse/all>
- [The BING from Microsoft](#)
- [The Yahoo Boss from Yahoo](#)
- [Faroo.com](#)

The first three options have a small initial amount of free search queries and the rest of queries are paid. The last option is free. We decided to utilize the google CSE based on the following comparison table.

²³<https://github.com/jnicklas/capybara>

²⁴<https://github.com/mikeal/spider>

²⁵<http://mechanize.rubyforge.org/Mechanize.html>

Table 1. Comparison table of search engines with programmatic interfaces.

Search Engine	Limited search	Synonyms	Image search	Ads exclusion	Advanced Search	Internationalized Search
Google CSE	Y	Y	Y	Y	Y	Y
BING	N	N	Y	-	Y	N
Yahoo	N	N	Y	-	Y	N
Faroo	N	N	Y	-	N	N

The advantage of using programmable search engines instead of scraping tools and methods is that one needs not worry about infrastructure resources such as periodic crawling and scraping which would constitute the capital expenditure of the approach. One can just focus on the results.

This technical solution is bound to be replaced by a purpose-built software tool, which incorporates storage, computing (retrieval, indexing etc) and API specification for searching.

The Google leverage

Within this project's approach all text, multimedia and source code of a web site are considered as Internet content. Using Google CSE provides us with access to already structured content. This could act as a significant leverage to the project's goals but also to future efforts on automated Internet data collection enriching statistics indicators and their methodologies.

The second area of Google CSE leverage is the set of available search operators. These operators act as a simplified, easily accessible, regular expressions language: adding symbols or words to search terms in the Google search box allows for more specific results²⁶. Through this feature, text patterns, lists of words, complete words can be approached and better understood.

Image search is another potential capability. This could be beneficial when searching for Facebook or Twitter logos in web sites.

The implementation of web sites in different national languages is another potential barrier in search activities. Google CSE can return results no matter what is the language of a specific web site, provided that the search query includes terms in that language.

Sitemap indexing is another capability of Google CSE. As we discussed earlier the sitemap functionality allows for indexing of dynamic web sites. In this context it is possible to use either pre-crawled content or to have on demand indexing.

²⁶ Punctuation and symbols in search available

²⁶ <https://support.google.com/websearch/answer/2466433>

²⁶ Using Google Search Operators

²⁶ http://www.googleguide.com/advanced_operators_reference.html

HTML snippets

For verification of its results the CSE can provide HTML snippets together with URLs. The snippets are screenshots of a small part of the page corresponding to each search result which are designed to give a sense of what is on the page and why it is relevant to the query. For instance while searching for sites that contain telephone numbers the CSE provided among others the following results:

Box 1. Small sample of Custom Search Engine's list of results with HTML snippets.

[Shipping Companies - Patras Port Authority](http://www.patrasport.gr/?section=1638&language=en_US)

www.patrasport.gr/?section=1638&language=en_US



Central Agents: For Patras: PatraikaNautiliakaPraktoreiaS.A..Address: ΗρώωνΠολυτεχνείου 50. Post Code: 26441, Patra Phone: 2610-426000-10. Fax: 2610- ...

[EBETAM A.E. - Contact](http://www.ebetam.gr/?contact&lang=en)

www.ebetam.gr/?contact&lang=en



MIRTEC S.A. (Headquarters), A' Industrial Area, P.O.Box 13, GR-38500 Volos Tel : +30 24210 95340-2 Fax: +30 24210 95364, e-mail: volos.office@eb

In need of meta-search engines

The specific indicators that the present project deals with require scraping of separate web sites. In other domains, e.g. the estimation of price indices or of average prices a more evolutionary approach can be followed. It will focus on product catalogues and price lists generated by meta-search engines (i.e. aggregators). The data collection tools used by the producers of official statistics can target these aggregators, instead of individual web sites, and further process their results for every category under concern. For instance in house prices, aggregators prove quite more effective than official agencies. Facing that situation, NSIs could examine the possibility of using either some sort of proprietary backend API that the aggregators publicize, or standardizing some form of API in cooperation with other European agencies. It goes without saying that the aggregators' data should be examined beforehand to establish whether they are suitable as a data source for official statistics²⁷.

²⁷ This topic is the subject of deliverable D5 of the project.

European meta-aggregators

Some product markets show unexpected maturity in their European presence, which allows for Europe-wide price comparisons. For instance in the European e-bay market area it is possible to search for products and compare prices. E-bay provides feeds for product searches. So in general it is possible to orchestrate a specific multinational search. This capability can be exploited in a really convenient programming manner utilizing Yahoo Pipes²⁸. For instance http://pipes.yahoo.com/pipes/pipe.info?_id=trGzfqUY3RGDm_UvLO2fWQ gives an indicative result for multiple country search. If the share of eBay or other similar aggregators in the European consumer-to-consumer e-commerce transactions is sufficiently high it should be examined as a potential data source. At the moment this share is not publicly available.

2.4. User-centric methods

User-centric methods are by far the most interesting to watch because their dynamics are not limited by researchers or vendors. They are influenced by individuals who develop interesting applications in order to attract users and ultimately generate value. On a different aspect, if a national statistical institute (NSI) would like to invest on this type of Internet-based indicator collection they just need to order a suitable application from the market.

There is a major clustering between user-based methods focusing on desktop users and those focusing on mobile users. Mobile users are going to overtake fixed users in terms of number of end-devices²⁹. This argument can be backed by the recent ITU ICT statistics³⁰ where mobile subscriptions have overrun the fixed ones. Thus, modern user-centric collection methods have to be developed for both areas with mobile desktop application sector slowly overtaking the fixed based desktop environment. However the volume of consumption is largely devoted to fixed users while the duration of consumption is still shared between fixed (corporate – daily) usage and mobile (personal – leisure) usage. Hence in order to capture usable user-centric statistics there is need of coverage of diverse set of environments.

Another important clustering happens also with the smartphone shares, which is affected by the OS market share shown in the following table.

²⁸<http://pipes.yahoo.com/pipes/>

²⁹<http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/>

³⁰http://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2012/MIS2012_without_Annex_4.pdf

Table 2. Share of operating systems in smartphones sold.**OS % Share of Smartphone Sales**

	3 mo. ending Feb 12	3 mo. ending Feb 13
U.S MARKET	100%	100%
iOS	47.0	43.5
Android	45.4	51.2
RIM	3.6	0.7
Windows	2.7	4.1
Symbian	0.5	0.1
Other	0.8	0.4

(Credit: Kantar WorldpanelComTech)

This table shows that the major mobile environments are the iOS³¹ and Android³² one, with the last one slowly overtaking the first one.

A specific user-centric indicator collection method

Consider the bandwidth meter open source application (www.measurementlab.net/), which was mentioned in deliverable D1 of the present project. This application measures the available instantaneous bandwidth and logs the results on open repositories. This application has been build around a Java³³-based applet. Similar end-user applications such as *speedtest*³⁴ built on top of flash³⁵-based applications are also common in order to address the diversity of browser implementation. None of these applications are available in their own form in the mobile application environment, because neither flash, nor java are available in the most popular mobile platforms. Hence in order to account for a user-centric ICT indicator an NSI should account for a multitude of application environments including iOS and Android for mobile devices and java at least for desktop usage. There are various ways to achieve this: a) use an existing traffic monitoring service, b) installing a custom service that is application sensitive and c) propose a user customized DNS traffic service approach. The first two options share the same technical characteristics, meaning that they answer to multiple application environments by tracking internet user traffic coming from each active application. The third one is using the network Domain Name Service request of every user activity in order to categorize this traffic.

³¹<http://www.apple.com/ios/>

³²<http://www.android.com/>

³³www.java.com/

³⁴<http://www.speedtest.net/>

³⁵Adobe Flash, <http://get.adobe.com/flashplayer/>

Data from smartphones

Acceleration sensor data generated by smartphones can be analysed with the help of computer software. One example is PySensor³⁶ which can receive data from Symbian (e.g. N95) and Android phones. These software tools typically consist of an environment to work with acceleration sensor data as emitted by mobile devices. Acceleration data and key-press data are sent to a server (PC) where further processing, logging, recording, and visualization of the data is made possible. The data are then distributed to client applications. Optionally, the client applications may send image data at any rate and the device retrieves these images as fast as possible. Recorded input data can be replayed at original or altered speed. De Souza et al [Souza, 2010] introduced a framework to collect, modify, and distribute acceleration sensor data from multiple smartphones and integrate them with a medical imaging system which results in an environment suitable for e.g. doctors reviewing and explaining diagnostic findings.

Accelerometers along with GPS trackers and other information can be used as data source in social statistics especially for statistics and indicators related to citizens' use of time, physical activity, sedentary behavior and more generally statistics related to health and quality of life issues. Recent research entitled "Guidelines for Harmonising Time Use Surveys" prepared by the United Nations Economic Commission for Europe Task Force³⁷ on Time Use Surveys, mentions that:

*"Smart phones offer the possibility to bridge these various techniques, as applications.... can equip most phones to act as accelerometers, GPS trackers and collect other information automatically while also serving as the platform through which participants might complete their diaries. Already, examples of such surveys are in the field (for instance the **London School of Economics Mappiness**³⁸ project, which uses a smart phone application to collect GPS, diary and emotion information. For the moment, the distribution of smart phones varies significantly across countries, and across regions within some countries, but this distribution will change over time. Also, variety of the platforms on which future devices operate will pose challenges for the design of survey apps. Nevertheless, developments in technology will open opportunities to collect further types of information not presently available for official statistics or research. National Statistical Offices will need to consider the level of personal intrusion such advanced collection methods have and ensure this is compliant with their national privacy legislation."*

A review of **smartphone applications** for collecting behavioral and **psychological data** can be found in [Miller2012]. The paper gives an extensive comparison with traditional methods and presents challenges and problems which are more or less the same as the ones that have to be addressed in collecting data for official statistics.

An interesting paper, [Rofouei2012], describes a technique for associating multi-touch interactions to individual users and their **accelerometer-equipped** mobile devices. Real-time device accelerometer data and depth camera-based body tracking are analysed in order to **associate each phone with a particular**

³⁶ <http://code.google.com/p/pysensor/>.

³⁷ UNECE, www.unece.org/

³⁸ <http://www.mappiness.org.uk/>

user, while body tracking and touch contacts positions are analysed in order to **associate a touch contact with a specific user**. The technique is called ShakeID³⁹.

In [Beach2010] a system called **SocialFusion**⁴⁰ is presented, capable of systematically integrating diverse mobile, social, and sensing input streams and effectuating the appropriate context-aware output action. The interesting part of the paper is the explanation of some of the major challenges that SocialFusion must overcome. The paper discusses also new problems and therefore new research directions for preserving users' security and privacy and highlights:

“To collect data from users' mobile devices, a mobile application needs to be built that can identify the user's location and the phone's sensor values and pass on that information to SocialFusion. Given that a mobile device is typically used for a multitude of applications, this mobile application must be power efficient and power aware, preferably adaptive to the current remaining power of the device as well as non-intrusive to the user.”

LiveLab project is also proposing a methodology to measure real-world smartphone usage and wireless networks. The methodology is presented in [Shepard2011]. As noted, data regarding smartphone usage and user experience are imperative to the design and evaluation of techniques improving performance and efficiency of wireless Internet access and user experience. The paper has an interesting discussion of problems and challenges and how they are addressed by LiveLab, pointing out that

“...existing client-based network measurement solutions require time-intensive war-driving, ... which is unlikely to provide a fine-grained and dynamic network map. Wireless network and mobile users can also be measured from inside the network However, usage data collected by network operators are limited in both scope and detail. For example; they do not include applications that do not access the network. That is, cellular network carriers will be unable to collect data when a user is using WiFi. Furthermore, network operators rarely share their data with the research community, citing privacy and commercial concerns.”

The proposed **LiveLab** methodology aims at addressing these challenges by logging smartphone usage in the field, leveraging mobile users as a network sampling tool, and allowing the logger to be dynamically reprogrammed in the field.

Mobile device forensics

This area is not directly related to the collection of data for official statistics (actually it concerns data related to crime solving) but the advanced tools that have been developed for data acquisition could be possibly used for other purposes, such as studies or surveys.

In [Mokh2007] the use of an on-phone forensic tool to collect the contents of the device and store them on removable storage is proposed. Two other commercial products are mentioned and compared with the proposed tool: The **XRY** forensic software toolkit (<http://www.msab.com/>) and the **Oxygen Phone Manager** (http://download.cnet.com/Oxygen-Phone-Manager-II-for-Nokia-Phones/3000-2074_4-10054658.html) which are now commercially available.

³⁹<http://phys.org/news/2012-06-shakeid-tracks-action-multi-user.html>

⁴⁰<http://www.cs.colorado.edu/~rhan/mosonets.html>

Usage patterns of smartphones

The paper of [Kang2011] presents a usage pattern analysis of smartphones, i.e. how users use their phones. The methodology is interesting for extended use and for official statistics:

*“First, we define possible **smartphone states** based on their basic functions, e.g., voice call and data communication. Second, we define **log metrics** to measure time and battery spent in each operational state. Third, we **develop a mobile application** (called a **battery logger**) for collecting log data from real smartphones, deploy this application to our campus and online sites, and observe how it is used by real users. Finally, we analyze the collected data to show that each user has his/her own usage pattern. In this research, we **develop a battery logger** based on an **Android** mobile platform and **collect log data** from Android smartphone users. In all, we collect real smartphone usage logs from 20 users over a two month period.”*

The researchers collected the following data from smartphones with a mobile application they developed:

- Voice call status (Ringing, Waiting, Calling)
- Screen status (On/Off)
- 3G data communication status (In/Out/InOut)
- Active network (3G, WiFi)
- WiFi status (On/Off)
- Battery level (0–100 %)
- Battery status (Charging, Discharging, Full)
- Battery plugged status (Battery, AC, USB)

Device Analyzer

A large scale mobile data collection project [Wagner2013] was carried out by researchers from University of Cambridge. The results about usage information from 12,500 Android devices over the course of nearly 2 years (the contributors are increasing). The dataset contains 53 billion data points from 894 models of devices running 687 versions of Android⁴¹. This is an example of how academia is capable of conducting large scale projects which can produce invaluable usage data and research results. Obviously these projects need technical and human resources.

3. Feasibility within the conditions of the ESS

The current organisation of European information society statistics calls for the data to be compiled by National Statistical Institutes (NSIs) according to specifications agreed between the Member States and Eurostat. These specifications are issued in the forms of annual legal acts and methodological manuals. In view of these arrangements the feasibility of using Internet as a source of information society statistics is examined from the point of view of the NSIs in the present chapter.

The statistical legal basis of European information society statistics does not prescribe any specific mode for the collection of the data. Regulation (EC) 808/2004 as amended by Regulation (EC) 1006/2009 of the European Parliament and the Council and the annual Commission implementing regulations specify the

⁴¹<http://deviceanalyzer.cl.cam.ac.uk/keyValuePair.htm>

topics that will be covered in each data collection, the reference population of the statistics and the period of data collection. They do not forbid any particular collection mode and therefore automatic methods are a priori acceptable. Moreover, there are no clauses in these legal acts about privacy or protection of confidential data that could serve as basis for not allowing Internet data as a source for the statistics⁴².

The experimental nature of Internet data-based methods is not a forbidding factor for their adoption either. NSIs are on the lookout for more efficient statistical production methods. Ongoing projects that try to produce official statistics using Internet data or other big data as data sources demonstrate the NSIs' willingness to try new ways of doing their work. Some examples of such pilot projects are the following⁴³:

- **Tourism statistics:** mobile positioning data from mobile phones are used in order to estimate the trips of individuals to or from Estonia and the numbers of nights they spend in or outside the country. These short-term indicators are calibrated with official accommodation and travel statistics. The monthly statistics are used in the calculation of the national balance of payments. Statistics Netherlands also tests a similar method.
- **Collection of price statistics from e-shops:** an ongoing collaborative project to which participate several European NSIs tests methods and software tools for the automatic collection of price data from e-shops that will be used as input in the computation of the Consumer Price Index (CPI).

Discussions about the feasibility of Internet-data based methods were held with four NSIs. They revolved around the experiences they might have had with such methods and around their opinions about these methods in general (irrespective of whether they have applied such methods or not). The list of topics that was sent to the NSIs to prepare them for the discussion is shown in appendix 10.3. Details about each NSI are provided in the boxes that follow.

Box 2. Discussion with the Office for National Statistics (ONS).

Office for National Statistics (ONS)
<p>The discussion was held with members of staff from the department responsible for Information Society statistics. There is a different department which houses a “big data team” but it is located in different premises and therefore they did not take part in the discussion.</p> <p>It appears that the ONS takes very careful steps in its move toward new methods of statistical production. The major ongoing change in the office is the move from paper to electronic questionnaires, with the surveys remaining “traditional” in other respects.</p> <p>There have been however some experimental activities utilizing Internet or other big data but not in the domain of Information society statistics.</p> <p>The “Beyond 2011” programme seeks ways to exploit available and emerging data sources in order to</p>

⁴² The more general legal context of the protection of personal data is discussed in chapter **Error! Reference source not found.**

⁴³ Karlberg, M., Skaliotis, M. (2013) Big data for official statistics - strategies and some initial European applications. *UNECE, Conference of European Statisticians, Seminar on statistical data collection, Geneva, 25-27 September 2013, working paper nr 30.*

Office for National Statistics (ONS)

improve the production of population statistics. As part of the investigations

- a literature review on the uses of big data in official statistical production has been carried out;
- discussions have been held with private companies that process big data on the potential of using their results in the production of official statistics;
- the correlations of search data with particular statistical indicators have been examined.

More specifically regarding the work on search data

- the volume of searches about products correlates satisfactorily with the retail sales index for some product groups but not satisfactorily for the rest. The purpose of this investigation was to examine whether data on searches could be used for quality assurance of the index
- the correlation of the volume of searches from the UK in specific languages with the number of people in the UK that speak these languages has been investigated as a possible source for immigration statistics.

Another activity that was at an early stage when the discussion took place is the use of retail stores' loyalty card data as a source of data on retail sales. Unfortunately no more details could be provided about any of the activities.

There are however legal barriers to the use for statistical purposes of data that have been compiled for other, governmental or private reasons. In order to obtain even government data the ONS has to state with precision the uses that will be made of them.

Legislation in the UK obliges statistical units to provide data for official statistics. It is not clear however whether the law allows the statistical producers to choose any means of data collection they wish. Concerning the potential recording of individuals' activities in the Internet the feeling of our correspondents was that there will be reactions to it unless there are incentives. The "privacy lobby" in the UK is very strong and the discussants could not foresee what the impact would be if the ONS tried such recording, even with the consent of the individuals.

Box 3. Discussion with the Hellenic Statistical Authority (ELSTAT).**Hellenic Statistical Authority (ELSTAT)**

The team responsible for information society statistics in ELSTAT has not been aware of any ongoing activities in the NSI having to do with Internet data or big data in general. They have the impression that asking enterprises or individuals for automatic collection of data would cause negative reactions.

Moreover, in view of the workload of the team and the resources available to it they do not find feasible the introduction of such automated methods. Finally, there was negative reaction to the option of using for example big retailers' data as input; no argumentation was provided however.

Box 4. Discussion with ISTAT.**ISTAT**

ISTAT is pursuing specific Internet data and big data related activities. The most intensive relevant work is carried out by the Consumer Price Index (CPI) team. ISTAT participates in the project about price collection from e-shops mentioned in the main body of the chapter.

Product characteristics and price data about consumer electronics items are collected with web scraping

ISTAT

from the sites of big e-vendors and are used as input for the production of the CPI, purchasing power parities (PPPs) and the so-called “detailed average prices”. This work was previously carried out by humans who visited the sites regularly and compiled the price and characteristics data manually.

The automation of this operation has been achieved with the help of the iMacros software (see section 2.3).

The need to record human operations means that automation is achieved one site at a time. Moreover, even the structure of a site changes a human operator must visit it and carry out collection manually for iMacros to record it anew. At the moment the discussion with ISTAT took place (September 2013) the software had replaced humans for half of the consumer electronics e-shops included in the CPI sample.

ISTAT is very pleased with the performance of the automatic collection. It has sped data collection up by 30% and has increased the volume of collected data without deterioration in quality. Moreover, the software tools used are well within the capabilities of its IT staff. It plans to extend the use of these methods to all consumer electronics and then to other products whose prices are collected from the Internet by human operators. These items in total represent 23% of the CPI’s basket.

The product groups being investigated after consumer electronics are rail and plane tickets. The behaviour of a human inserting travel information and requesting ticket prices can be simulated by the software. A large technical obstacle however is the use of CAPTCHAs by the websites. At the moment there is no obvious way to overcome it.

The legality of this scraping is not clear to the CPI team. The owners of the e-shops have consented to the compilation of data from their sites by human operators; it is not clear whether they would be as consenting to the use of automated software.

ISTAT is also examining the use of retail stores’ scanner data as a source of price statistics for the CPI. This investigation is at its first stages, with discussions taking place with the association of retailers. A possible incentive to retailers for providing their data is the calculation by ISTAT of store-specific or retailer-specific inflation with official statistical methodology.

Besides possible legal issues and the problem with CAPTCHAs, the use of Internet or big data raises also methodological issues for the production of the CPI. One is the combination of Internet or big data with “traditional” price data. A second issue is that traditional sampling of price-taking locations breaks down with the abundance of price data available in the Internet; new approaches to sampling must be examined. Finally, the legislation concerning the coverage of the index might need to change. At present, special-offer prices are not allowed in the computation of the index. Scanner data however do not distinguish them from regular prices; this should not be a factor preventing the use of the data.

In the domain of information society statistics work with Internet data began in 2013. All enterprises included in the 2013 sample of the ICT enterprise survey which provided a website address in their response have been “visited” by a web spider. The software analyses the content of the site and tries to identify whether specific functionalities are available. The results are compared with the answers of the enterprises, to the regular ICT survey, about availability of the functionalities. The site owners were not informed about this collection.

If the results of the pilot are encouraging this spider-based collection will be applied to the sites of public administration authorities. At the moment of the discussion the collected data were still being analysed.

Box 5. Discussion with the Central Bureau of Statistics (CBS) - Netherlands.**Central Bureau of Statistics (CBS) - Netherlands**

The CBS is the more active of the four NSIs included in this feasibility assessment, in terms of the number of domains in which Internet and big data based methods have been investigated and in terms of extent of investigations. Seven domains of applications were discussed.

Automatic collection of Internet use data from smartphones and tablets. The CBS has contracted several studies about the use of monitoring software for the collection of Internet use data. Their results have been mixed and the main problem has been the very low response rate. The activity has seized due to lack of funding.

Housing prices. Statistics on the price per square metre asked by sellers of housing accommodation (“asking price”) are compiled based on data from the websites of large real estate agents. These data are combined with register and administrative data. The current activity follows on the steps of earlier experiments. It started with the evaluation of approximately 40 websites that could become data sources. The sites were assessed for the quality of the housing information that they contain. Five sites were selected for regular data collection and crawler software was built specifically for each one. At the moment one site prepares and delivers the data itself, with a financial grant from CBS covering its costs, while the others are visited by the crawlers. The CBS has signed agreements with two of these sites that allow data collection on a weekly basis. The coverage of the housing market by these data is not perfect but the volume of data is very large. The collected data amount to approximately 220 thousand houses per week.

Turnover of customer-to-customer (C2C) Internet sales. Six years’ product advertisement, auction and sales data were acquired as a batch by CBS from the Dutch equivalent of eBay. The acquisition cost 60 thousand Euros. Due to concerns about privacy the data of the individuals involved in the transactions were anonymised at the level of postcode. SBS has analysed the characteristics and prices of products advertised and sold and has combined the information with the location of sellers / buyers and socio-economic characteristics of the locations (from other sources).

Road traffic statistics. Several thousands of traffic detection loops have been installed on the Dutch road network and detect the passing vehicles and their length. Based on length CBS categorises the vehicles into length classes that are considered to correlate well with classes such as “private vehicles”, “trucks”, etc. The data enable the production of statistics about the number of vehicles by type of vehicle (length class), location (break down of Netherlands into four regions) and hour of the day. At the moment the statistics are produced on a quarterly basis but there are plans for monthly statistics at a finer regional division. The coverage of the road network is still not perfect but is expanding.

Tourism statistics. The CBS is collecting mobile phone positioning data from mobile telephony companies and uses them in order to produce statistics about inbound tourism to the Netherlands. At the moment, no data are compiled on outbound traffic. Due to privacy concerns the data are delivered to the CBS in the form of aggregates that comprise at least 15 persons per data item.

Job vacancy statistics. This is an example of the CBS receiving processed data from a private company. The company in question compiles data about jobs being advertised in jobseeker websites and produces clean micro-data in the form of one record per vacancy without duplications. The data cover the “Dutch Internet”, i.e. all websites with content in the Dutch language. It is not clear whether this excludes Flemish sites or sites in Afrikaans (S. Africa). The quality of the micro-data appears very good to the CBS and it is overall very satisfied with the results. The trends of statistics produced from these data are very comparable to those of regular job vacancy statistics; the levels of the series on the other hand are

Central Bureau of Statistics (CBS) - Netherlands

not comparable. There are still some concerns also about the representativeness and sectoral coverage of the data. Finally it should be noted that the processing carried out by the private company for the production of the clean micro-data is not known to the CBS.

Consumer confidence index. The CBS is producing a sentiment index based on social media messages. Similarly to the case of job vacancy micro-data, a private company analyses all public messages posted in Dutch, removes those consisting of “pointless babble” and assigns a sentiment score to the remaining ones: 1 for positive messages, 0 for neutral ones and -1 for negative ones. The average of the scores is the index. The messages analysed are public messages only, overwhelmingly consisting of Facebook status updates and Twitter tweets. Daily movements of the index are very volatile but weekly and monthly movements are stable and correlate strongly with the regular Dutch Consumer Confidence Index especially in what regards consumer’s confidence about the situation of the economy.

The legal context is not clear, at least in the opinion of the correspondents. Statistical legislation does not contain clear statements concerning the types of data collection envisaged in these activities. Moreover, there is a prevailing attitude that scraping large volumes of data, which have been produced for profit by a private company, might be problematic.

According to our correspondents there is no obvious way to assess the quality of statistics produced on the basis of Internet or big data other than comparing them with corresponding regular official statistics.

The correspondents think that a new skill-set, that of the so-called “data scientist”, combining statistics, mathematics and IT skills is required for the usage of such methods by NSIs.

The picture that emerges from these discussions is firstly one of “no objection” to the new methods. Excluding the attitude of ELSTAT, which probably is due to its lack of acquaintance with them, the other NSIs view the new methods favourably as production tools, in principle not different from the other methods they use. They experiment with them and assess them with the same procedures they assess the quality of production processes. They are concerned about the accuracy of their results but in most cases they find it satisfactory, while they recognise the gains in timeliness they offer.

The legal setting is not clear for any of the NSIs. It is not clear to them if the consent of individuals or enterprises whose data are collected or of the owners of the data is sufficient to make the methods “legal”. Some of the scraping experiments in fact have been conducted without the site owners being aware of the scraping.

Leaving legal feasibility aside, the new methods seem feasible in the context of the ESS. They should be discussed with ESS partners and be “promoted” by their exponents like any other production method. This, together with methodological support should go a long way in ensuring their adoption by the Member States.

4. Methodological approach

In this chapter we act as if processes are in place for the production of statistics on the facilities of business web sites and on the use of Internet by individuals. Under this assumption we examine the

processes from the methodological point of view. Do they produce the indicators that they are intended to produce? Is their quality at the required level?

The framework for assessment of statistical quality used in the ESS⁴⁴ is a suitable vehicle for the assessment in this chapter. The quality dimensions employed in the ESS will be used:

1. Relevance. *'Relevance is the degree to which statistical outputs meet current and potential user needs. It depends on whether all the statistics that are needed are produced and the extent to which concepts used (definitions, classifications etc.) reflect user needs.'*
2. Accuracy. *'The accuracy of statistical outputs in the general statistical sense is the degree of closeness of estimates to the true values.'*
3. Coherence and comparability. *'The coherence of two or more statistical outputs refers to the degree to which the statistical processes by which they were generated used the same concepts - classifications, definitions, and target populations – and harmonised methods. Coherent statistical outputs have the potential to be validly combined and used jointly. Comparability is a special case of coherence and refers to the ψασ€ where the statistical outputs refer to the same data items and the aim of combining them is to make comparisons over time, or across regions, or across other domains.'*
4. Accessibility and clarity. *'Accessibility and clarity refer to the simplicity and ease with which users can access statistics, with the appropriate supporting information and assistance.'* Accessibility is not shaped by the methodology used for the production of statistics. It is a matter of the means used to disseminate the statistics and is therefore of no relevance for this assessment. We therefore assess only clarity under this heading.
5. Timeliness and punctuality. *'The timeliness of statistical outputs is the length of time between the event or phenomenon they describe and their availability. Punctuality is the time lag between the release date of data and the target date on which they were scheduled for release as announced in an official release calendar, laid down by Regulations or previously agreed among partners.'* Punctuality refers to the respect of production and dissemination deadlines by the NSIs and Eurostat and is also not of relevance for the present assessment. We therefore assess only timeliness under this heading.

The assessment is carried out separately for each of the two envisaged processes.

4.1. Production of statistics on the characteristics of business web sites

The process analysed in this section produces statistics about the number or (equivalently) the proportion of enterprises whose web site possesses a number of characteristics. A separate indicator (number or proportion) is defined per characteristic. One example: proportion of enterprises whose web site contains a list of its products or services. The statistics belong to the domain of Information Society.

Data collection is automated and achieved with the use of software tools called crawlers. A random sample of enterprises is drawn from the business register, which should contain the URLs of its entries among the available contact information. The owners of the enterprises or the managers of their site are informed about the survey and give their consent to participate in it; they also provide the enterprise's

⁴⁴ Eurostat (2009) ESS Standard for Quality Reports. Luxembourg: Office for Official Publications of the European Communities.

URL if it had not been listed in the register. The crawler visits the web site and extracts or analyses in real time the content of all pages and a list of the web technologies implemented in the pages.

The purpose of the analysis of these data (content and technologies) is to identify specific keywords or technologies, which the producer of official statistics considers as proxies for the existence of target characteristics. If at least one keyword or technology is found in at least one page then a binary variable corresponding to the characteristic takes value 'YES' for the enterprise in question. If no keyword or technology is detected the binary variable takes value 'NO'. The number of 'YES' in the sample is used for the estimation of the corresponding indicator.

Part II of deliverable D6 of the project is a proposed cookbook for a possible implementation of the process. Moreover, chapter 3 of deliverable D3 presents a pilot implementation of the process.

4.1.1. Relevance

The number of indicators is equal to the number of target characteristics. A list of characteristics, by no means exhaustive, is the following.

Table 3. List of enterprise web site characteristics.

Characteristic	Definition	Comments
Contact information - URL	The site lists a URL (web address) among the contact information that it provides to visitors; this may or may not be the same as the main URL of the site	
Contact information - Email address	The site lists an email address among the contact information that it provides to visitors	
Contact information - Telephone number	The site lists a telephone number among the contact information that it provides to visitors	
Contact information - Postal address	The site lists a postal address among the contact information that it provides to visitors	
Availability of the web site in the national language	At least one of the pages of the web site is provided in the national language.	
Availability of the web site in English	At least one of the pages of the web site is provided in English.	
Availability of "last updated" date	The site lists the date on which it was last updated.	
Availability of privacy policy	The site displays (or provides a link to a document containing) the privacy policy of the site. This is a description of the use of personal information - particularly personal information collected via the website - by the website owner. It also describes measures taken to guarantee secure handling of financial information.	Relevant indicator produced from the regular ICT survey too.
Availability of registration facility	The web site has facility for users to sign up and then sign in.	
Availability of personalised content for regular/repeated visitors	The web site has the ability to recognise the user from previous visits (login/password) and adapt the content of the pages accordingly.	Relevant indicator produced from the regular ICT survey too.
Availability of site map	A site map is a list of pages of the web site accessible to crawlers or users. It can be either a document in any form, or a web page that lists the pages, typically	

Characteristic	Definition	Comments
	organized in hierarchical fashion.	
Display of the number of visitors	At least one page of the web site displays the number of visitors since a - listed too - given point in time.	
Availability of product catalogues	The web site provides lists of products or services offered by the enterprise to its clients. They might include also the characteristics of these products or services. The information may be static or dynamic (extracted online from a database and as such always updated).	Relevant indicator produced from the regular ICT survey too.
Availability of price lists	The web site provides provides a product catalogue which includes prices.	Not common for certain types of enterprises, e.g. in the services sector. Relevant indicator produced from the regular ICT survey too.
Possibility for site visitors to customise or design the products	The web site provides an interactive interface where users can choose from several possible characteristics of the products (colour etc.) or services and see online in the site the impact, for instance, on the price. The interface might also include the possibility for the user to visualise the appearance of the product with the options that were selected. The carrying out of simulations or any calculations (e.g. what-if calculations) for products like loans in the financial sector, belongs here as well.	Relevant indicator produced from the regular ICT survey too.
Availability of online ordering or reservation or booking facility	<p>The web site provides a facility which allows the user to order products or services with no additional contact offline or via e-mail required (for the ordering). A shopping cart and checkout facility is such an example. It includes also the facility for reservation of hotel rooms or the booking of flights.</p> <p>It does not include a link in the website which directs the user to an e-mail application which requires the user to send the order via e-mail. Payment may or may not be included in the ordering facility, e.g. payment may be made on reception of the product or by other means other than electronic payment.</p> <p>Carrying out a transaction via online banking in general does not qualify as online ordering; specific cases however, e.g. when buying shares (with a commission to be paid to the bank), qualify as online orders in the banking sector.</p>	Relevant indicator produced from the regular ICT survey too.
Availability of online order tracking facility	The web site provides facility that aims to keep the customer informed on the progress of the ordering and delivery process.	Relevant indicator produced from the regular ICT survey too.
Listing of open job positions or availability of	This item includes both cases where just simple information on job vacancies is provided in the web site	Relevant indicator produced from the

Characteristic	Definition	Comments
online job application facility	as well as those where the site provides also an online facility for candidates to apply for the jobs.	regular ICT survey too.
Number of open job positions in the enterprise, listed in the web site	The number of job opening listed in the web site.	
Availability of links to multimedia content (audio, videos, etc)	The web site provides links to multimedia content hosted in the servers of the enterprise.	
Availability of links to content in multimedia sharing sites (YouTube, Flickr, etc)	The web site provides links to multimedia content hosted in multimedia sharing sites.	
Availability of links to social networks or blogs (Facebook, LinkedIn, Yammer, Twitter, etc)	The web site provides links to social networks or blogs.	
Availability of links to wikis and wiki-based sharing tools	The web site provides links to wikis and wiki-based sharing tools.	

Due to the mode of data collection used, data collection can be over for a sample of some thousands of enterprises within a matter of days. The reference period of the indicators is the moment of data collection and statistics can be disseminated at quarterly intervals.

The statistics measure the richness of content and the sophistication of business web sites. Nowadays, web sites are one of the main channels of communication between enterprises and consumers. They are used for dissemination of information and marketing, for transactions with customers, for receiving feedback from customers. They serve as an additional storefront, especially in sectors that have embraced the digital economy. Their sophistication reveals the attention paid to them by the owners of the enterprises and the importance they attribute to the Internet as a driver for profits. Furthermore, by examination of the technologies used in them, the extent of innovation in web site design that is diffused in the economy can be studied. From this point of view statistics about the web sites are highly relevant Information Society indicators.

In addition, the information extracted by the crawlers can be post-processed and additional indicators, not foreseen at the time of collection, can be computed. Such computation may be needed to extend backwards, towards our present, time series of indicators that will be deemed important in the future.

On the other hand, these indicators do not cover and cannot cover all relevant information about ICT in the business world. They correspond to a small subset of the statistics produced by the Community Survey on ICT usage and E-commerce in Enterprises. This does not diminish their relevance but shows that they must be complemented by at least a) indicators based on data extracted from proprietary, offline⁴⁵, servers of the enterprises and b) questionnaire-based data.

⁴⁵ Offline as in not being accessible by external visitors via the enterprise web site.

A more serious disadvantage of the process is its reliance on proxy variables. The presence of the target characteristics is implied by the presence of specific keywords or technologies. While the latter could sometimes have an 1-to-1 mapping with specific characteristics this is rarely the case with keywords. As the results presented in Table 4 below show, the keywords cannot be very characteristic-specific; they can appear quite frequently even when the assumed characteristic is absent. To the extent therefore that the statistics measure the presence of keywords instead of characteristics their relevance is reduced.

The conclusion is that the detection of technologies shows more promise in producing well-specified, relevant statistics about the characteristics of web sites.

4.1.2. Accuracy

The accuracy of survey-based statistics is the result of the joint effects of sampling and non-sampling errors. As the discussion will make clear, the major concern are non-sampling errors caused by a) the use of keywords as proxies for the presence of specific characteristics, and b) non-response, mainly refusals from enterprises to participate. For some indicators the bias caused by keywords is very high. Refusals will lead to unit non-response affecting equally all indicators.

4.1.2.1. Sampling error

The pilot survey that was implemented in the context of the present project did not manage to obtain a random sample from a proper business register. The sample was not even random, because the only frame that was obtained contained a small number of enterprises and they were all included in the pilot. Therefore there are no estimates of the standard error of the statistics.

On the other hand, the envisaged implementation of the process relies on sampling from the business register of the NSI. Its sampling error will therefore be comparable to those of the other national business surveys. Its magnitude will be the outcome of the sample design and the sample size chosen and of the prevalence of the different characteristics amongst the national enterprise web sites.

4.1.2.2. Non-sampling errors

There are several possible types of non-sampling error.

Coverage errors are caused by imperfections in the sampling frame or in the sample selection procedure, which cause the population represented by the sample (called the ‘frame population’) to differ from the desired target population. In the envisage case of relying on the business register of the NSI and using a sampling procedure that is also used in other business surveys, the coverage errors will be comparable with those of the other surveys.

The pilot survey relied on a very small list of enterprises and therefore there is no point in even assessing coverage.

Measurement errors *‘are errors that occur during data collection and cause the recorded values of variables to be different from the true ones’*⁴⁶. As the pilot study carried out in this project showed, the process that relies on the detection of keywords suffers from such errors.

⁴⁶ Eurostat (2009) ESS Standard for Quality Reports. Luxembourg: Office for Official Publications of the European Communities.

It turns out that keywords do not allow neither sensitive nor specific measurements. The findings are summarised in Table 4. The table shows the percentage of sites that had each characteristic and divides into those where it was detected with the help of keywords and those where it was not. Moreover, it divides those that did not have each characteristic into those that were wrongly indicated as possessing it and the rest.

Sensitivity is measured by the proportion of sites with the characteristic, which was indeed detected. It ranges from 0% for links to social networks, blogs and multimedia content up to almost 90% for site map and contact email address.

Specificity on the other hand is the lack of ‘false positives’, i.e. wrong identification of sites as possessing a given characteristic. The share of false positives ranges from 2% for contact URLs up to 100% for links to wikis.

Table 4. Specificity and sensitivity of keyword-based detection of web site characteristics: shares (%) of a pilot sample of 281 enterprises.

Characteristic	Characteristic is present		Characteristic is absent	
	Detected	Wrongly not detected	Wrongly detected	Not detected
Contact URL	14.8	8.2	52.5	24.6
Contact email address	77.0	9.8	9.8	3.3
Contact telephone number	75.4	21.3	1.6	1.6
Contact postal address	55.7	36.1	3.3	4.9
Pages in the national language	65.6	13.1	14.8	6.6
Pages in English	59.0	24.6	4.9	11.5
Date of last update	0.0	0.0	0.0	100.0
Privacy policy of the web site	16.4	11.5	1.6	70.5
Site map	34.4	4.9	3.3	57.4
Use of web analytic tools	1.6	0.0	8.2	90.2
Announcement of open positions or provision of forms for applying for a job online	21.3	11.5	9.8	57.4
Links to social networks or blogs	0.0	29.5	0.0	70.5
Links to wikis and wiki-sharing tools	0.0	0.0	14.8	85.2
Links to multimedia content	0.0	27.9	0.0	72.1

This is the result of the use of keywords. Examples are given in Table 5. Keywords may be present without the respective characteristic being present. For example, the word ‘telephone’ will be used in a page listing contact information but it may also be used in a different context, e.g. the company apologising it its site ‘... for our helpdesk telephones not been operational yesterday morning’. On the other hand keywords not thought of may be used in other web sites which have the desired characteristic and their presence will go undetected.

Table 5. Examples of web site characteristics and matched keywords.

Characteristic	Keywords
Contact information - URL	url, Website
Contact information - Email address	e-mail, Email, E-mail, email, eMail, E
Contact information - Telephone number	telephone, telephone number, Phone, Tel., Fax, Tel/Fax, T:, tel, TELEPHONE
Availability of the web site in English	Language, English, EN
Availability of "last updated" date	Last Update, Last Updated Dated
Availability of privacy policy	privacy policy, terms of use, Privacy Statement, Conditions of use, Terms and Conditions, Terms & Conditions, Privacy, Legal, DISCLAIMER, Disclaimer, Copyright
Availability of registration facility	Signin, login, Login, register, Create an Account, openID, registration, Subscribe
Availability of links to multimedia content (audio, videos, etc)	mpeg,
Availability of links to wikis and wiki-based sharing tools	wikis

The use of detected technologies as proxies for web site characteristics was not tested in the pilot survey due to lack of resources and to time constraints. Therefore there are no indications about possible measurement errors.

Processing errors do not affect the process. All variables are binary, indicating presence or absence of keywords or technologies. The processing is similar to the processing carried out in other business surveys.

Non-response errors are caused by respondents not providing any data or providing only a subset of the data. The latter case would correspond to only parts of the web sites being accessible by the crawlers, which is not very common. The former case however, which amount to ‘unit non-response’ can be common and in fact more common than in other business surveys. The use of crawlers may look like the use of malignant software of the kind that spies on web sites or makes denial-of-service attacks. Therefore refusal rates may be higher than in other business surveys as the site owners may start asking for use of a questionnaire instead of software. The NSI must make efforts to reverse the negative climate by explaining the nature of the collection and of the data and by offering possible incentives.

4.1.3. Coherence and comparability

The envisaged survey is a business survey, which, with the exception of the measurement process and the definitions of the variables on which data are collected, operates with the same concepts and procedures

as other business surveys. It is under this light that its coherence with other surveys and its comparability across countries and over time are assessed.

4.1.3.1. Coherence

The survey uses the same definition of enterprise as the other business surveys of the ESS, classifies the target population by region, economic activity and company size, using the same nomenclatures as the other business surveys and draws its sample from the national business register.

Its ‘novelty’ lies exclusively in the survey variables, indicating presence or absence of characteristics and mainly in the data collection mode. These however, are not affecting its coherence with other surveys. Coherence with other business surveys is very high; coherence with non-business surveys is at the same level as that of the other business surveys with them.

4.1.3.2. Comparability

The approach of using keywords as proxies for characteristics of the web sites leaves too much room for differences between countries. This is the opinion of the project team, although it could not be tested in the pilot survey, and the rest of the section provides our reasoning.

The selection of the right keywords, with all the shortcomings of the approach in what regards sensitivity and specificity, is kind of an ‘art’. It requires that keywords are determined by experts with good knowledge of the way national web sites are designed. It is not certain that this activity will be implemented equally well in all countries and therefore each country may be measuring a different subset of each target characteristic: aspects of the characteristic not expressed with the selected keywords will remain undetected. A remedy for this problem could be the ‘central’ determination of keywords, e.g. by Eurostat or a working group.

This however faces the obstacle of linguistic differences between countries. The terms being used in each country, which could be the proxies for a characteristic, may not be direct translations of the terms of other countries. In other words, we believe that the recourse to national experts cannot be avoided.

Comparability over time will suffer too but for different reasons. Similarly to the current ICT survey, the set of target characteristics will probably need to change often, to reflect changes in information technologies. However, if the data collected in previous rounds are available, they will be re-processed for the computation of the new indicators, as long as they too rely on keywords. Historical series will then be re-constructed, resolving the comparability problem. On the other hand administrative or legal reasons, e.g. the requirement to delete the collected data (site content) after a given amount of time, will make impossible the re-construction of time series.

4.1.4. Clarity

The produced indicators are quite straightforward to understand, even for laymen. Some technical expertise is required for their definition and their subsequent expression in more simple terms. We do not foresee any major issues in this respect.

4.1.5. Timeliness

The speed of data collection, processing and production of statistics is un-matched by the current ICT survey. Data collection and processing could be over in less than a month. The slowest stage of the

process is expected to be the communication with enterprises in order to get their permission for data collection. Even with this stage included, the whole process could take less than three months. The timeliness of the statistics will therefore be very high.

4.1.6. Conclusions about the statistics on the characteristics of business web sites

The envisaged statistics are very relevant for the measurement of the information society, since they express the sophistication of business web sites and their role in the activities of the enterprises owning them. Moreover, the use of crawlers for data collection automates their production and greatly reduces the time required for one production cycle. This leads to very timely statistics, available in very few months after the end of the reference period. Relevance and timeliness are the two great strengths of the approach.

The drawbacks of the approach are three. The reliance on keywords as proxies for the possession of the target characteristics by the web sites can cause serious bias in the statistics. Moreover, the use of crawlers for data collection may cause concerns to site owners and lead to large refusal rates and therefore unit non-response. Finally, linguistic differences between countries and varying expertise in the selection of keywords between countries may reduce the geographical comparability of the statistics.

The conclusion of the project team is that a survey encompassing all possible characteristics of a business web site will suffer from reduced accuracy. The approach should be used only for carefully selected characteristics, which can be mapped, with an 1-to-1 mapping, to specific technologies rather than keywords. Only then can accuracy improve to a point that the approach is appropriate for official statistics. This however requires further testing.

4.2. Production of statistics on the use of Internet by individuals

The process analysed in this section produces statistics about the number or (equivalently) the proportion of individuals who engage in a number of activities on the Internet, the time (duration or share of total time) that they spend on them and the type and amount of data downloaded from or uploaded to online sites. Some examples:

- Proportion of individuals who are taking an online course.
- Amount of time that the average individual spends per day on online gambling.
- Amount of music data that the average individual downloads per day.

The statistics belong to the domain of Information Society.

Data collection is automated and achieved with the help of monitoring software tools installed on the users' devices (computers, smartphones, tablets). A random sample of individuals is drawn from the usual sampling frame of the NSI. The selected persons are informed about the survey and give their consent to participate in it. They also answer some screening questions about whether they use the Internet and the devices they use to access it. The individuals then receive the installation files and instructions from the NSIs and install and activate the software for a fixed, specified period of time. During this time the software records the applications launched and the sites visited and the times on which these activities start and finish. The users have the ability to switch it off and on at will; therefore they can avoid to have certain activities monitored.

Applications and web sites are mapped, in advance or during processing of the data, into target activities that are of interest for measurement, e.g. education, health, games, gambling, news, etc.

Part I of deliverable D6 of the project is a proposed cookbook for a possible implementation of the process. Moreover, chapter 2 of deliverable D3 presents a pilot implementation of the process.

4.2.1. Relevance

The number of indicators depends on the number of target categories of activities and the number of types of data that can be downloaded or uploaded. A list of categories, by no means exhaustive, is the following.

Box 6. List of categories of Internet activities.

- Using cloud storage facilities
- Doing an online course (in any subject)
- Education activity, other: time spent on online activities / web sites related to education, but not to doing an online course, e.g. searching for information about courses.
- Email
- Employment: time spent on online activities / web sites related to employment.
- Entertainment
- Finding information about goods or services
- Forums
- Gambling
- Games, unspecified
- Government: time spent on government web sites.
- Listening to web radio
- Networked games
- Social networks
- Playing online, but not networked games
- Adult content
- Reading news
- Shopping
- Sports
- Technology
- Telephoning / video calling (via webcam) over the internet
- Using services related to travel or travel related accommodation
- Viewing / listening to online images, videos, music
- Internet, other: time spent on online activities / web sites that cannot be classified in one of the other categories.
- Offline: time spent on offline activities.
- Not clear: time spent using applications for which it cannot be distinguished whether they involve online activity or not.

Due to the mode of data collection used, data collection runs in parallel for the whole sample irrespective of size. The reference period of the statistics is the period of data collection and they can be disseminated very quickly after it, e.g. at quarterly intervals.

The Internet is omnipresent nowadays and occupies an increasing share of individuals' time, especially younger persons and persons doing desk-bound work. It is usually “on” in the background of other activities people do on computers or (lately at an increasing pace) on smartphones and tablets. We resort

to it less or more intensively in order to read news, check and send emails, search for information, do transactions with authorities, interact socially with acquaintances, etc. Depending on the time spent on it, the activities carried out and the time of day when they are carried out it can boost or reduce the productivity of employees. As a major factor of social and economic life it is certainly worthy of statistical measurement and indeed it attracts a lot of attention.

Statistics based on automatically collected monitoring data cover a large share of relevant information about the use of Internet by individuals. They correspond to a substantial subset of the statistics produced by the Community Survey on ICT usage in Households and by Individuals. Types of information that cannot be covered are users' opinions and reasons for engaging on or avoiding certain activities. These would still need to be collected with the help of questionnaires.

The types of activities and products can be discerned at great detail and therefore rich classifications can emerge for statistical use. Moreover, the fact that data are recorded at great detail also allows the change of the classifications to fit changing statistical needs. Historical data can be converted easily to the new classifications.

The conclusion is that the statistics produced by the envisaged process are highly relevant Information Society indicators.

4.2.2. Accuracy

As in the case of the statistics about business web sites, the discussion will deal with both sampling and non-sampling errors. As the discussion will make clear, the major concern are non-sampling errors caused by a) ambiguity in what is 'usage' of an application or web site, and b) non-response, due to refusals to participate and the ability of users to switch the software off at will.

4.2.2.1. Sampling error

The pilot survey that was implemented in the context of the present project did not manage to obtain a random sample from a proper sampling frame. The only frame that was obtained was a web panel maintained by a market research and opinion polling company in Greece. The sample was not even random: the whole panel was informed about the pilot survey and those who wished to take part were included in the sample. Therefore there are no estimates of the standard error of the statistics.

On the other hand, the envisaged implementation of the process relies on sampling from the regular population sampling frames of the NSI. Its sampling error will therefore be comparable to those of the other national social surveys. Its magnitude will be the outcome of the sample design and the sample size chosen and of the prevalence of the different activities amongst users.

4.2.2.2. Non-sampling errors

There are several possible types of non-sampling error. Some arguments or brief definitions introducing the different types in section 4.1.2.2, concerning business web sites, hold here too but are not repeated in order to save space.

Coverage errors. In the envisage case of relying on the regular population sampling frames of the NSI and using a sampling procedure that is also used in other social surveys, the coverage errors will be comparable with those of the other surveys.

A potential source of undercoverage is to use software that does not run on all operating systems. NSIs must clearly strive to cover all operating systems with large market shares.

The pilot survey relied on a very small sample drawn from a very small web panel and therefore there is no point in even assessing coverage.

Measurement errors. The basic data recorded by the software are the times when the user starts and finishes using specific applications or visiting specific web sites. If we assume that the ‘name’ of the application or web site is recorded accurately, measurement errors can affect recorded times only. One possible error is to record starting times to the nearest minute before starting and end times to the nearest minutes after finishing, thus causing a positive bias. The size of this bias will depend on the number and duration of activities: the more and shorter they are, the greater bias will be as a share of true durations.

The definition of ‘usage’ of an application or web site can also cause biases. If a user stops typing in his blog for 10 minutes because he is thinking about what he is about to write is this measured as usage? If he has stopped typing because he is doing something else but the blog’s window is still open? The technicalities of the monitoring software and the definition of ‘usage’ are therefore crucial. It is not even simple to have an overall conclusion about whether bias will be positive or negative.

Bias will also be caused if more than one user share a device and if, moreover, there is no way of distinguishing who is using it on any time. This lack of separation of users will overestimate the shares of users carrying out each activity and will over- or underestimate the amounts and shares of time spent on them depending on each user’s pattern of use.

The wrong mapping of specific applications or web sites to categories of activity is also a measurement error, if mapping is automated by the monitoring software and if the recorded data mention only categories. It causes bias in the shares of users and amounts of time per category of activity.

On the positive side, the envisaged process does not suffer from any deficient recollection of activities, which reduces the accuracy of information collected with questionnaires. Furthermore, considerably richer information is recorded by the software than with questionnaires: individual applications and web sites, exact recorded starting and finishing times and separate recording of activities running in parallel. These are positive aspects unattainable by traditional surveys.

Processing errors. If the mapping between applications / web sites and categories of activities, mentioned before, is carried out in the NSI, errors that occur in are classified as processing errors. All other processing, e.g. conversion of starting and finishing times into durations, is automated and any errors will be discovered during testing.

Non-response errors. There is one source of unit non-response and two sources of item non-response in the process. Unit non-response mainly corresponds to refusal of individuals to participate in the sample and at present can be expected to be very extensive. Earlier studies have found that usable data are obtained from approximately 5% of the initially contacted samples of individuals⁴⁷. If those that accept to

⁴⁷ 5.8% of the chosen sample according to ‘Bouwman, H., Heerschap, N., de Reuver, M. (2012) Mobile handset study 2012. The Hague: Statistics Netherlands’ (p.10); 3.8% of the sample according to ‘European Commission (2012) Internet as a data source. Luxembourg: Publications Office of the European Union.’ (p. 148).

participate differ from those that do not, e.g. they are persons more accustomed to using the Internet, younger, etc. the bias can be serious. The NSI must make efforts to reverse the negative climate by explaining the nature of the collection and of the data and by possibly offering incentives.

The equivalent of item non-response is caused by two reasons. Firstly, the ability of users to switch the software off at will. This causes certain types of activity, whatever each user does not want to reveal, to be under-represented. The impact on average durations is harder to assess as it depends on the specific usage patterns of each person. The second possible cause of item non-response is the inability of the chosen software to record certain types of activity on certain operating systems. For example, the software used in the study by Bowman *et al* (2012), mentioned in the previous footnote, could not record URLs of sites visited with the Safari browser in iOS devices. This latter type of non-response is however avoidable by the selection of a different software.

4.2.3. Coherence and comparability

The envisaged survey is a social survey, which, with the exception of the measurement process and the definitions of the variables on which data are collected, operates with the same concepts and procedures as other social surveys.

4.2.3.1. Coherence

The survey samples individuals as the other social surveys of the ESS, classifies the target population by region, sex, age, level of education, employment status, etc., using the same nomenclatures as the other social surveys and draws its sample from the regular national sampling frame.

Its ‘novelty’ lies exclusively in the survey variables and mainly in the data collection mode. These however, are not affecting its coherence with other surveys. Coherence with other social surveys is very high; coherence with non-social surveys is at the same level as that of the other social surveys with them.

4.2.3.2. Comparability

The mapping of applications and web sites to different categories of activity makes possible differences between countries and therefore can reduce geographical comparability. This is the opinion of the project team, although it could not be tested in the pilot survey.

A large number of ‘standard’ applications and web sites are however universal. Moreover, local variants of them (e.g. the national equivalents of eBay) are usually well known to local experts and it is very clear which international ‘standard’ they resemble. As a consequence there is a lot of potential for a ‘centralised’ mapping in the ESS and its use by all NSIs. What will be left out will be national applications and sites, which will arguably have small user bases; otherwise they would have already been included in the popular local variants mentioned earlier. Therefore, problems may appear in geographical comparability but it seems that they will not be serious.

The same goes for comparability over time. Similarly to the current ICT survey, the set of target categories of activities will probably need to change often, to reflect changes in the Internet’s place and usage in society. However, if the data collected in previous rounds are available, they will be re-processed for re-classification, as long as they report individual applications and web sites. Historical series will then be re-constructed, resolving the comparability problem. On the other hand administrative or legal reasons, e.g. the requirement to delete the collected data after a given amount of time, will make

impossible the re-construction of time series. Judging from the legal feasibility assessment reported in chapter 6 of the present document this deletion of data may be imposed on NSIs in the end.

4.2.4. Clarity

The produced indicators are quite straightforward to understand, even for laymen. Some technical expertise is required for their definition and their subsequent expression in more simple terms. We do not foresee any major issues in this respect.

4.2.5. Timeliness

The speed of data collection, processing and production of statistics is un-matched by the current ICT survey. Data collection can last as long or little as desired while processing could be over in a matter of days. The slowest stage of the process is expected to be the communication with individuals in order to get their permission for data collection. Even with this stage included, the whole process could take less than three months. The timeliness of the statistics will therefore be very high.

4.2.6. Conclusions about the statistics on the use of Internet by individuals

The envisaged statistics are very relevant for the measurement of the information society, since they describe, in very rich detail, the interactions of society with the Internet. The use of software allows the timely recording of activities with details that cannot be matched by traditional methods. Moreover, the processing of the data is very quick and very timely statistics can be available in very few months after the end of the reference period. Relevance, degree of detail and timeliness are the great strengths of the approach.

Non-response on the other hand is the major drawback of the approach. Monitoring software resembles spyware, which is clandestinely installed on devices and which, rightly, users have learnt to fear. Moreover, the recorded data are personal and most users do not want to share them with third parties.

The expected extent of non-response is so large that it makes the approach look impractical. Pilot studies however are not surveys run by NSIs. The latter generally have institutional credentials and legal backing to engage in data collection and should be trusted to protect the data they collect. With suitable legal arrangements to accommodate digital personal data it can be expected that the reluctance of the public to participate in surveys following this approach will decrease gradually.

5. Cost-benefit balance

5.1. Web site-centric methods

The site search-based approach (or Search Engine API), besides its technical feasibility, addresses a series of reported drawbacks and makes efficient and fair use of online and financial resources.

Firstly, the major disadvantage of the use of automated agents to collect data from web sites is their inability to interpret and read texts on web pages the way human agents do. This is offset by their ability to cover large samples of web sites. Hence a combination of human and automated agents seems to be the best of both worlds. The site search-based approach is characterized by the facilitation of the aforementioned combination because it can extract useful information through the usage of keywords (in the pilot survey described in section 3 of deliverable D3 of the project) and regular expressions in the future.

Another issue with scraping tools is that while they are a lot faster than human agents they have high costs because they need to be reprogrammed for each target web site and for each change of a target web site.

The site search-based approach, as it will be argued in the next sections is both economic feasible and scalable for more than one indicators.

5.1.1. The site search market

For the last decade, Google is dominating the worldwide search engine market. Indicatively, in April 2010 Google controlled 86.3% of the global search market⁴⁸, while in July 2013 its market share dropped slightly in 84%⁴⁹. The main two markets in which Google is not leading are China and Russia. In China, Baidu holds a market share over 60%⁵⁰ and Yandex performs similarly in Russia.

According to economic theory, the Google ecosystem of services is based on the most profound indirect network effect, the so-called “multi-sided platforms”. As [Vafopoulos 2011] explains:

“A multi-sided platform provides services to two or more distinct groups of customers who not only need each other in some way, but also rely on the platform in order to intermediate transactions among them. Multi-sided platforms emerge when there is underlying value from getting multiple sides together but transactions costs are high (e.g. eBay decreased the exchange cost for buyers and sellers). In the Web, multi-sided platforms primarily perform three interrelated core functions. First, they serve as matchmakers to facilitate exchange among users. Second, they build communities because in that way users are more likely to find a suitable match (e.g. Facebook). Third, they provide shared resources and reduce the cost of providing services to multiple consumer segments⁵¹. This practice has resulted in an ecosystem that consists of interconnecting multi-sided platform businesses (e.g. Google’s advertising platform) with excessive market power. In economics, the theory of network externalities and effects has extensive applicability and importance in diverse issues like competition, anti-trust policy and regulation, business strategy, innovation and intellectual property.”

In this context, during 2010 the European Commission introduced an antitrust investigation into allegations that Google had abused a dominant position in online search to impose preferential placement of its own services in the advertising market⁵².

On the other hand, Google’s unwavering dominance over the years has a solid ground on better quality search results. In such a case, market shares could serve as an accurate proxy for quality in the online search industry. For that reason, Google has been selected from our consortium to provide the site search platform for the pilot survey described in section 3 of deliverable D3.

⁴⁸<http://marketshare.hitslink.com/report.aspx?qprid=5&qpcustom=Google%20-%20Global&qptimeframe=M&qpsp=120&qnp=25>

⁴⁹<http://www.academicads.ca/seo/search-engine-market-share-july-2013/>

⁵⁰<http://thenextweb.com/asia/2013/09/17/baidu-still-tops-chinas-search-market-with-63-share-as-merger-shakes-up-chasing-pack/>

⁵¹European Commission DG Communications Networks, Content & Technology Internet as data source. Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering

⁵²Antitrust: Commission probes allegations of antitrust violations by Google: 2010.

<http://europa.eu/rapid/pressReleasesAction.do?reference=IP/10/1624&format=HTML&aged=0&language=EN&guiLanguage=en>.

The Site search or the Search Engine API market segment is a small part of online search industry⁵³. Its core functionality focuses on providing search facilities inside a website (e.g. a search box that is powered by Google) and it is mainly used from web administrators and web analysts.

The main market players are the established search engines like Google, Yahoo! and Bing.

[Google Site Search](#) is part of the Google Enterprise Search service bundle. It offers the widest range of features such as multiple languages, top results biasing, XML feeds, synonyms, flexible indexing and much more.

Regarding pricing policies, the first 100 queries per day are free. For smaller sites, Google Site Search starts at \$100 for up to 20000 annual searches. For usage above one million searches, enterprise-level support and offline purchasing are available (refer to Table 6 for unit price comparisons).

The [Yahoo! BOSS Search API](#) is the open search and data services platform of Yahoo!, which offers, among other features, flexible search options in parts of the Web such as images, blogs etc. The pricing policies of the services are far lower from all the other competitors (except the free plans).

The [Bing Search API](#) is offered through the Azure marketplace and provides similar functionality to Yahoo! (e.g. users can request web, images, news, video and other source types for a single search query), but in higher prices per query.

The [Baidu P4P \(Paid Search\) API](#) only offers search marketing APIs and no organic search API.

The rising [DuckDuckGo](#)'s Instant Answer API offers free access to most of its services such as topic summaries, categories, disambiguation and definitions. The use of the service is restricted to non-commercial use (commercial use requires email approval from the company) and by providing attribution in each place the API is employed. According to our initial tests DuckDuckGo's Instant Answer API is only relevant for English websites.

[Faroo.com API](#) is promoted as the free alternative to established search engines since no registration is required for an upper limit of 1 million queries per month or more than 1 request per second. Custom pricing policies apply beyond these limits. On the other hand, there are no extensive and official reports for its quality standards.

Table 6. Comparative analysis of pricing policies in the site search market.

Provider	Free plan	Cost per search query (\$)		comments
		low volume plans	high volume plan	
Google	100 queries / day	0.005	0.004	Top quality results
Yahoo!	Does not exist	0.0008	0.0008	Low quality - flexible options
Bing	5000 queries	0.002	0.002	Average

⁵³To the best of our knowledge there is no yet a quantitative estimate of the total sales in this market segment.

Provider	Free plan	Cost per search query (\$)		comments
		low volume plans	high volume plan	
	/ month			quality
Baidu	no organic search API			
DuckDuckGo	All queries	Only relevant for English sites		
Faroo	Less than 1 million queries / months AND 1 query / second	-	Custom pricing	No quality reports

5.1.2. Costs

Software costs

Building and reprogramming crawlers for each website separately, is not only very costly for the interested parts (approximately 200 human hours are needed to develop, test and reprogram twice a web robot) but also costs to the society as whole, since bandwidth and processing power are scarce resources that cannot be stored for future use.

As noted by [Koster 1995] the use of Web robots induces the following indirect costs: (a) network resources, as robots require considerable bandwidth and operate with a high degree of parallelism during a long period of time, (b) server overload, especially if the frequency of accesses to a given server is too high (c) poorly written robots, which can crash servers or routers, or which download pages they cannot handle and (d) personal robots that, if deployed by too many users, can disrupt networks and Web servers.

On the other hand, site search actually, re-cycles the already fulfilled crawl and analysis done by the search engines. Thus, in the case of the proposed approach, site search comes with virtually zero social cost and avoids Web “pollution”.

In the same context, the financial cost, even for the scenario of 100 indicators for 30 countries and keywords per indicator would not cost more than \$450 in the case of Google (for a detailed analysis refer to Table 7). This solution is also very elastic and adjustable and can be repeated several times in different periods during a calendar year.

Table 7. Analysis of costs for various scenarios for the site search approach.

				Cost per search query	
				Low volume plans	High volume plan
			Google	\$0.0050	\$0.0040
			Yahoo!	\$0.0008	\$0.0008
			Bing	\$0.0020	\$0.0020
SCENARIOS					
Countries	1	1	10	20	30
Indicators	10	30	40	40	100
Keywords per indicator	10	20	20	20	30
Total queries	100	600	8000	16000	90000
COST					
Google	\$0.00	\$3.00	\$40.00	\$80.00	\$450.00
Yahoo!	\$0.08	\$0.48	\$6.40	\$12.80	\$72.00
Bing	\$0.00	\$1.20	\$16.00	\$32.00	\$180.00

These costs can be contrasted with the cost of a government web site scanner software reported in a recent study for the European Commission⁵⁴:

- 210000 euros initial cost for software development and licences
- 32500 euros cost per country for national adaptations.

Sampling costs

The sampling of enterprises will be no different than that in other business surveys. The sample will be drawn from the national business register, or other frame used by the NSI, and suitable contact persons will be contacted in order to be informed about the survey and give their consent.

It can be expected that more effort will be required to obtain consent than is usually the case in business surveys. There is no information though on how much greater this effort will be: the expected reluctance of site owners to allow a crawler search their site might not be great if the request is posed by the NSI with the possible backing of legal obligation for enterprises to provide the data. It seems reasonable that in such a case not much extra effort will be needed.

⁵⁴ European Commission (2012) Internet as a data source. Luxembourg: Publications Office of the European Union.

In this respect there are no financial gains or losses in sampling, compared to a possible implementation of the survey by questionnaire.

Data collection costs

One component of collection costs is the preparation of the mapping between web site characteristics and keywords or technologies. This is a kind of “art”, as stated earlier, and the effort required depends on the number of characteristics, their complexity and the expertise of the staff that will undertake it. Some piloting may also be needed in order to test and improve the mapping. Judging from the pilot survey implemented in this project, two person-months are sufficient for this exercise.

The actual collection of data has very little cost as shown above. This cost however is additional to that of the Community Survey on ICT usage and E-commerce in Enterprises because the automatic collection covers a very small subset of the information collected by that survey.

Therefore the new method offers no collection cost savings; its costs can only be juxtaposed with the expected benefits in terms of statistical information produced.

Processing costs

The processing of the data is automated and quite straightforward since it involves, at least for the currently envisaged indicators, the tabulation of simple proportions.

There is only one exception: the validation of the mapping between characteristics of web sites and keywords or technologies. Validation can only be achieved by human operators re-visiting a subset of the sample and assessing themselves whether the target characteristics are present. Depending on the size of a site, on the number of target characteristics and on their “complexity”, the checking of a single site can take a few hours. Assuming that 5% of the sample will be checked like this and that each site takes 1/3 of a person day, the amount of person-days needed for validation will be $0.05 \cdot n / 3$, where n is the size of the sample. In 2012 the total achieved sample size in the EU28 member states was 164655 enterprises⁵⁵. Assuming that 80% of them have a web site a rough estimate of validation effort is 2195 person-days or a little more than 100 person-months over the EU.

5.1.3. Benefits and conclusion

As discussed also in section 4.1 the main benefits of the approach are that it produces very relevant indicators in a very timely way. No monetary value can be put however on them.

Moreover, the benefits are offset by the insufficient accuracy of the produced statistics.

To our opinion the costs (especially validation effort) are too high for the obtained benefits. Unfortunately, lacking more detailed cost information, no more precise assessment can be made.

5.1.4. To the future

As of October 2013 at least 4.45 billion webpages (excluding the Deep Web)⁵⁶ have been uploaded, 2.4 billion people are online worldwide⁵⁷, including a billion active Facebook users⁵⁸, 400 million

⁵⁵ Eurostat (2013) *Methodological manual for statistics on the Information society*, v. 3. Luxembourg: Eurostat.

⁵⁶ <http://www.worldwidewebsite.com>

tweets were generated per day as of March 2013⁵⁹, 71 million Wordpress sites were available as of October 2013⁶⁰ and 52 billion Resource Description Framework (RDF) triples were published and linked in OpenLink Software's Linked Open Data Cloud Cache as of March 2012⁶¹.

This inexhaustible flood of data remains untapped from the majority of companies, public agencies and citizens and it is actually monopolized from a small number of gigantic multinational private and government entities.

On the other hand, European and international standards on critical aspects of everyday life (e.g. food safety, public health, competition) are considered to be a catalyst for economic development and wellbeing. For instance, any packaged food is required to clearly state its ingredients on the package. Lately, in many countries, a new regulation has been imposed for some categories of foods to include a [nutrition facts label](#) in order to increase further food quality and safety.

In the case of the Web, W3C is continuously working in this direction of developing technical quality standards (e.g. <http://validator.w3.org/>). But these standards are not mandatory and are not designed to provide accurate and comprehensive reports of their major characteristics and functionalities. Practically, they cannot work as reporting mechanisms that will allow NSIs and other agencies with public interest to produce and publish information about online activity, in reasonable cost, which can be used for social and economic policy-making. This remains, mainly, an exclusive privilege to the Web giants with the difference that they share a tiny part of the accumulated information and resulted knowledge.

More generally, the rapid popularity and penetration of the Web, has raised the issue of information accountability in the sense that information usage should be transparent so it is possible to determine whether a use is appropriate under a given set of rules [Weitzner et al 2008]].

Information accountability crosses a wide range of social and economic issues, such as privacy, security, freedom, self-determination and so on. But it is also relevant to the collection and analysis of online activity by NSIs since it reflects the lack of any social obligation for companies from sharing and occupying part of the online public space.

Our proposal is focused on developing procedures and employing technologies that will enable the direct and automatic transfer of information from companies that are based or have Web presence to NSIs.

This set of information could be both relevant to existing questionnaires (e.g. URL, web commerce options etc.) and newly introduced such as open data exploitation and policy. The first application of the aforementioned proposal could be developed for e-government websites and services. In such case, the reported variables may reflect the range, the quality and actual use and user's feedback of their services.

The site search approach that has been followed in a part of this pilot project could be considered as an intermediate step towards building auto-reporting processes for Official Statistics. In general, Web

⁵⁷<http://www.internetworldstats.com/stats.htm>

⁵⁸<http://www.bbc.co.uk/news/technology-19816709>

⁵⁹<http://www.youtube.com/watch?feature=playerembedded&v=Bl-FpuehWGA>

⁶⁰<http://en.wordpress.com/stats>

⁶¹<http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

scrapping or other similar techniques to acquire online information are suboptimal solutions because they are built after and apart from the original code of a web page. This detachment between the original code and the mechanism about reporting it causes many inefficiencies and inaccuracies and demands additional effort and financial resources.

In our point of view, a long-term, sustainable solution will include the legal obligation of any company and service residing in the Web to provide a minimum amount of *embedded* information in the form of Open Data (e.g. machine-readable and freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control) both to every user and to public authorities for statistical purposes. This data provision (let us call it “online information accountability by design”) will be based on a flexible architecture that will take account of a series of key parameters such as the industry, the user base, the influence etc. of the online service. Thus, for instance, a giant social networking site should be obliged to provide more thorough data reports. Data reporting could be divided to one part open to all users and a second more detailed part, which may include sensitive business information, and will be directed to NSIs. NSIs will only publish aggregated information related to this sensitive information in order to safeguard the motivation for new investments in the Web.

The technological background for such solutions is here, including the ontological schemes for statistical modelling (e.g. SDMX, data cube ontology) and Web 3.0 technologies for data processing and provision in global scale. As [Weitzner et al 2008] argue “Drawing on semantic Web techniques, larger and larger overlapping communities on the Web can develop shared policy vocabularies in a bottom-up fashion. A lack of perfect global interoperability of these policies is not a fatal flaw. Just as human societies learn to cope with overlapping and sometimes contradictory rules, so too are policy-aware systems likely to develop at least partial interoperability.” In this context, during the last years Web 3.0 technologies (similar terms are Semantic Web, Web of Data, Linked Data) have prevailed as an effective way to manage content that includes complex concepts and meanings and which requires inference using multiple channels, including the Web. The main advantages of Web 3.0 are summarized in scalability, interoperability, lower costs, useful searches and the creation of rich context in human interaction (e.g. social networking).

Apart from a potential paradigm shift in Official statistics, an “online information accountability by design” initiative will act as a catalyst in realizing the vision of an Open Web characterized by the self-determination of users, transparent public data flows and fair barriers to lock-in practices.

5.2. User-centric methods

A user centric application akin to crowdsourcing will typically involve setting up a sample of individuals randomly selected from a population of interest who will install an application in all devices they use to access the Internet. The application will be working in the background consuming minimal resources while collecting specific data for device and Internet usage which it consequently sends to the NSI for processing.

This is a process that is very different from the typical workflow for data collection from either statistical surveys or administrative sources that NSIs are using. Most of the effort is required to recruit the sample and also to develop or acquire and maintain the application software. The rest of the process (data entry, editing etc) can be more or less automated as many steps are or will be fully automated.

If an NSI opts for setting the survey up as a panel survey, with only gradual and partial refreshment of the sample of individuals, the process will have quite different cost structure with high initial costs and minimal repetitive ones compared to other production processes. In the presentation we refer to both options, namely a) a renewed sample in each round of the survey and b) a panel.

5.2.1. Costs

The main costs that need to be assessed include software development or acquisition and maintenance as well as sample recruitment (and retention in the case of a panel).

Software costs

An application needs to be installed in all devices that a user may use to access the Internet. This will include computers, tablets and smartphones. The application(s) should be also able to cover different operating systems (Windows, OSX, iOS, Android, etc). The coverage of all devices with Internet access will become more complicated in the future as the gamut of types used increases, e.g. Smart TVs.

Monitoring applications that are intentionally installed by users (as opposed to spyware) are already used for various purposes (parental control, employee monitoring, etc.). They are sometimes called benevolent spyware because they capture, store and share private information but, unlike spyware, they do that for a legitimate purpose and the device owner is aware of their presence.

The development of a monitoring application that captures and transmits information need not start from scratch, as possible components are already available, some of them for free and also as open source software. For example *kidlogger* is distributed with its source code and is able to monitor time of active use as well as applications used, web history etc. *Web filter / parental control* is another, although less developed, open source project for the same purpose. Monitoring applications are also available for traffic measurement and analysis.

However, the development / acquisition of software to cover different platforms, the development / acquisition of server-side software for receiving and processing data and the integration of components can amount to substantial software development effort and cost. Therefore, depending on the scope of the survey software development may require substantial financial resources and time before it is implemented even in pilot setups. On the other hand, this expenditure will be capital expenditure.

Some indications of software costs, culled from literature, are as follows:

- The aforementioned study for the European Commission⁶² estimated (roughly) that operating system monitoring software costs between 25000 and 125000 euros per year.
- The same study estimated that software monitoring activities through the users' browser would have 400000 euros setup costs and 100000 euros per year monitoring costs. *It must be noted that the software envisaged in this project functions as a combination of the two types of monitoring software studied in the Commission study.*

⁶² European Commission (2012) Internet as a data source. Luxembourg: Publications Office of the European Union.

Sampling costs

Sampling need not cost more than what it does in a regular household survey. Installing an application developed or purchased by the NSI in a selected respondent's device should have marginal costs. Selected persons receive instructions and support to download and install applications in an automatic way. However, a personal session, face-to-face or over the telephone, is required in order to record background characteristics of the respondent (demographic information, income, computer skills etc). It will also be useful for ensuring compliance, i.e. installation on all devices used by respondent. So the cost of a brief personal interview for each sample unit at recruitment needs to be budgeted. After that however no more effort is required apart from maintenance and relatively infrequent events (software malfunction, new devices etc.).

If the sample of individuals is retained as a panel, with gradual refreshment in a rotational scheme, the sampling costs will be substantially smaller. Assuming that

- the survey is quarterly,
- a nominal sample size of n persons is required per quarter,
- one quarter of the panel is renewed each quarter, and
- there is no panel attrition

the annual sample of a cross-sectional survey will be $4 \cdot n$, whereas that of a panel survey will be $1.75 \cdot n$ in the first year and n in every subsequent year. In the long run therefore sampling costs of the panel survey will tend to be 25% of those of a cross-sectional survey.

Financial incentives

The use of incentives is an old concern in Official Statistics. In general the view that was adopted in the 90's was that surveys for official statistics do not need incentives to boost response rates and should not use them except when respondents face actual costs by participating in the survey or the survey is too intrusive⁶³.

Installing software and being continuously monitored is quite intrusive but it may not be necessarily perceived as such. So it is not clear if monetary incentives should be used or not. Limited experience in Greece showed that an incentive of between 30 and 50 euros per person would be needed. Incentives however can be offered in kind (e.g. purchase coupons for an electronics mega-store) and they will then cost to the NSI less than their nominal value since they will be bought in bulk. We believe that 20 euros is a reasonable unit cost for a 30-euro coupon. In 2011 the achieved sample size in the EU28 member states was 185141 individuals⁶⁴. If an incentive of 30 euros was to be given to each one of them, the total cost at EU level would be 3.7 million euros.

⁶³ COPAFS (1993) Providing Incentives to Survey Respondents, Final Reports, Submitted to the Regulatory Information Service Center General Services Administration Contract Number GS0092AEM0914 by the Council of Professional Associations on Federal Statistics September 22, 1993 available at http://www.copafs.org/reports/providing_incentives_to_survey_respondents.aspx

⁶⁴ Eurostat (2013) *Methodological manual for statistics on the Information society*, v. 3. Luxembourg: Eurostat.

The pilot investigations however did not have the “backing” of an official statistical authority and therefore it cannot be ascertained whether an incentive would still be needed.

Data collection and processing costs

Collection and processing costs are expected to be very little, due to the automation of the process. There will still be some interviewing effort for collecting the background and possibly opinion and perception information but the length of the questionnaire will be a fraction of what would be needed in order to collect all the information by an interview.

If the software used for monitoring is priced per person, as is the case with parental-control tools, the data control costs include this cost too. *Qustodio*, which was used in the pilot survey carried out in this project, costs 285 euros per year for 50 individuals and 50 devices. Assuming that each user has two devices on average and that the licences (priced annually) can be transferred to different sample members in each quarter, the 2011 sample size of the ICT survey would cost roughly 530000 euros per year.

5.2.2. Benefits and conclusion

The benefits of the approach have also been summarised in section 4.2.

It can collect automatically a large part of the information currently collected with questionnaires in the regular ICT survey and therefore it reduces response burden considerably. It can also collect information, which could not be easily collected with a questionnaire, e.g. the volumes of data received or transmitted by the individuals.

Furthermore, the data are recorded with great precision because they do not depend on the individuals' recall and reporting of activities but are recorded digitally. This also enables their recording in very rich detail that cannot be matched by traditional methods: individual applications and web sites, exact recorded starting and finishing times and separate recording of activities running in parallel.

The time required for data collection is also reduced considerably due to the automation and the reduced need for interviewers. Statistics can be available a lot faster than with traditional methods.

However, it is not easy to put a monetary value on these benefits so as to juxtapose it with the costs.

We have given some indications or very rough approximations of the cost of the automated data collection.

The only indication of the cost of the current ICT survey comes from the grants that Eurostat gave to national authorities. Anonymized data provided to the project team report the total data collection cost, for both the households and the enterprise surveys, over the EU in 2012 at almost 3900000 euros.

The new method can present considerable savings in data collection, if a solution that is not priced by user is adopted. On the other hand it has considerable setup costs and possibly costs for the provision of incentives. Based on the limited available data it seems that the new method is overall most costly than the current survey.

6. Legal feasibility

6.1. Introduction

The aim of this assessment is to examine whether the automatic data collection methods examined by the project are feasible from the legal point of view.

The issue of collecting and aggregating statistical data has legal implications that relate both to Data Protection and Privacy regulations, and to areas of Intellectual Property Rights and particularly the sui generis Database right in the EU context.

We start by analyzing whether the methods of statistical data collecting and aggregating proposed are compatible with the existing legal framework (section 6.2). The analysis has been based on the exploration of the EU data protection legal framework concerning the processing of statistical data (section 6.3) and of the provisions concerning the sui generis Database right in the EU context (section 6.4).

6.2. Legal compatibility analysis

The object of the present legal analysis is a set of methods under which the Internet shall be used as a data source suitable for statistical purposes and relevant research. More precisely, the examination of the legal feasibility concerns a project that involves:

- (a) the installation of a software mechanism in several types of personal computing devices (i.e. desktop computers, tablets, smartphones etc.) with the aim of collecting information on the user's online activities on the Internet, such as duration of Internet usage, hours per day, days per week of Internet usage, visits on web pages etc.
- (b) use of a “crawler”-type software to collect and analyse content of corporate web sites, such as the kind of facilities and several categories of information, such as open vacancies for employment, that the site provides to end users.

Overall conclusion: In both cases the user and the private entity (corporation, enterprise etc.) must give their explicit consent for the data collection and processing. If this is received and moreover the sample members have been informed about the data that will be collected and the uses to which they will be subjected, the electronic collection does not differ, from the legal point of view, from the collection of similar data with questionnaires.

The legal assessment will focus on the stages of: data creation, data aggregation or collection stage, enrichment stage and dissemination stage. In each of the stages the aim is to identify the degree to which:

- property rights are created and how their transfer is effected
- if personal data are involved, who conducts their processing, for how long and how they are to be used.

6.2.1. Data protection terms and conditions

The data protection legal framework recognizes the consent of the data subject generally as an appropriate legal basis for the collection and processing of personal data. Nevertheless, there are two

crucial factors that should be taken into account in order to ensure that the data subject's consent is an adequate condition for all four stages of the methodology in hand. The first factor refers to the circumstances the data subject opted in and the content of his/her consent. The second factor refers to the cases of data collection and processing that even the proper consent forms only one part of the overall procedure for the lawfulness of the project.

1. The adequate consent

According to the European data protection legal framework, the data subject's consent is defined as “*any freely given specific and informed indication of his wishes by which the data subject signifies his agreement to personal data relating to him being processed*”⁶⁵. The relevant provisions on the lawfulness of data collection and processing are referring to the existence of the “*unambiguous*” consent. For consent to be unambiguous, the procedure to seek and to give consent must leave *no doubt* as to the data subject's intention to deliver consent. In other words, the indication by which the data subject signifies his agreement must leave no room for ambiguity regarding his/her intent. If there is a reasonable doubt about the individual's intention, there is ambiguity.

There are in principle no limits as to the form consent can take. However, for consent to be valid, in accordance with the Directive, it should be an indication. Even if it can be “any” form of indication, it should be clear what exactly can fall within the definition of an indication. The form of the indication (i.e. the way in which the wish is signified) is not defined in the EU Data Protection Framework. For flexibility reasons, “written” consent has been kept out of the final text. It should be stressed that the Directive includes “any” indication of a wish. This opens the possibility of a wide understanding of the scope of such an indication. The minimum expression of an indication could be any kind of signal, sufficiently clear to be capable of indicating a data subject's wishes, and to be understandable by the data controller. The words “indication” and “signifying” point in the direction of an action indeed being needed (as opposed to a situation where consent could be inferred from a lack of action)⁶⁶.

More specifically, in the field of personal data collection and processing for statistical purposes the data subject's “informed” consent requires⁶⁷ that the persons questioned shall be informed of the following elements:

- (a) the compulsory or optional nature of the response and the legal basis, if any, of the collection,
- (b) the purpose or purposes of the collection and processing
- (c) the name and position of the person or body in charge of the collection and/or processing,
- (d) the fact that the data will be kept confidential and used exclusively for statistical purposes,
- (e) the possibility of obtaining further information on request.

At their request and/or according to the ways and means defined by domestic law, data subjects shall also be informed of the following:

⁶⁵ Article 2 (h) of the Data Protection Directive. Article 2 (g) of the Data Protection Framework Decision in the Framework of the Police and Judicial Cooperation in Criminal Matters. Article 2 (f) of the e-Privacy Directive. Article 2 (h) of the Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data Official Journal L 008 , 12/01/2001 P. 0001 - 0022

⁶⁶ Article 29 Working Party Opinion 15/2011 on the definition of consent, p. 11.

http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf

⁶⁷ According to Chapter 5 of the Appendix to Council of Europe's Recommendation No. R (97) 18 concerning the protection of personal data collected and processed for statistical purposes

- (f) the way in which consent can be refused or withdrawn, in the case of optional surveys and, in the case of compulsory surveys, the possible sanctions this would entail;
- (g) where applicable, the conditions of the exercise of the rights of access and rectification,
- (h) the categories of persons or bodies to whom the personal data may be communicated;
- (i) the guarantees to ensure the confidentiality and the protection of personal data;
- (j) the categories of data collected and processed.

When the data subjects are not directly questioned, they shall be informed of the existence of the collection unless this is manifestly unreasonable or impracticable. They shall be able to inform themselves appropriately of the elements listed above. The persons questioned shall be informed at the latest at the time of collection. Under the title “Secondary collection”, the Chapter reads that cases of processing or communication for statistical purposes of personal data collected for non-statistical purposes shall receive suitable publicity. The data subjects shall be able to obtain in a suitable way all abovementioned information, unless:

- (a) this is impossible or involves a disproportionate effort; or unless
- (b) the processing or communication of the data for statistical purposes is expressly provided for under domestic law.

The data subject shall be able to withdraw his or her consent for a single survey, as long as, identification data have not been separated from other data collected, or to suspend at any time and without retroactive effect his or her co-operation in a survey which extends over a period of time. Refusal to reply shall not be penalized unless domestic law provides for sanctions⁶⁸.

Personal data processed for a given statistical purpose may be communicated for other statistical purposes as long as these are specified and of limited duration. Communication in accordance with this principle shall be the subject of a written document setting out the rights and obligation of the parties, unless safeguards are provided for by domestic law. The controller shall in particular:

- (a) stipulate that the third party may communicate these data only with the express agreement of the said controller;
- (b) stipulate that the third party take appropriate security measures and
- (c) ensure that any publication of statistical results obtained by this party will anonymize the data unless dissemination or publication manifestly presents no risk of infringing privacy rights.

Sensitive data communication is allowed where provided for by the law, or where the data subjects have given their explicit consent and provided domestic law does not prohibit the giving of the consent.

Consent can only be valid if the data subject is able to exercise a real choice, and there is no risk of deception, intimidation, coercion or significant negative consequences if he/she does not consent. If the consequences of consenting undermine individuals' freedom of choice, consent would not be free. An example of the above is provided by the case where the data subject is under the influence of the data controller, such as an employment relationship. In this example, although not necessarily always, the data subject can be in a situation of dependence on the data controller - due to the nature of the relationship or to special circumstances - and might fear that he could be treated differently if he does not consent to the data processing.

To be valid, consent must be specific. In other words, blanket consent *without* specifying the exact purpose of the processing is not acceptable. To be specific, consent must be intelligible: it should refer clearly and precisely to the scope and the consequences of the data processing. It cannot apply to an open-

68 According to Chapter 6 of the Appendix to Recommendation No. R (97) 18

ended set of processing activities. This means in other words that the context in which consent applies is limited⁶⁹.

Consent must be given in relation to the different aspects of the processing, clearly identified. It includes notably which data are processed and for which purposes. This understanding should be based on the reasonable expectations of the parties. “Specific consent” is therefore intrinsically linked to the fact that consent must be informed. There is a requirement of granularity of the consent with regard to the different elements that constitute the data processing: it cannot be held to cover “all the legitimate purposes” followed by the data controller. Consent should refer to the processing that is reasonable and necessary in relation to the purpose. It should be sufficient in principle for data controllers to obtain consent only once for different operations if they fall within the reasonable expectations of the data subject.

According to a preliminary ruling regarding Article 12(2) of the e-Privacy Directive⁷⁰, concerning the need for renewed consent of subscribers who had already consented to have their personal data published in one directory, to have their personal data transferred to be published by other directory services the EU Court of Justice held that where the subscriber has been correctly informed of the possibility that his personal data may be passed to a third-party undertaking and s/he has already consented to the publication of those data in such a directory, renewed consent is not needed for the transfer of those same data, *if it is guaranteed that the data in question will not be used for purposes other than those for which the data were collected with a view to their first publication (paragraph 65)*.

2. Where the consent is not enough

The Data Protection Directive foresees in Article 8.2(a) that in some cases, to be determined by Member States, the prohibition of the processing of special categories of personal data may not be lifted by the consent of the data subject. This is the case when the operation contains “special categories” of personal data (revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life.).

Collecting data of an end user's visits on the Internet may also contain collection and processing of these sensitive data categories. In several Member States, the appropriate safeguards that allow the collection and processing of sensitive data are formulated as a prior permission issued by the independent Data Protection Authority⁷¹. Article 8 of the Data Protection Directive obliges the data controller to comply with national law procedures, in the case of sensitive data collecting.

In conclusion, the mere consent will not be the appropriate legal ground for collecting sensitive data. The controller must make sure that all national law procedures applicable to any territory exposed to the project are followed. It must be examined carefully whether the recording of sensitive data abides to the specific national laws or whether the data that will be recorded must be tweaked appropriately.

b. Database right dimension

The copyright issues relating to the methodologies of the project are less complex, since there will be a consent for collecting data from corporate webpages. The webpages may form a “database” of the owner company. In the case of software that pulls data from the webpage, the mere permission of the company

⁶⁹ Article 29 Working Party Opinion 15/2011 on the definition of consent, p. 17.

⁷⁰ Judgment of the Court of 5 May 2011, Deutsche Telekom AG (Case C-543/09). This case started with the referral made by the German Federal Administrative Court regarding telecom directories and in particular the interpretation of Article 25(2) of the Universal Service Directive (2002/22/EC) and Article 12(2) of the e-Privacy Directive (2002/58/EC). It is clearly linked to the special role of directories in the Universal Service Directive.

⁷¹ This is the case according to the Greek Law Nr. 2472/1997.

will legalize the whole operation. It should be mentioned in the relevant contracts the categories of data that will form part of the operation and the confirmation that the company owns all copyright data of its webpage. In the case of intellectual property rights reservations to third parties (i.e. webpage developers etc.), their consent should be also demanded.

6.2.2. Course of action for NSIs

The NSIs envisaging the application of IaD methods must therefore make sure that all steps of the production processes are compatible with the relevant national and EU legal framework. The following steps must be taken:

1. The legal service of the NSI carries out a thorough review of national and European legislation concerning the collection, storage and processing of personal and enterprise data for statistical purposes.
2. The production units of the NSI that will utilise the IaD methods prepare detailed descriptions of the “business cases”. They contain a description of the data sources, of the means that will be used for data collection, of the data that will be collected, of the statistical purposes that will be served, of the processing they will be subjected too, of possible re-uses in the future (always for statistical purposes), e.g. re-coding for reconstruction of historical data series of new indicators or with new codelists, of the means taken to ensure and protect the anonymity of the statistical units (persons or enterprises).
3. The descriptions are scrutinised by the legal service and revisions are proposed.
4. The descriptions are finalised and are submitted to the national bodies responsible for data protection issues.
5. Taking these bodies’ comments into account revised descriptions are produced and the production units examine whether the resulting production processes are still satisfactory from the statistical point of view.

6.3. Data protection legal framework

a. The Data Protection Directive

The main piece of personal data protection legislation at EU level is Directive 95/46/EC⁷². According to article 3 para. 1, the Directive shall apply to the processing of personal data wholly or partly by automatic means, and to the processing otherwise than by automatic means of personal data which form part of a filing system or are intended to form part of a filing system. Furthermore, according to Article 3, the Directive shall not apply to the processing of personal data:

- *in the course of an activity which falls outside the scope of Community law, such as those provided for by Titles V and VI of the Treaty on European Union and in any case to processing*

⁷² Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal L 281, 23/11/1995 P. 0031 - 0050

operations concerning public security, defense, State security (including the economic well-being of the State when the processing operation relates to State security matters) and the activities of the State in areas of criminal law,

- *by a natural person in the course of a purely personal or household activity.*

Article 2 of the Directive contains a list of definitions regarding the concept of the terms used at its provisions. The most important definition clarifies the mere notion of personal data. According to Article 2 (a),

“personal data’ shall mean any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity”.

The concept of personal data has been extensively analyzed by the Working Party composed by the representatives of the European data protection authorities, the European Commission and the European Data Protection Supervisor that was established by Article 29 of the Directive (“The Article 29 Working Party”). According to the Working Party, there are four essential elements that should be examined in order to clarify whether the information in hand is “personal data”: i) “...any information...”, ii) “...relating to...”, iii) “... identified or identifiable...”, iv) “...natural person...”⁷³. In the course of the analysis of the third element, the Working Party concluded that in general terms, a natural person can be considered as “identified” when, within a group of persons, he or she is “distinguished” from all other members of the group. Accordingly, the natural person is “identifiable” when, although the person has not been identified yet, it is possible to do it (that is the meaning of the suffix “-able”). This second alternative is therefore in practice the threshold condition determining whether information is within the scope of the third element. Identification is normally achieved through particular pieces of information which we may call “identifiers” and which hold a particularly privileged and close relationship with the particular individual. Examples are outward signs of the appearance of this person, like height, hair colour, clothing, etc... or a quality of the person which cannot be immediately perceived, like a profession, a function, a name etc. The Directive mentions those “identifiers” in the definition of “personal data” in Article 2 when it states that a natural person *“can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity”*.

In the same Opinion, the Working Party gave an example on the gray areas between personal data and statistical data:

73 Opinion 4/2013 on the concept of personal data, Article 29 Data Protection Working Party, WP 136, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

“Apart from their general obligation to respect data protection rules, in order to ensure anonymity of the statistical surveys, statisticians are subjected to a specific duty of professional secrecy, and under those rules it is forbidden for them to publish non anonymous data. This obliges them to publish aggregated statistical data which cannot possibly be attributed to an identified person behind the statistics. This rule is particularly relevant concerning the publication of census data. In each situation a threshold should be determined under which it is deemed possible to identify the persons concerned. If a criterion appears to lead to identification in a given category of persons, however large (i.e. only one doctor operates in a town of 6000 inhabitants), this “discriminating” criterion should be dropped altogether or other criteria be added to “dilute” the results on a given person so as to allow for statistical secrecy.”

Turning back to the Directive, there are specific provisions that relate to the processing of personal data for statistical purposes. Article 6 contains principles relating to “data quality”. According to these legally binding principles, Member States shall provide that personal data must be, inter alia, collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, *statistical* or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards. Furthermore, according to the same Article, personal data must be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, *statistical* or scientific use.

According to the Article 29 Working Party interpretation of these provisions⁷⁴, they “*should not be read as providing an overall exception from the requirement of compatibility, and it is not intended as a general authorisation to further process data in all cases for historical, statistical or scientific purposes. Just like in any other case of further use, all relevant circumstances and factors must be taken into account when deciding what safeguards, if any, can be considered appropriate and sufficient. In addition, as in other situations, a separate test must be carried out to ensure that the processing has a legal basis in one of the grounds listed in Article 7 and complies with other relevant requirements of the Directive*”. The Article 29 Working Party concludes that there may be three different scenarios for further analysis:

- Scenario 1: unidentifiable personal data: data are anonymised or aggregated in such a way that there is no remaining possibility to (reasonably) identify the data subjects. Full anonymisation (including a high level of aggregation) is the most definitive solution. It implies that there is no more processing of personal data and that the Directive is no longer applicable.
- Scenario 2: indirectly identifiable personal data: partial anonymisation or partial de-identification may be the appropriate solution in some situations when complete anonymisation is not practically feasible. In these cases, various techniques (including pseudo-anonymisation, key-coding, keyed-hashing, using rotating salts, removal of direct identifiers and outliers, replacing unique IDs, introduction of 'noise', and others) should be used to reduce the risk that data subjects can be re-identified, and subsequently, that any measures or decisions can be taken in their regard. In addition, there will also often be a need to complement these techniques with other safeguards in order to adequately protect the data subjects. These include data minimisation, as well as appropriate organisational and technical measures, including effective 'data silo-ing', to ensure functional separation.

⁷⁴ Opinion 3/2013 on purpose limitation, adopted on 2 April 2013, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

- Scenario 3: situations where directly identifiable personal data are needed due to the nature of the research. Directly identifiable personal data may be processed only if anonymisation or partial anonymisation is not possible without frustrating the purpose of the processing, and further provided that other appropriate and effective safeguards are in place. Among the appropriate safeguards which may bring additional protection to the data subjects, the following could be considered:
 - taking specific additional security measures (such as encryption);
 - in case of pseudonymisation, making sure that data enabling the linking of information to a data subject (the keys) are themselves also coded or encrypted and stored separately;
 - entering into a trusted third party (TTP) arrangement in situations where a number of organisations each want to anonymise the personal data they hold for use in a collaborative project;
 - restricting access to personal data only on a need-to-know basis, carefully balancing the benefits of wider dissemination against the risks of inadvertent disclosure of personal data to unauthorized persons. This may include, for example, allowing read-only access on controlled premises. Alternatively, arrangements could be made for limited disclosure in a secure local environment to properly constituted closed communities. Legally enforceable confidentiality obligations placed on the recipients of the data, including prohibiting publication of identifiable information, are also important. It is important to note that in high-risk situations, where the inadvertent disclosure of personal data would have serious or harmful consequences for individuals, even this type of access or restriction may not be suitable.

In addition,

- further processing of personal data concerning health, data about children, other vulnerable individuals, or other highly sensitive information should, in principle, be permitted only with the consent of the data subject;
- any exceptions to this requirement for consent should be specified in law, with appropriate safeguards, including technical and organisational measures to prevent undue impact on the data subjects (in case of doubt, the processing should be subject to prior authorisation of the competent data protection authority); exceptions should only apply with regard to research that serves an important public interest, and only if that research cannot possibly be carried out otherwise.

In Article 7 the Directive sets out the criteria for making data processing legitimate. There are six different legal grounds that permit the processing of personal data:

- (a) the data subject has unambiguously given his consent; or
- (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or
- (c) processing is necessary for compliance with a legal obligation to which the controller is subject; or
- (d) processing is necessary in order to protect the vital interests of the data subject; or
- (e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed; or

- (f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject.

In the case of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life, there is a specific regime for the lawful processing. According to Article 8 of the Directive, processing of such special categories of data shall be prohibited by the Member States, with five concrete exemptions:

- (a) the data subject has given his explicit consent to the processing of those data, except where the laws of the Member State provide that the prohibition may not be lifted by the data subject's giving his consent; or
- (b) processing is necessary for the purposes of carrying out the obligations and specific rights of the controller in the field of employment law in so far as it is authorized by national law providing for adequate safeguards; or
- (c) processing is necessary to protect the vital interests of the data subject or of another person where the data subject is physically or legally incapable of giving his consent; or
- (d) processing is carried out in the course of its legitimate activities with appropriate guarantees by a foundation, association or any other non-profit-seeking body with a political, philosophical, religious or trade-union aim and on condition that the processing relates solely to the members of the body or to persons who have regular contact with it in connection with its purposes and that the data are not disclosed to a third party without the consent of the data subjects; or
- (e) the processing relates to data which are manifestly made public by the data subject or is necessary for the establishment, exercise or defence of legal claims.

Directive 95/46 provides for specific obligations to data controllers. One of the general transparency obligations is to provide information to the data subject, when the data have not been obtained from him or her. According to Article 11, when the data have not been obtained from the data subject, Member States shall provide that the controller or his representative must at the time of undertaking the recording of personal data or if a disclosure to a third party is envisaged, no later than the time when the data are first disclosed, provide the data subject with at least the following information, except where he already has it:

- (a) the identity of the controller and of his representative, if any;
- (b) the purposes of the processing;
- (c) any further information such as
 - the categories of data concerned,
 - the recipients or categories of recipients,
 - the existence of the right of access to and the right to rectify the data concerning the data subject

in so far as such further information is necessary, having regard to the specific circumstances in which the data are processed, to guarantee fair processing in respect of the data subject.

According to Article 11 para. 2, the abovementioned obligation shall not apply where, in particular for processing for *statistical purposes* or for the purposes of historical or scientific research, the provision of such information proves impossible or would involve a disproportionate effort or if recording or disclosure is expressly laid down by law. In these cases Member States shall provide appropriate safeguards.

Data processing for statistical purposes is therefore recognized as a legitimized interest that may restrict data protection principles, according to national legislation. This is stipulated in Article 13 para. 2 of the Data Protection Directive, which states that subject to adequate legal safeguards, in particular that the data are not used for taking measures or decisions regarding any particular individual, Member States may, where there is clearly no risk of breaching the privacy of the data subject, restrict by a legislative measure the rights provided for in Article 12 when data are processed solely for purposes of scientific research or are kept in personal form for a period which does not exceed the period necessary for the sole purpose of creating statistics.

b. The e-Privacy Directive

While Directive 95/46 is of a general nature, there are specific EU provisions for the protection of privacy and data protection in the field of electronic communication. The e-Privacy Directive 2002/58/EC⁷⁵ contains a set of legally binding rules concerning some fields of data processing in the electronic communications sector. The e-Privacy Directive was amended by Directive 2009/136/EC⁷⁶. There are no specific rules governing data collection for statistical purposes in this legal framework. As a result, the general provisions on data collection for statistical purposes apply also in the electronic communications network.

Nevertheless, one should keep in mind that the e-Privacy Directive contains specific rules on mechanisms of data collection in the digital environment. From this point of view, there are provisions that may have a direct impact in assessing mechanisms that collect data from the Internet or other digital networks.

According to Article 1 para. 1 of the e-Privacy Directive, its provisions provide for the harmonization of the national provisions required to ensure an equivalent level of protection of fundamental rights and freedoms, and in particular the right to privacy and confidentiality, with respect to the processing of personal data in the electronic communication sector and to ensure the free movement of such data and of electronic communication equipment and services in the Community.

Article 3 defines the scope of the e-Privacy Directive as follows:

75 Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). Official Journal L 201 , 31/07/2002 P. 0037 - 0047

76 Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws Text with EEA relevance. Official Journal L 337 , 18/12/2009 P. 0011 - 0036

This Directive shall apply to the processing of personal data in connection with the provision of publicly available electronic communications services in public communications networks in the Community, including public communications networks supporting data collection and identification devices.

Article 4 para. 1 (“Security of processing”) states that the provider of a publicly available electronic communications service must take appropriate technical and organizational measures to safeguard security of its services, if necessary in conjunction with the provider of the public communications network with respect to network security. Having regard to the state of the art and the cost of their implementation, these measures shall ensure a level of security appropriate to the risk presented. According to para. 2, in case of a particular risk of a breach of the security of the network, the provider of a publicly available electronic communications service must inform the subscribers concerning such risk and, where the risk lies outside the scope of the measures to be taken by the service provider, of any possible remedies, including an indication of the likely costs involved. According to para. 3, in the case of a personal data breach, the provider of publicly available electronic communications services shall, without undue delay, notify the personal data breach to the competent national authority.

Article 5 (“Confidentiality of the communications”) obliges the Member states to prohibit listening, tapping, storage or other kinds of interception or surveillance of communications and the related traffic data by persons other than users, without the consent of the users concerned, except when legally authorized to do so in accordance with Article 15 para. 1. This provision does not affect any legally authorized recording of communications and the related traffic data when carried out in the course of lawful business practice for the purpose of providing evidence of a commercial transaction or of any other business communication. Member States shall ensure that the storing of information, or the gaining of access to information already stored, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned has given his or her *consent*, having been provided with clear and comprehensive information, in accordance with Directive 95/46/EC, *inter alia*, about the purposes of the processing. This shall not prevent any technical storage or access for the sole purpose of carrying out the transmission of a communication over an electronic communications network, or as strictly necessary in order for the provider of an information society service explicitly requested by the subscriber or user to provide the service.

Specific provisions of the e-Privacy Directive regulate the processing of traffic data and location data. According to Article 6 data relating to subscribers and users processed and stored by the provider of a public communications network or publicly available electronic communications service must be erased or made anonymous when it is no longer needed for the purpose of the transmission of a communication. Traffic data necessary for the purposes of subscriber billing and interconnection payments may be processed. Such processing is permissible only up to the end of the period during which the bill may lawfully be challenged or payment pursued. According to Article 9, where location data other than traffic data, relating to users or subscribers of public communications networks or publicly available electronic communications services, can be processed, such data may only be processed when they are made anonymous, or with the consent of the users or subscribers to the extent and for the duration necessary for the provision of a value added service. The service provider must inform the users or subscribers, prior to obtaining their consent, of the type of location data other than traffic data which will be processed, of the purposes and duration of the processing and whether the

data will be transmitted to a third party for the purpose of providing the value added service. Users or subscribers shall be given the possibility to withdraw their consent for the processing of location data other than traffic data at any time. Where consent of the users or subscribers has been obtained for the processing of location data other than traffic data, the user or subscriber must continue to have the possibility, using a simple means and free of charge, of temporarily refusing the processing of such data for each connection to the network or for each transmission of a communication.

c. Council Framework Decision on data protection in the framework of police and judicial cooperation in criminal matters

The Data Protection Directive and the e-Privacy Directive contain provisions that apply to the former “First Pillar” according to a former version of the European Union Treaty (namely: the European Community law). After the Lisbon Treaty, the scope of the secondary community legislation obtains a new dimension, which does not fall within the aim of this study to describe. Under the three-pillars system, the European Union adopted a specific set of data protection rules applying in the framework of police and judicial cooperation in criminal matters. This is the Data Protection Framework Decision⁷⁷, which contains specific provisions for data protection in this field.

According to Nr. 6 of the preamble, the Data Protection Framework Decision applies only to data gathered or processed by competent authorities for the purpose of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties. This Framework Decision should leave it to Member States to determine more precisely at national level which other purposes are to be considered as incompatible with the purpose for which the personal data were originally collected. In general, further processing for historical, statistical or scientific purposes should not be considered as incompatible with the original purpose of the processing.

The non-incompatibility principle is stipulated in Article 3 of the Decision (“Principles of lawfulness, proportionality and purpose”):

“1. Personal data may be collected by the competent authorities only for specified, explicit and legitimate purposes in the framework of their tasks and may be processed only for the same purpose for which data were collected. Processing of the data shall be lawful and adequate, relevant and not excessive in relation to the purposes for which they are collected.

2. Further processing for another purpose shall be permitted in so far as:

(a) it is not incompatible with the purposes for which the data were collected;

⁷⁷ Council Framework Decision 2008/977/JHA of 27 November 2008 on the protection of personal data processed in the framework of police and judicial cooperation in criminal matters

(b) the competent authorities are authorised to process such data for such other purpose in accordance with the applicable legal provisions; and

(c) processing is necessary and proportionate to that other purpose.

The competent authorities may also further process the transmitted personal data for historical, statistical or scientific purposes, provided that Member States provide appropriate safeguards, such as making the data anonymous.”

One more exceptional provision for statistical purposes is contained in Article 11 (“Processing of personal data received from or made available by another Member State”)

“Personal data received from or made available by the competent authority of another Member State may, in accordance with the requirements of Article 3(2), be further processed only for the following purposes other than those for which they were transmitted or made available:

(a) the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties other than those for which they were transmitted or made available;

(b) other judicial and administrative proceedings directly related to the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties;

(c) the prevention of an immediate and serious threat to public security; or

(d) any other purpose only with the prior consent of the transmitting Member State or with the consent of the data subject, given in accordance with national law.

The competent authorities may also further process the transmitted personal data for historical, statistical or scientific purposes, provided that Member States provide appropriate safeguards, such as, for example, making the data anonymous.”

d. Council of Europe Treaties

The Council of Europe was established in 1949 to enable governments of the European states to co-operate *"to achieve a greater unity between its members for the purpose of safeguarding and realising the ideals and principles which are their common heritage and facilitating their economic and social progress"* (Article 1 of the Statute of the Council of Europe). The international organization is governed by the Committee of Ministers of Foreign Affairs of the member states, which is advised by the Parliamentary Assembly, and many intergovernmental committees of experts dealing with most aspects of the daily life of European citizens, except defence: human rights, harmonization of law, culture and education, social affairs, public health and the economy. The Council of Europe's activities focus in particular on "topical issues" such as problems linked to drugs, terrorism, refugees and the prevention of torture.

The Council of Europe Convention for the Protection of Human Rights and Fundamental Freedoms was opened for signature in 1950. Article 8 of this Convention states that "everyone has the right to respect for his private and family life, his home and his correspondence". This right can be restricted by a public authority only in accordance with domestic law and in so far as it is necessary, in a democratic society, for the defence of a number of legitimate aims. But the Convention also lays down, in Article 10, the fundamental right to freedom of expression. This right includes explicitly the "freedom to receive and impart information and ideas without interference by public authority and regardless of frontiers". The "freedom to receive information" set out in Article 10 is considered as implying the "freedom to seek information". Articles 8 and 10 are not contradictory but complementary. However, in practice, the exercise of one of these rights can be restricted by the exercise of the other. For this reason, the European Commission and Court of Human Rights have defined in case-law the limits to the exercise of each of these rights and, in particular, the extent to which public authorities have the right to interfere. This case-law has been - and still is - of great importance to the Council of Europe in its work on data protection as the source of criteria for the development of national regulations on data protection. Nevertheless, in the years following the adoption of the European Convention on Human Rights, it became apparent that efficient legal protection of privacy required more specific and systematic development.

The first international legally binding text on data protection was adopted by the Council of Europe Member States in 1981. The European Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data⁷⁸ is a "first generation" international treaty that has been ratified by all countries of the European area. Even in this primary piece of legislation, restrictions to national data protection rules for statistical purposes were expressly considered as acceptable. According to the Convention's Article 9, restrictions on the exercise of the rights specified in Article 8, paragraphs b, c and d, may be provided by law with respect to automated personal data files used for statistics or for scientific research purposes when there is obviously no risk of an infringement of the privacy of the data subjects. The Council of Europe Convention served as a model for the drafting of Directive 95/46/EC.

The current impact of the Data Protection Convention with regard to the processing of personal data for statistical purposes is connected mainly to a secondary Council of Europe text that applies the Convention's principles to the special sector of statistical activities. Recommendation No. R (97) 18 concerning the protection of personal data collected and processed for statistical purposes⁷⁹ was adopted by the Council of Europe's Committee of Ministers on 30 September 1997. This text replaced Recommendation No. R (83) 10 on the protection of personal data used for scientific research and statistics in so far as that recommendation applies to the collection and automatic processing of personal data for statistical purposes.

According to its preamble, the Recommendation reads that the Committee of the Ministers recognizes that "the production of reliable statistics depends to a great extent on the collection of the most detailed information possible and on the processing of this information by means of increasingly effective automatic data processing technology", while it is also "aware of the fact that such information may concern identified or identifiable persons ("personal data")" and "aware of the need to develop techniques

78 European Treaty Series, No. 108, <http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm>

79 Text available on <https://wcd.coe.int/com.instranet.InstraServlet?command=com.instranet.CmdBlobGet&InstranetImage=2001724&SecMode=1&DocId=578856&Usage=2>

making it possible to guarantee the anonymity of the data subjects” and of “the concern of the international community of statisticians for the protection of personal data, and the development of international recommendations with regard to the professional ethics of statisticians”.

The Appendix to Recommendation No. R (97) 18 contains the substantial contribution of this secondary Council of Europe text to the subject matter of data protection in the statistical sector. The Appendix contains a definitions chapter. According to this, “personal data” *“means any information relating to an identified or identifiable individual (“data subject”). An individual shall not be regarded as “identifiable” if the identification requires an unreasonable amount of time and man-power. Where an individual is not identifiable, data are said to be anonymous.”* As “identification data”, the Appendix defines those personal data *“that allow direct identification of the data subject, and which are needed for the collection, checking and matching of the data, but are not subsequently used for drawing up statistical results.”* As “sensitive data” the Appendix defines the ones that have been defined as “special categories” of data by the Data Protection Convention”: racial origin, political opinions, religious or other beliefs, health, sexual life, criminal convictions “and other data defined as sensitive by domestic law”. As “processing” the Appendix defines any operation or set of operations carried out partly or completely with the help of automated processes and applied to personal data, *“such as storage, conservation, adaptation or alteration, extraction, consultation, utilization, communication, matching or interconnecting and erasure or destruction.”* The Appendix contains an additional definition, for the term of “communication”. It refers to the act of *“making personal data accessible to third parties, regardless of the means or media used”*. There are two different definitions for the terms “statistical purposes” and “statistical results”. The first term refers to *“any operation of collection and processing of personal data necessary for statistical surveys or for the production of statistical results. Such operations exclude any use of the information obtained for decisions or measures concerning a particular individual”*. The second term means information which has been obtained by processing personal data *“in order to characterize a collective phenomenon in a considered population”*.

Chapter 2 to the Appendix defines the scope of the recommendation, which includes the collection and automated processing of personal data for statistical purposes and extends to the statistical results, to the extent that they permit identification of data subjects. The scope chapter provides that no personal data shall be processed in a non-automatic manner in order to avoid the provisions of this recommendation.

Chapter 3 to the Appendix (“Respect for privacy”) contains three general principles concerning the right to privacy.

- (a) Privacy should be respected in all three stages of personal data collection and processing:
 - when these data are kept for future use;
 - when statistical results are disseminated;
 - when, for reasons of better ensuring that statistical records are representative or for reasons of confidentiality, personal data need to be modified.
- (b) Persons involved in a statistical activity that contains personal data collection and processing shall be subject to a duty of professional secrecy by domestic law or practice.
- (c) Personal data collected and processed for statistical purposes shall be made anonymous as soon as they are no longer necessary in an identifiable form.

Chapter 4 to the Appendix contains general conditions for lawful collection and processing for statistical purposes. Under the title “*Purpose*”, this Chapter stipulates a more concrete application of the purpose limitation principle: “personal data collected and processed for statistical purposes shall serve only those purposes. They shall not be used to take a decision or measure in respect of the data subject, nor to supplement or correct files containing personal data which are processed for non – statistical purposes. Processing for statistical purposes of personal data collected for non-statistical purposes is not incompatible with the purpose(s) for which the data were initially collected if appropriate safeguards are provided for, in particular to prevent the use of data for supporting decisions or measures in respect of the data subject. Under the title “*Lawfulness*”, the Chapter reiterates the legality criteria that were previously stipulated in Article 6 of the Data Protection Directive and the transparency obligations set forth in Section IV of the Data Protection Directive. Consent plays a crucial role when examining the legality of data processing for statistical purposes, while the Appendix adds a provision according to which “personal data may be collected on a compulsory basis with a view to their being processed for statistical purposes only if required by domestic law”. According to the proportionality principle, “*only those personal data shall be collected and processed which are necessary for the statistical purposes to be achieved. In particular, identification data shall be collected and processed only if this is necessary.*” Under the title “*Sensitive data*”, the Appendix reiterates that if these data are to be processed for statistical purposes, they should be collected in a form in which the data subject is not identifiable. In the case the statistical purposes necessitates the identification of the data subjects, domestic law shall provide appropriate safeguards including specific measures to separate identification data as from the stage of collection unless it is manifestly unreasonable or impracticable to do so.

Chapter 5 to the Appendix provides extensive conditions of information to be given to the data subject. Under the title “*Primary collection*”, the text reads that the persons questioned shall be informed of the following elements:

- (a) the compulsory or optional nature of the response and the legal basis, if any, of the collection,
- (b) the purpose or purposes of the collection and processing,
- (c) the name and position of the person or body in charge of the collection and/or processing,
- (d) the fact that the data will be kept confidential and used exclusively for statistical purposes,
- (e) the possibility of obtaining further information on request.

At their request and/or according to the ways and means defined by domestic law, data subjects shall also be informed of the following:

- (f) the way in which consent can be refused or withdrawn, in the case of optional surveys and, in the case of compulsory surveys, the possible sanctions this would entail,
- (g) where applicable, the conditions of the exercise of the rights of access and rectification,
- (h) the categories of persons or bodies to whom the personal data may be communicated,
- (i) the guarantees to ensure the confidentiality and the protection of personal data,
- (j) the categories of data collected and processed.

When the data subjects are not directly questioned, they shall be informed of the existence of the collection unless this is manifestly unreasonable or impracticable. They shall be able to inform

themselves appropriately of the elements listed above. The persons questioned shall be informed at the latest at the time of collection. Under the title “Secondary collection”, the Chapter reads that cases of processing or communication for statistical purposes of personal data collected for non-statistical purposes shall receive suitable publicity. The data subjects shall be able to obtain in a suitable way all abovementioned information, unless:

- (a) this is impossible or involves a disproportionate effort,
- (b) the processing or communication of the data for statistical purposes is expressly provided for under domestic law.

Chapter 6 to the Appendix (“*Consent*”) reiterates that consent of the data subject, when required, shall be free, informed and unambiguous and that the data subject shall be able to withdraw his or her consent for a single survey, as long as, identification data have not been separated from other data collected, or to suspend at any time and without retroactive effect his or her co-operation in a survey which extends over a period of time. Refusal to reply shall not be penalized unless domestic law provides for sanctions.

Chapter 7 to the Appendix provides for the rights of access and rectification. Any person may obtain the personal data concerning him or her held by the data controller and, as the case may be, have them rectified. However, where there is clearly no risk of breaching the privacy of the data subject, this right may be restricted in accordance with domestic law when the personal data are processed solely for statistical purposes and specific appropriate measures exist to prevent any identification by a third party on the basis of individual data or of statistical results.

Under the title “Rendering data anonymous” (Chapter 8), the Appendix introduces the principle that personal data collected for statistical purposes shall be made anonymous immediately after the end of data collection, checking or matching operations, except:

- (a) if identification data remain necessary for statistical purposes and the measures prescribed by principle 10.1 have been taken; or
- (b) if the very nature of statistical processing necessitates the starting of other processing operations before the data have been made anonymous as long as the safeguards envisaged in principles 15.1. to 15.3 are in force.

Reiterating the fairness of data collection principle, Chapter 9 (“Primary collection of personal data for statistical purposes”) to the Appendix underlines that personal data shall be collected only from a person other than the data subject if domestic law provides for it and includes appropriate safeguards, or there is manifestly no risk of infringement of the rights and fundamental freedoms of the data subject. Exemptions are recognized where domestic law includes appropriate safeguards and:

- (a) provides for the collection with identification data or
- (b) permits the linking of the data collected to identification data for the construction of samples.

According to this Chapter, data on non-respondents relevant to the planning or carrying out of the survey, or information on the reasons for non – response, may be used only in order to ensure the representative

quality of the survey. The controller shall take appropriate measures to allow the persons questioned to assure themselves of the authority to act of the person collecting the data.

The Appendix contains also two principles on “Identification data” (Chapter 10). When these data are collected and processed for statistical purposes, they shall be separated and conserved separately from other personal data, unless it is manifestly unreasonable or impracticable to do so. These data may, however, be used to create a file of addresses for statistical purposes if provided for by domestic law, if the data subject has been informed and has not opposed it, or if the data come from a file accessible to the public.

With regard to the conservation of data, Chapter 11 provides that, unless they have been made anonymous, or domestic law provides for these data to be kept for archiving purposes subject to appropriate safeguards, personal data collected and processed for statistical purposes shall be destroyed or erased when they are no longer necessary for those purposes. In particular, identification data shall be destroyed or erased as soon as they are no longer necessary:

- (a) for the collection, checking and matching of the data; or
- (b) to ensure the representativeness of the survey; or
- (c) to repeat the survey with the same people.

Under the title “Communication”, Chapter 12 to the Appendix states that personal data collected for statistical purposes shall not be communicated for non-statistical purposes. Nevertheless, personal data processed for a given statistical purpose may be communicated for other statistical purposes as long as these are specified and of limited duration. Communication in accordance with this principle shall be the subject of a written document setting out the rights and obligation of the parties, unless safeguards are provided for by domestic law. The controller shall in particular:

- (a) stipulate that the third party may communicate these data only with the express agreement of the said controller;
- (b) stipulate that the third party take appropriate security measures, in accordance with principles 15.1 to 15.3 of this recommendation and
- (c) ensure that any publication of statistical results obtained by this party will conform with principle 14 of this recommendation.

Sensitive data communication is allowed where provided for by the law, or where the data subjects have given their explicit consent and provided domestic law does not prohibit the giving of the consent.

According to Chapter 13, the principles of this recommendation shall be applicable to the transborder communication of personal data for statistical purposes, under the relevant provisions of the Data Protection Convention (and its Protocol on transborder data flows, that had not entered into force when the recommendation was adopted).

Statistical results shall be published or made accessible to third parties only if measures have been taken to ensure that the data subjects are no longer identifiable on the basis of these results, unless dissemination or publication manifestly presents no risk of infringing the privacy of the data subjects (Chapter 14).

With regard to security of personal data, Chapter 15 reiterates general principles concerning the relevant obligations of the data controller. If data must be retained in an identifiable form, organisational and technical resources, in particular automated resources, shall be used to prevent unauthorized identification of the data subject. Measures shall be taken to prevent re-identification of data subjects and use for non-statistical purposes of personal data collected for statistical purposes. Professionals, firms or bodies in charge of producing statistics shall develop techniques and procedures ensuring the anonymity of data subjects. According to Chapter 16 (“Codes of ethics”), entities in charge of producing statistics should adopt and publish codes of professional ethics which meet the principles set out in this recommendation, in particular:

- (a) on the other categories of persons and bodies which have access to the personal data;
- (b) on the measures to be taken for the protection, confidentiality and security of these data as well as measures to respect statistical ethics;
- (c) on the controllers of statistical processing.

According to Chapter 17, in order to ensure broad access of information tools and to technical knowledge appropriate to effective protection of personal data collected for statistical purposes, competent governmental bodies should collaborate closely in the development of these tools and technical knowledge, and should set up international programmes of co-operation, exchanges of experience, transfer of knowledge and technical assistance. According to Chapter 18, member states give one or more independent authorities responsibility for ensuring the application of the provisions of domestic law giving effect to the principles laid down in the recommendation.

For a deeper analysis of the principles laid down in the recommendation, an Explanatory Memorandum⁸⁰ is also available.

6.4. The *sui generis* Database Right

A database right is a special formulation of copyright legal provisions that exist to recognize the investment that is made in compiling a database, even when this does not involve the “creative” aspect that is reflected by copyright. In European Union law, database rights are specifically coded laws on the copying and dissemination of information in computer databases. These rights were first introduced in 1996. The relevant legally binding instrument is Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

The Database Directive contains a definition for the “database”, according to which this term shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. It is expressly stipulated that protection under the Database Directive shall not apply to computer programs used in the making or operation of databases accessible by electronic means.

According to Article 2, there are some limitations on the scope of the Database Directive: it shall apply without prejudice to Community provisions relating to:

80 [http://www.coe.int/t/dghl/standardsetting/dataprotection/EM/EM_R\(97\)18_EN.pdf](http://www.coe.int/t/dghl/standardsetting/dataprotection/EM/EM_R(97)18_EN.pdf)

- (a) the legal protection of computer programs;
- (b) rental right, lending right and certain rights related to copyright in the field of intellectual property;
- (c) the term of protection of copyright and certain related rights.

The object of the legal protection provided for by the Directive is described in Article 3. Databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation shall be protected as such by copyright. No other criteria shall be applied to determine their eligibility for that protection. The copyright protection of databases provided for by the Directive shall not extend to their contents and shall be without prejudice to any rights subsisting in those contents themselves.

Article 4 to the Directive defines the database authorship. The author of a database shall be the natural person or group of natural persons who created the base or, where the legislation of the Member States so permits, the legal person designated as the right holder by that legislation. Where collective works are recognized by the legislation of a Member State, the economic rights shall be owned by the person holding the copyright. In respect of a database created by a group of natural persons jointly, the exclusive rights shall be owned jointly.

According to Article 5 ("Restricted acts") in respect of the expression of the database which is protectable by copyright, the author of a database shall have the exclusive right to carry out or to authorize:

- (a) temporary or permanent reproduction by any means and in any form, in whole or in part;
- (b) translation, adaptation, arrangement and any other alteration;
- (c) any form of distribution to the public of the database or of copies thereof. The first sale in the Community of a copy of the database by the rightholder or with his consent shall exhaust the right to control resale of that copy within the Community;
- (d) any communication, display or performance to the public;
- (e) any reproduction, distribution, communication, display or performance to the public of the results of the acts referred to in (b).

Article 6 provides for exceptions to restricted acts. The performance by the lawful user of a database or of a copy thereof of any of the acts listed in Article 5 which is necessary for the purposes of access to the contents of the databases and normal use of the contents by the lawful user shall not require the authorization of the author of the database. Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part. Member States shall have the option of providing for limitations on the rights set out in Article 5 in the following cases:

- (a) in the case of reproduction for private purposes of a non-electronic database;
- (b) where there is use for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved;
- (c) where there is use for the purposes of public security or for the purposes of an administrative or judicial procedure;
- (d) where other exceptions to copyright which are traditionally authorized under national law are involved, without prejudice to points (a), (b) and (c).

In accordance with the Berne Convention for the protection of Literary and Artistic Works, this Article may not be interpreted in such a way as to allow its application to be used in a manner which unreasonably prejudices the rightholder's legitimate interests or conflicts with normal exploitation of the database.

The “sui generis” database right is stipulated in Article 7 of the Directive. Member States shall provide for a right for the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database. The sui generis right may be transferred, assigned or granted under contractual license. It shall also apply irrespective of the eligibility of that database for protection by copyright or by other rights. Moreover, it shall apply irrespective of eligibility of the contents of that database for protection by copyright or by other rights. Protection of databases under the right provided for in paragraph 1 shall be without prejudice to rights existing in respect of their contents. For the purposes of this Directive:

- (a) 'extraction' shall mean the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form;
- (b) 're-utilization' shall mean any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission. The first sale of a copy of a database within the Community by the rightholder or with his consent shall exhaust the right to control resale of that copy within the Community; public lending is not an act of extraction or re-utilization.

The repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database shall not be permitted.

Article 8 provides for rights and obligations of lawful users. The maker of a database which is made available to the public in whatever manner may not prevent a lawful user of the database from extracting and/or re-utilizing insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever. Where the lawful user is authorized to extract and/or re-utilize only part of the database, this paragraph shall apply only to that part. A lawful user of a database which is made available to the public in whatever manner may not perform acts which conflict with normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database. A lawful user of a database which is made available to the public in any manner may not cause prejudice to the holder of a copyright or related right in respect of the works or subject matter contained in the database.

Exceptions to the sui generis right are mentioned in Article 9 to the Directive. Member States may stipulate that lawful users of a database which is made available to the public in whatever manner may, without the authorization of its maker, extract or re-utilize a substantial part of its contents:

- (a) in the case of extraction for private purposes of the contents of a non-electronic database;
- (b) in the case of extraction for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved;

- (c) in the case of extraction and/or re-utilization for the purposes of public security or an administrative or judicial procedure.

The right provided for in Article 7 shall run from the date of completion of the making of the database. It shall expire fifteen years from the first of January of the year following the date of completion. In the case of a database which is made available to the public in whatever manner before expiry of the period provided for, the term of protection by that right shall expire fifteen years from the first of January of the year following the date when the database was first made available to the public. Any substantial change, evaluated qualitatively or quantitatively, to the contents of a database, including any substantial change resulting from the accumulation of successive additions, deletions or alterations, which would result in the database being considered to be a substantial new investment, evaluated qualitatively or quantitatively, shall qualify the database resulting from that investment for its own term of protection.

According to Article 11 of the Directive, the right provided for in Article 7 shall apply to database whose makers or rightholders are nationals of a Member State or who have their habitual residence in the territory of the Community. This shall also apply to companies and firms formed in accordance with the law of a Member State and having their registered office, central administration or principal place of business within the Community; however, where such a company or firm has only its registered office in the territory of the Community, its operations must be genuinely linked on an ongoing basis with the economy of a Member State. Agreements extending the right provided for in Article 7 to databases made in third countries and falling outside the abovementioned provisions shall be concluded by the EU Council acting on a proposal from the EU Commission. The term of any protection extended to databases by virtue of that procedure shall not exceed that available pursuant to Article 10.

According to Article 13, the Database Directive shall be without prejudice to provisions concerning in particular copyright, rights related to copyright or any other rights or obligations subsisting in the data, works or other materials incorporated into a database, patent rights, trade marks, design rights, the protection of national treasures, laws on restrictive practices and unfair competition, trade secrets, security, confidentiality, data protection and privacy, access to public documents, and the law of contract.

According to Article 14, protection pursuant to the Database Directive as regards copyright shall also be available in respect of databases created prior to the date referred to in Article 16 which on that date fulfill the requirements laid down in this Directive as regards copyright protection of databases. Notwithstanding the abovementioned provision, where a database protected under copyright arrangements in a Member State on the date of publication of the Directive does not fulfill the eligibility criteria for copyright protection laid down in Article 3 the Directive shall not result in any curtailing in that Member State of the remaining term of protection afforded under those arrangements. Protection pursuant to the provisions of the Directive as regards the right provided for in Article 7 shall also be available in respect of databases the making of which was completed not more than fifteen years prior to the date referred to in Article 16 (1) and which on that date fulfill the requirements laid down in Article 7. The protection provided for in the abovementioned provisions shall be without prejudice to any acts concluded and rights acquired before the date referred to in those paragraphs. In the case of a database the making of which was completed not more than fifteen years prior to the date referred to in Article 16 (1), the term of protection by the right provided for in Article 7 shall expire fifteen years from the first of January following that date.

Article 15 states that any contractual provision contrary to Articles 6 (1) and 8 shall be null and void.

The Database Directive provides no mandatory public-interest exceptions comparable to those recognized under domestic and international copyright laws. An optional exemption concerning “illustrations for teaching or scientific research” applies to extractions but not reutilization⁸¹.

6.5. Conclusion

The overall project seems to be compatible with relevant data protection and database right rules. The prior consent and permissions should comply with the abovementioned provisions. The compliance is a matter of properly drafted Terms of Service to which the end user and the companies may opt in, before the installation / operation of the data collection software to their devices or web pages. The examination of the Terms of Service by the independent Data Protection Authorities in the territories exposed to the project would also provide for an additional confirmation of the legal compatibility.

7. Socio-political acceptance

In Europe the right to privacy is enshrined in the European Convention of Human rights⁸² reflecting an approach in society that values privacy and personal dignity on par with freedoms unlike in other regions like the US. Recent events, like the Snowden revelations for large government operations that collect data on individuals at a huge scale worldwide have increased public awareness on the issue of privacy with respect to governments and big data holders, especially in Europe.

In this section we will examine attitudes of stakeholders (individuals and businesses) towards a system of data collection that collects data for statistical purposes from their day to day actions.

User centric approach

In user centric approaches the individual is the reference unit and the stakeholder. Based on our pilot exercise participating users were also asked whether they had reservations about installing a data collection application and to describe them. Most of the respondents (38/40 i.e. 79%) did not have reservations and 10 (21%) provided some.

We have identified four issues that should be considered as generating (justified or not) reservations for participating in a user centric data collection system.

- Intrusiveness. Obtaining too much information. Statistical data collections for official statistics, while handling sensitive information about individuals and households, are rarely intrusive. One common exception is the information on income that is well known that generates both frustration and has relatively low response rate⁸³. The low response rate for some aspects of income indicates that some respondents have limits on the kind of information that they are ready to provide

81 “The role of scientific and technical data and information in the public domain”, Proceedings of a symposium, National Research Council of the National Academies, Washington DC 20001, 2003, *Jerome Reichman*, Discussion Framework, p.82

82 Article 8 stipulates that “Everyone has the right to respect for his private and family life, his home and his correspondence” subject to certain restrictions.

83 An assessment of survey errors in EU-SILC, ISSN 1977-0375, p.32, available at http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-10-021/EN/KS-RA-10-021-EN.PDF

in official statistics surveys. An application that is installed in the devices that a person uses to access the internet may record and dispatch information that is really too sensitive for many persons to accept even if they trust that it will only be used for statistical purposes. This issue has been brought up by 2 members of our sample, one was worried whether the content of chats and other personal communication with family and friends was recorded and another gave a general statement that the computer is used for personal matters and was sceptical about sharing these uses with others. It is therefore important that information that is collected and transmitted to the NSI is as little sensitive as it can eg. reporting category of websites visited and not individual sites.

- Confidentiality is also an issue that concerns users. Confidentiality protection is enshrined in statistical law in all countries and for all statistics produced. From comparative results in certain countries, there is more trust on behalf of the public that their data is kept confidential than to Statistical institutes in general⁸⁴ but it is nevertheless an issue that needs to be addressed.
- Security. Installing an application that collects information in the background and then sends it to another computer over the internet poses the security risk that it can be intercepted by a third party or that it can be used as a back door to gain access to their computers. This worry has been reported by three respondents that report concerns about their computer security (two were questioning whether their passwords are safe) as well as their concern that third parties could obtain personal information. It is essential that these fears are taken seriously in the design of the software as well as assuring users for the safety of the operation.
- Transparency. What exactly the software does after it is installed in a device may also worry respondents. Two respondents express them, revealing a need to explain the way the application functions in a clear, comprehensive and specific manner (what it does and what it does not). It may also be useful to allow for verification of these claims by revealing the source code of the software although this may compromise security.

When users were asked to name conditions required for accepting to participate in such a survey they mostly named confidentiality issues (16/26), i.e. preserving anonymity. Security was also reported by some respondents (3/26). Issues related with the use of the device (slowing down, leaving traces, ease of installing and uninstalling) were indicated by three respondents (3/26). Two respondents noted that their participation depended on whether they were interested on the scope of the survey and one mentioned the degree of trust to the responsible institution. Another issue that was brought up by several respondents (7/26) was a potential incentive that they required in order to participate.

Incentives, whether monetary or nonmonetary can be considered as an inducement offered by the survey designer to compensate for the absence of factors that might otherwise stimulate cooperation--e.g., interest in the topic of the survey or a sense of civic obligation.⁸⁵ Although Official Statistics Institutes are reluctant to use incentives, which, among other issues, may render a survey very expensive they should contemplate their use when requiring installation of software in respondents devices.

- Such software uses computational resources of the device. Although it should have only a small effect on device performance this is certainly not zero. Incentives can be seen as some sort of partially renting the respondent's equipment.
- Transmission of data via 3G/4G networks may entail actual costs for cooperation that users are entitled to ask compensation for.

⁸⁴ Public image survey of Statistics Denmark, 2011, available at <http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2759>

⁸⁵ Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1), 112-141.

Site centric approach

The data collection tool chosen for the pilot used Google's Custom Search Engine in order to detect keywords in the enterprises websites' content that has already been indexed by Google. 63 enterprises (approximately 20%) have been randomly selected among those listed in our inventory. Among the 63 randomly selected enterprises, two enterprises do not have active websites and have been excluded from the analysis. Thus, the eligible sample is 61 enterprises. In the course of this assessment, we have contacted the owners of the 61 randomly selected websites in order to investigate whether they are willing to accept and implement the proposed new method of data collection. We have prepared a questionnaire, which outlined the proposed method and indicators and posed five questions in order to collect their opinions about them.

Out of the 61 selected websites that were contacted, 27 (44,3%) websites' owners have replied and 16 (26,2%) have refused to take part. The rest of the 18 (29,5%) websites owners never replied.

On first question asking for preference for data collection (automatic vs questionnaire based) the responses from website managers was divided. Considering that computing the indicators and filling up the questionnaire results in some burden we can ascertain that website managers have some reservations about allowing such automatic data collection tools.

1st question: Having read the accompanied document that lists the specific indicators related to your website, which way would you prefer in order to provide data for those indicators to an Official Statistical Institution?	N	%
Via automatic collection from my website without my interference	12	44.4
Via an appropriate questionnaire	12	44.4
Via either of the first two ways	1	3.7
None of the first two ways	1	3.7
I do not know/No answer	1	3.7
Total	27	

Some of those that opposed automatic collection (3 out of 13) did not mind if collection was implemented manually yet still from the statistical institute and not themselves, although most retained their objection and wanted to have responsibility for data referring to their sites.

2nd question: If an employee from an Official Statistical Institution manually visited your website and recorded the requested data, would you still be opposed? (only for those who answered “Via an appropriate questionnaire” or “None of the first two ways” in the 1st question)	N	%
Yes	10	76.9
No	3	23.1
I do not know/No answer	-	-
Total	13	

When asked about the reasons for opposing automatic collection most respondents (7/13) did not elaborate, two refused citing general reasons while the rest suggested that they want to be fully informed of the data content as well as the data collection process, while some also noted the need some verification.

3rd question: Can you please specify the reasons, why you do not wish to automatically collect data from your website? (only for those who answered “Via an appropriate questionnaire” or “None of the first two ways” in the 1st question)	N	%
No reason	7	53.8
I do not think it is necessary	1	7.7
I do not want the collected data from my site to be published by an Official Statistical Institution or to be known to my competitors	1	7.7
I want to be informed every time about which data will be used and the nature of the survey	1	7.7
I would agree to an automatic data collection if only the requested data was the one that it is described in your document. If more data is going to be collected, such as measuring website’s traffic then I am opposed.	1	7.7
In order always to be able to verify the information/data is going to be requested	1	7.7
We want to know, every time, the requested information	1	7.7
Total	13	

4th question: In order to give your permission for an automatic data collection from your website, would you require some kind of a confidentiality guarantee?	N	%
--	----------	----------

Yes	14	51.8
No	10	37.0
I do not know/No answer	3	11.1
Total	27	

Respondents willing to cooperate mostly required some sort of bilateral agreement. Only two were satisfied with general confirmation and assurances on behalf of the statistical institute. Most of those requiring some sort of agreement wanted a cooperation agreement (9) rather than a confidentiality agreement (3). Only one respondent required financial compensation as part of the cooperation agreement.

	5th question: What kind of confidentiality guarantee would you require? (only for those you answered "Yes" in 4th question)	N	%
Confidentiality Assurance	Written confirmation that the data will be used only for the purposes of this research and will not be used for other purposes or disclosure to third parties	1	7.1
	Assurance of anonymity	1	7.1
Confidentiality agreement	Confidentiality agreement	2	14.3
	Privacy policy agreement	1	7.1
Cooperation agreement	Cooperation agreement	7	50.0
	Written agreement that data will not be used for commercial purposes and copywrite will be protected	1	7.1
	Financial compensation and a cooperation agreement	1	7.1
	Total	14	

From our small sample of website managers it seems that about half will not cooperate with an automatic survey (although some of them might be turned if they have full information on the collection process and access to the data transmitted). Those that can potentially agree see themselves as partners and not just respondents and require bilateral agreements rather than self-imposed rules and commitments from the National Statistical Institute.

8. Conclusions

Two separate production processes, one web site-centric and the other user-centric have been examined in this report:

- the production of statistics on the characteristics of business web sites, based on data collected with the help of crawlers or search engines that rely on earlier crawling from the said web sites.
- the production of statistics on the use of Internet by individuals, based on data collected with the help of monitoring software installed on the users' devices.

The two processes have been examined from several angles.

Technically they are both feasible. Software components are available in several forms and the software technologies needed for development from scratch are commonplace. The capacities needed for development and maintenance are quite easy to find in the job market even if not already available to the NSIs.

The processes are also acceptable in the ESS, according to the small sample of NSIs that were interviewed. The NSI most opposed to these processes was mainly not aware of their details and potential, and expressed concerns about the additional workload that they would impose. In general however, NSIs are at least curious about these methods and see their potential. Some of them are already studying them.

The two processes diverge in the conclusions about their methodological feasibility. The both produce very relevant, timely and rich-in-detail statistics. Compared to the current ICT surveys the web-site centric process has a much narrower scope: it substitutes and expands a small subset of the current survey's indicators, while the user-centric process can reproduce most current indicators. The user-centric process thus also offers great savings in response burden. Both have accuracy issues: the web site-centric one suffers from measurement errors, in its keyword-based implementation and possibly by non-response. The user-centric one mainly suffers from non-response, manifested as refusals to participate or switching off of the monitoring software occasionally.

The two processes also achieve different cost-benefit balance. The web site-centric process seems to have too high costs for the benefits it offers, especially if one takes into account that it covers a small subset of current indicators and has reduced accuracy. The user-centric approach seems to be more expensive than the current ICT survey but reduces response burden and production times considerably. Unfortunately there was no detailed cost information about these processes or the current ICT surveys so as to make a more precise assessment.

The processes are compatible with current European legislation, as long as NSIs inform explicitly individuals and enterprises about the collected data and the uses they will be subjected to and they obtain the sample units' consent. In principle the processes do not differ from traditional surveys that collect sensitive business or personal data.

In user centric approach we found that most users want to cooperate and will do so if they are satisfied that their privacy and anonymity will be preserved and their use of their devices will not be affected in a substantial way. Incentives may help to further increase cooperation. Regarding the site centric approach, a large part of websites (about half) will refuse cooperation and those that can potentially agree see themselves as partners and not just respondents and require bilateral agreements rather than self-imposed rules and commitments from the National Statistical Institute.

Overall, the user-centric process is the more feasible of the two. It can replace the current ICT survey to a great extent for a not much higher cost. The same cannot be said for the web-site process. As envisaged it collects a small subset of the current survey's indicators. A variation, namely the collection of data from enterprise servers, which was outside the scope of the project, can supplement this process and can deliver a much larger set of highly relevant ICT and other enterprise data.

9. References

- [Beach2010] Beach, A., Gartrell, M., Xing, X., Han, R., Lv, Q., Mishra, S., & Seada, K. (2010, February). Fusing mobile, sensor, and social data to fully enable context-aware computing. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications* (pp. 60-65). ACM.
- [Kang2011] Kang, J. M., Seo, S. S., & Hong, J. K. (2011, September). Usage pattern analysis of smartphones. In *Network Operations and Management Symposium (APNOMS), 2011 13th Asia-Pacific* (pp. 1-8). IEEE.
- [Koster 1995] Martijn Koster, Robots in the Web: threat or treat? *ConneXions*, Volume 9, No. 4, April 1995. <http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html>
- [Miller 2012] Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, 7(3), 221-237.
- [Mokh2007] Mokhonoana, P. M., & Olivier, M. S. (2007, September). Acquisition of a Symbian smart phone's content with an on-phone forensic tool. In *Proceedings of the Southern African Telecommunication Networks and Applications Conference* (pp. 1-7).
- [Rofouei 2012] Rofouei, M., Wilson, A., Brush, A. J., & Tansley, S. (2012, May). Your phone or mine?: fusing body, touch and device sensing for multi-user device-display interaction. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (pp. 1915-1918). ACM.
- [Shepard2011] Shepard, C., Rahmati, A., Tossell, C., Zhong, L., & Kortum, P. (2011). LiveLab: measuring wireless networks and smartphone users in the field. *ACM SIGMETRICS Performance Evaluation Review*, 38(3), 15-20.
- [Souza 2010] de Souza, M., Carvalho, D. D. B., Barth, P., Ramos, J. V., Comunello, E., & von Wangenheim, A. (2010, August). Using acceleration data from smartphones to interact with 3D medical data. In *Graphics, Patterns and Images (SIBGRAPI), 2010 23rd SIBGRAPI Conference on* (pp. 339-345). IEEE
- [Vafopoulos 2011] Vafopoulos, M. (2011). The Web economy: goods, users, models and policies. *Foundations and Trends® in Web Science*, 3(1-2), 1-136. doi:<http://dx.doi.org/10.1561/18000000015>
- [Wagner2013] Wagner, D. T., Rice, A., & Beresford, A. R. Device Analyzer: Large-scale mobile data collection.
- [Weitzner et al 2008] Weitzner, D., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., & Sussman, G. J. (2008). Information accountability. *Communications of the ACM*, 51(6), 82-87

10. Annex

10.1. Appendix 1 - Synonym XML definition

```
<?xml version="1.0" encoding="UTF-8"?>
<Synonyms start="0" num="20" total="20">
  <Synonym term="B8b">
    <Variant>privacy policy</Variant>
    <Variant>terms of use</Variant>
    <Variant>Privacy Statement</Variant>
    <Variant>Conditions of use</Variant>
    <Variant>Terms and Conditions</Variant>
    <Variant>Terms & Co</Variant>
  </Synonym>
  <Synonym term="B8g">
    <Variant>jobs</Variant>
    <Variant>vacancies</Variant>
  </Synonym>
  <Synonym term="B8p1">
    <Variant>cart</Variant>
    <Variant>shopping basket</Variant>
  </Synonym>
  <Synonym term="N11a">
    <Variant>widgets</Variant>
    <Variant>Facebook</Variant>
    <Variant>LinkedIn</Variant>
    <Variant>Yammer</Variant>
    <Variant>Twitter</Variant>
    <Variant>Follow us</Variant>
    <Variant>Share this page</Variant>
    <Variant>Follow</Variant>
    <Variant>Like us</Variant>
  </Synonym>
  <Synonym term="N11b">
    <Variant>Blogs</Variant>
    <Variant>Follow</Variant>
  </Synonym>
  <Synonym term="N13">
    <Variant>Wiki</Variant>
  </Synonym>
  <Synonym term="N14">
    <Variant>Creative commons (licence)</Variant>
    <Variant>rss (feed)</Variant>
  </Synonym>
  <Synonym term="N18">
    <Variant>Workflow Engine</Variant>
  </Synonym>
  <Synonym term="N1a">
    <Variant>url</Variant>
    <Variant>Website</Variant>
  </Synonym>
  <Synonym term="N1b">
```

<Variant>e-mail</Variant>
<Variant>Email</Variant>
<Variant>E-mail</Variant>
<Variant>email</Variant>
<Variant>eMail</Variant>
</Synonym>
<Synonym term="N1c">
<Variant>telephone</Variant>
<Variant>telephone number</Variant>
<Variant>Phone</Variant>
<Variant>Tel.</Variant>
<Variant>Fax</Variant>
<Variant>Tel/Fax</Variant>
<Variant>T:</Variant>
<Variant>tel</Variant>
<Variant>TELEPHONE</Variant>
</Synonym>
<Synonym term="N1d">
<Variant>address</Variant>
<Variant>Postal Address</Variant>
<Variant>Post code</Variant>
<Variant>P.O. box</Variant>
</Synonym>
<Synonym term="N22">
<Variant>Online chat</Variant>
</Synonym>
<Synonym term="N2a">
<Variant>Language</Variant>
<Variant>Greek</Variant>
<Variant>EL</Variant>
</Synonym>
<Synonym term="N2b">
<Variant>English</Variant>
<Variant>EN</Variant>
</Synonym>
<Synonym term="N3">
<Variant>Last Update</Variant>
<Variant>Last Updated Dated</Variant>
</Synonym>
<Synonym term="N4">
<Variant>Signin</Variant>
<Variant>login</Variant>
<Variant>Login</Variant>
<Variant>register</Variant>
<Variant>Create an Account</Variant>
<Variant>openID</Variant>
<Variant>registration</Variant>
<Variant>Subscribe</Variant>
</Synonym>
<Synonym term="N5">
<Variant>sitemap</Variant>
<Variant>site map</Variant>
<Variant>SITEMAP</Variant>
<Variant>Sitemap</Variant>
<Variant>Site Map</Variant>
</Synonym>

```
<Synonym term="N6">  
<Variant>analytics</Variant>  
<Variant>googleanalytics</Variant>  
</Synonym>  
<Synonym term="N9">  
<Variant>mpeg</Variant>  
</Synonym>  
</Synonyms>
```

10.2. Appendix 2

Tools (open/free) for mobile data collection

iPhone Analyzer:(<http://www.crypticbit.com/zen/products/iphoneanalyzer>) allows you to forensically examine or recover data from an iOS device. It principally works by importing backups produced by iTunes or third party software, and providing you with a rich interface to explore, analyse and recover data in human readable formats. Because it works from the backup files everything is forensically safe, and no changes are made to the original data.

BitPim:(<http://www.bitpim.org/>) is a program that allows you to view and manipulate data on many CDMA phones from LG, Samsung, Sanyo and other manufacturers. This includes the PhoneBook, Calendar, WallPapers, RingTones (functionality varies by phone) and the Filesystem for most Qualcomm CDMA chipset based phones. To see when phones will be supported, which ones are already supported and which features are supported

VIAFORENSICS: (<https://viaforensics.com/resources/tools/>) viaForensics has developed a number of free mobile and computer forensics tools.

Mobile Internal Acquisition Tool (MIAT): (<http://computerforensics.champlain.edu/blog-tags/mobile-internal-acquisition-tool>). The tool is presented in [Distefano2008]. It seems that it is freely available after request to authors.

TULP2G(<http://tulp2g.sourceforge.net/>): forensic framework for extracting and decoding data.

Commercial tool:

Lantern: (<http://katanaforensics.com/>): Well-known tool for iPhone, iPod, iPad. New releases support Android devices.

10.3. Appendix 3 – Topics for discussion with the NSIs for the assessment of feasibility in the ESS

Introduction

The project has several objectives related to the employment of modern and enhanced methodologies for producing official statistics from non-traditional data sources such as the Internet or Big Data.

The discussion with a selected group of National Statistical Institutes (NSIs), indicated by Eurostat / Unit G6, will provide input for assessing the feasibility of producing official statistics about the information society based on data obtained with two specific types of measurement:

1. **User-centric:** Automatic recording, with some sort of benevolent monitoring software, of data generated while individuals use the internet with personal devices such as computers, tablets and smartphones.
2. **Enterprise website-centric:** Automatic extraction, with some sort of benevolent web crawler, of data available in the websites of business enterprises about functionalities the websites offer to users and about characteristics of the enterprises (e.g. engagement in e-sales, price lists of products, vacancies, etc.).

Both types of measurement would be used only with the explicit consent of the targeted individuals or enterprises respectively. Moreover, data collected with them could be complemented with data collected with more “traditional” methods (e.g. surveys, data extraction from registers, etc.).

The following list contains the topics to be discussed with the NSIs. It is not a questionnaire but a roadmap of the discussion.

Activities of the NSI in this area

Discussion about statistical production activities of the NSI that involved user-centric or website-centric measurements similar to those described in the introduction. It does not matter whether they are still on-going or whether they are test activities or regular production ones.

1. Description of activities
 - a. Target indicators
 - b. Target population / statistical units
 - c. Collected variables
 - d. Sample design / sampling frame / sample selection procedure / sample size
 - e. Measurement mode → please identify cases where combinations of automatic measurements and traditional survey methods were used
 - f. Response rates

- g. Data processing and data analysis requirements
- h. Hardware and software used
- 2. Additional information
 - a. Reasons for undertaking these activities
 - b. Problems encountered
 - c. Notable experiences
 - d. Effort and cost
 - e. The NSI's / your "verdict" about the activities?
 - f. Are they still on-going?

Opinion about these methods of measurement

It is of interest to have the NSI's opinion about the feasibility and applicability of these methods in the context of the European Statistical System (ESS).

- 1. If there has been no such activity / If there were activities but they have been stopped, why is that?
- 2. Future plans, schedules
- 3. Opinion about the feasibility of the methods of measurement
 - a. Legal barriers foreseen
 - b. Quality of statistics (coverage of target population, coverage of the phenomena intended to be measured, non-response, precision, comparability, relevance of the produced indicators)
 - c. Did you have to deal with or have you thought about issues such as:
 - i. use of the regular individuals' and enterprises' sampling frames in surveys that will use the automatic measurement methods
 - ii. tracking of individual users in the case of multi-device use / multi-user use of the same device
 - iii. the possible association of one enterprise with multiple websites
 - d. Expected degree of acceptance by targeted users and enterprises and by the public in general. Potential to alleviate fears about breach of privacy.
 - e. Technical barriers, relevant competences required.
 - f. Which are the biggest advantages of these methods?
 - g. Which are their greatest problems?
- 4. Comparison with traditional surveys and production methods.
- 5. Likelihood of such methods being adopted for regular statistical production in the ESS.
- 6. Are other organisations in the country engaged in such activities, even if only for research?

12.4. D3 – Results of the testing of the two methods

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’

D3. Results of the testing of the two methods

April 2014

Document Service Data

Type of Document	D3. Results of the testing of the two methods		
Version:	5	Status:	Draft
Created by:	Sonia Chalkidou, Dimitris Kalogeras, Georgia Oikonomopoulou, Thanasis Priftis, Photis Stavropoulos, Alexandra Trampeli	Date:	25/4/2014
Distribution:	European Commission – Eurostat/G6, Agilis S.A., EELLAK		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	31/12/2013	Initial release
2	7/2/2014	Substantial extension with completion of the missing parts of the first version
3	18/2/2014	Modification of Table 5, addition of Table 6, revision of the commentary in section 2.5.3.
4	22/4/2014	Revisions based on comments received on 3/4/2014
5	25/4/2014	Revisions based on comments received on 25/4/2014

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

1. Introduction	4
2. Pilot survey of Internet usage by individuals	5
2.1. Scope of the pilot.....	5
2.1.1. Target population of the pilot survey	5
2.1.2. Statistical indicators produced.....	5
2.2. Sampling procedure	6
2.2.1. Sampling frame	6
2.2.2. Recruitment of sample members	6
2.3. Software tools.....	9
2.4. Implementation of the pilot	11
2.5. Results.....	12
2.5.1. Share of users that have engaged in each type of online activity	12
2.5.2. Share of time online allocated to different activities.....	16
2.5.3. Average amount of time per day, user and type of online activity.....	21
2.5.4. Questionnaire-based statistics on Internet use.....	26
2.6. Conclusions	29
3. Pilot survey of functionalities offered by the websites of business enterprises	32
3.1. Scope of the pilot.....	32
3.1.1. Target population of the pilot survey	32
3.1.2. Statistical indicators produced.....	32
3.2. Sampling procedure	34
3.3. Software tool used in the pilot	34
3.4. Implementation of the pilot	35
3.5. Results.....	37
3.5.1. Availability of contact information on the websites.....	39
3.5.2. Language options of websites	40
3.5.3. Website facilities	42
3.5.4. Other content of the websites	44
3.6. Assessment of the accuracy and specificity of the collected data	46
3.7. Conclusions	49
4. General conclusions	51
5. Annexes.....	52
5.1. Annex 1 – Information note sent to potential members of the sample of the pilot survey of individuals.....	52
5.2. Annex 2 – Screening questionnaire sent to interested potential members of the sample of the pilot survey of individuals.....	55

5.3. Annex 3 – Additional data questionnaire sent to members of the sample of the pilot survey of individuals	59
5.4. Annex 4 – Information note sent to owners of website enterprises, potential members of the sample of the pilot survey.....	64

1. Introduction

Deliverable D1 of the project has proposed a conceptual framework for the production of statistics on the usage of the internet by individuals and enterprises based on automated collection of data from the internet. It has also proposed a number of relevant indicators.

The present report presents the results of the pilot production of a subset of the indicators proposed in D1. It presents both the compiled indicators as well as the way in which the pilot was implemented. Two separate pilots were implemented, one targeting individuals and the other the websites of enterprises. Each pilot is the subject of a separate section: section 2 for individuals and section 3 for enterprises. Section 4 contains the conclusions of the pilot surveys.

2. Pilot survey of Internet usage by individuals

2.1. Scope of the pilot

The aim of the pilot collection was to replicate the current survey on ICT usage in households and by individuals as far as possible, with the following differences:

- the pilot targets usage of the internet only and not of ICT in general. Monitoring of usage of ICT would also be possible but it was outside the scope of the project.
- the pilot targets individuals only and not households. This is due to the way data collection was carried out, which is presented in sections 2.2 and 2.4.
- data collection in the pilot is automated via monitoring software installed on the users' devices; traditional questionnaires are used only for supplementary information.

However, the time and resources devoted to the pilot surveys were limited, due to the constraints of the project. Therefore, some compromises had to be made and focus had to be put on the most important aspects of the pilot, those that differentiate it from the regular ICT survey. This will be made clearer in the sections that follow.

2.1.1. Target population of the pilot survey

The target population of the survey consists of all users of computers, smartphones and tablets connected to the Internet who are resident in private households in Greece. It is therefore a subset of the target population of the regular ICT survey in households. The latter covers all persons, aged 16 or over, who reside in private households.

The restriction of having persons 16 years or over was not imposed explicitly in the pilot survey. It nevertheless applied since the sample was drawn from a panel constructed by a market research company, which consists only of persons aged 19 years or over.

2.1.2. Statistical indicators produced

The sites that users may visit while online have been grouped into approximately 50 categories by the makers of the software that was used in the pilot (see sec. 2.3). Examples of categories are: Educational, Government, Entertainment, Search Portal, News, Sports, Business, etc.

The same categorisation has been used for all activities that a user may perform online. For example, if a smartphone user uses the Youtube app we consider that (s)he is performing a “Viewing / listening to online images, videos, music” activity. The final list of activity types used in the pilot was dictated by the fact that some of the predefined activity categories were not carried out by any member of the sample.

Three indicators have been produced by the pilot survey for these types of activity:

1. Share of users that have engaged in each type of online activity
2. Percentage of time online that users devote on average to specific types of activities.

3. Amount of time that users devote on average per day to specific types of online activities.

All indicators are shown broken down by gender, age and level of education of the user and moreover by whether the day is a working day or not (weekends, holidays).

2.2. Sampling procedure

2.2.1. Sampling frame

The original intention was to draw a sample of households from the sampling frame of ELSTAT (Hellenic Statistics Authority, the Greek NSI), reproducing the stratified sampling scheme followed by the latter. Such a sample cannot be drawn by third parties; it has to be requested and prepared from ELSTAT. During a meeting with the responsible ELSTAT members of staff on 20/9/2013 it turned out that the provision of the sample would require at least one month from the moment a formal request would be submitted to the authority.

Secondly, ELSTAT is forbidden by law to provide to third parties contact details of individual persons or households. The most detailed sample units it can provide are coordinates of building blocks which must then be visited by the third parties for enumeration of households and selection of particular households and individuals.

The project team therefore decided to resort to other means of drawing a sample, even a non-random one. It was felt that the actual selection of the sample, carried out in the same manner as it is done in the regular survey, does not offer any input to the testing of the automated data collection method. The novel features of the method are found in the way it measures data; they can be tested on all kinds of samples.

A first thought was to try selecting the sample with random digit dialling methods. Such methods however produce a very large share of non-existent or non-eligible numbers (e.g. business phone numbers, fax machines, etc.) and require a lot of effort to check the numbers and recruit their owners to the survey. Due to the sensitive nature of the pilot survey (automatic recording of activities online) it was expected to have a very large rate of refusals, which would increase further the required time for completion of the pilot.

In the end, the project team chose as sampling frame a panel of persons compiled by a Greek market research company for use in opinion surveys. The characteristics of this panel are shown in section 2.2.2.

2.2.2. Recruitment of sample members

The panel comprises 1287 persons from the whole of Greece. Due to its small size and to the expected high rate of refusals to participate there was no random selection of sample members.

The market research company considered the provision of a monetary incentive to users as paramount to soliciting their cooperation. The reward for each participating member was €30.00. Due to this cost, as well as the cost of the monitoring software it was decided to restrict the sample to 150 persons and devices at most. As will be shown later however, due to a very low rate of cooperation, the final sample consisted of only 48 persons.

In order to attract more users we proposed to participants a “certificate of participation”, signed from EELLAK, stating their involvement to the pilot. Although, it was clearly stated that this is not a formal certification, certain younger participants responded positively to this.

As first contact a note was sent to all members of the panel informing them about the nature of the data collection, the anonymity of the data and the indicators that would be produced. The note, translated in English, is shown in section 5.1 in the annex of the report. It does not mention the financial incentive; this was mentioned in the email with which the note was sent. Together with the note the members of the panel received a screening questionnaire that asked about the types and number of devices which they use to access the Internet. The questionnaire is presented in section 5.2 in the annex of the report. Three reminders were sent and 145 persons in total accepted to participate.

In their replies to the screening questionnaire the individuals stated whether they use a personal computer, a smartphone or a tablet; moreover they indicated which of their smartphone or tablet they use more often for accessing the Internet. The project team selected for each user one device at random for installation of the software. The random choice was always made between the PC and the most frequently used of the mobile devices.

The users received the instructions for installing the software to their selected device (see section 2.4). In the end however, due to the difficulties in installing the software, or due to second thought perhaps, we managed to enlist only 48 persons in the sample. Due to the inability to install the software on iOS devices (see section 2.3) the random selection of devices led to the software being installed on 35 PCs, 12 Android smartphones and one Android tablet.

The characteristics of the panel, of the 97 persons that initially accepted to participate in the pilot but then withdrew and of the final sample of 48 persons are shown in the following table.

Table 1. Demographic characteristics of the sample, of the persons that finally refused to participate and of the complete panel.

Characteristic	Sample members N (%)	Persons who initially accepted but in the end refused to participate N(%)	Members of the panel N (%)
Gender			
Males	18 (38)	43 (44)	639 (50)
Females	30 (63)	54 (56)	648 (50)
Age			
<25	14 (29)	12 (12)	222 (17)
25-44	30 (63)	64 (66)	911 (71)
45+	4 (8)	21 (21)	154(12)

Characteristic	Sample members N (%)	Persons who initially accepted but in the end refused to participate N(%)	Members of the panel N (%)
Education (ISCED-11)			
ISCED 1	-	-	10 (1)
ISCED 2-4 ¹	20 (42)	52 (54)	576 (45)
ISCED 5A-5B	22 (46)	32 (33)	574 (45)
ISCED 5A-6 ²	6 (13)	13 (13)	127 (10)
Monthly Income (€)			
<1000	26 (53)	47 (48)	668 (52)
1001-2500	20 (41)	41 (41)	482 (37)
2501-5000	1 (2)	6 (6)	89 (7)
5001-10000	-	1 (1)	16 (1)
>10000	1 (2)	3 (3)	32 (2)
Occupation			
Employee	16 (33)	37 (38)	512 (40)
Self-employed	10 (21)	27 (28)	293 (23)
Student	15(31)	12 (12)	186 (14)
Unemployed	6 (13)	19 (20)	242 (19)
Other	1 (2)	2 (2)	54 (4)
Region (NUTS 2)			
Anatoliki Makedonia, Thraki (EL 11)	-	2 (2)	47 (4)
Kentriki Makedonia (EL 12)	32 (67)	44 (45)	445 (35)
Dytiki Makedonia (EL 13)	1 (2)	1 (1)	16 (1)
Thessalia (EL 14)	-	2 (2)	38 (3)

Characteristic	Sample members N (%)	Persons who initially accepted but in the end refused to participate N(%)	Members of the panel N (%)
Ipeiros (EL 21)	2 (4)	3 (3)	34 (3)
Ionia Nisia (EL 22)	-	1 (1)	13 (1)
Dytiki Ellada (EL 23)	1 (2)	1 (1)	42 (3)
Sterea Ellada (EL 24)	-	2 (2)	34 (3)
Peloponnisos (EL 25)	1 (2)	3 (3)	37 (3)
Attiki (EL 30)	10 (21)	35 (36)	491 (38)
Voreio Aigaio (EL 41)	1 (2)	1 (1)	16 (1)
Notio Aigaio (EL 42)	-	-	24 (2)
Kriti (EL 43)	-	2 (2)	50 (4)

1 The level of educational attainment of the members of the panel has been recorded, by the market research company, in a way that does not allow to separate ISCED level 4 from levels 2 and 3.

2 ISCED 5A has been split into post-graduate degrees (e.g. Masters) in this cell and tertiary non-postgraduate degrees in the cell above.

The comparison between the complete panel and the sample shows some issues. Women, young persons and students are over-represented in the sample, compared to the panel. The over-representation of the two latter categories is not surprising, given their greater familiarity with software tools and online interaction between persons. Education levels and income classes are quite fairly represented in the sample. Finally, there is a large over-representation of central Macedonia (EL 12) probably due to the fact that the market research company is located in Salonika, in central Macedonia.

2.3. Software tools

The software selected for monitoring and recording the users' activities was the online parental controls service Qustodio¹. We first explain its regular usage for monitoring children's activities online as this helps explain its deployment in the pilot (see section 2.4).

A parent signs in to the service and defines one "child" for each combination of child, device and user account on the device. For example, if a household has a desktop and a laptop PC and the two children have one account each on each device, their father will define four "children" in the service. This enables individualised monitoring. The parent uses separate passwords per "child" to activate monitoring. A second form of usage, suitable for schools that want to control activity over the school's computers, is to have a single administrator user name and password. The administrator then defines "children" for all the devices.

¹ www.qustodio.com

Qustodio records the time users spend on their device, on each specific website and while using specific applications. It reports to the user via a webpage the following information:

- Usage time: total and also “web time”, “social activity time” and “apps time”.
 - Total time denotes all time that the computer is active, even if not on the Internet.
 - Web time records only the time spent visiting websites; usage of Youtube app on a mobile device for example is not counted as web time. Web activity is computed as one minute per connection. If for example an open tab makes connections, each connection counts as one minute.
 - Social activity corresponds to Facebook activities only, e.g. chatting with friends and is not a subset of web activity, although visits to the Facebook page also count as web activity.
 - Apps time is the time spent using applications, even offline. In general, after 5 minutes of idle time of computer use, the usage of apps and websites stops counting.
- Share of total time spent per category of website. The following generic categories are recognised: Educational, Government, Entertainment, Search Portal, News, Sports, Business, Health, Technology, Games, Travel, Religion, Shopping, Employment, Webmail, Forums, Social Network, Chat, File Sharing, Gambling, Loopholes, Violence, Weapons, Profanity, Mature Content, Pornography, Alcohol, Drugs Tobacco.
- Share of total time spent per application. Individual applications are reported.

For each account Qustodio provides a separate web page with aggregated statistics. The aggregation period can be modified, spanning one, seven, 15 or 30 days. An indicative extract of results is shown in Box 1.

Box 1. Extract of Qustodio results for a single “child” and a single 24-hour period.

1181_pc.htm ← This is the “child”
 62.5% Using Microsoft Office Word Using Microsoft Office Word
 18.8% Surf on Search Portal websites Surf on Search Portal websites
 9.4% Surf on Social Network websites Surf on Social Network websites
 3.1% Surf on Shopping websites Surf on Shopping websites
 3.1% Surf on Webmail websites Surf on Webmail websites
 3.1% Using DUPLEX Using DUPLEX

 0:59 Total usage time during specified period
 0:08 Hours of Web activity
 0:00 Hours of Social activity²
 0:56 Hours of Apps usage

² We remind the reader that “Social activity” for Qustodio is only the interaction with friends in Facebook.

A look at the extract of results shows discrepancies between the reported shares of total usage time and the aggregated usage times reported in hours and minutes. For example, two applications, Microsoft Word and Duplex, are used for 65.6% at most³ of total usage time, but on the other hand total app usage is reported as 56 minutes, i.e. 95% of the total usage time. Due to such discrepancies and due to the more detail provided with times reported as shares, the analysis was based only on the shares and on total usage times.

An additional drawback, which however affected all available software options, was that Qustodio does not run on iOS devices (iphones and ipads). Therefore the coverage of the population suffers.

2.4. Implementation of the pilot

Two options were available for the installation of the software to the devices of the sample members:

1. Installation by the sample members themselves: detailed instructions were sent to each of the interested users indicating the device on which they had to install Qustodio and the way of installing it. They were instructed to choose a specific “child” name so that the project team could also access the results of monitoring. This approach had disappointing results because less than 20 people managed to complete the process successfully.
2. Generation of accounts on the users’ behalf: the project team generated accounts and emailed users about the device to be monitored and the chosen child name along with less instructions about the installation. This extra step increased the participation to 48 members, as this approach did not oblige them to log into the Qustodio web control panel. They just had to do the software installation after their account had been created.

It goes without saying that under both options the users knew that their activity would be reported to the project team for the production of statistical results.

A liaison to the project team was appointed by the market research company. This person, with the assistance of technical staff from EELLAK provided support to the members of the sample. The provided support was intense throughout the pilot. Personal emails were sent to each member. All sent emails were personally signed, so the recipients knew with whom they were communicating. This increased their confidence in the procedure itself. Many participants sent emails asking various questions. They all had their questions answered, usually immediately. The email signature contained contact telephone and the recipients were encouraged to call the liaison whenever they had anything to ask. After every telephone call a separate log was created, with the subject of the communication and contact details, in order to be able to catch up with them later and document their requests.

The pilot took place in December 2013. The first 10 days were spent deploying the software to the sample members. The remaining days were spent on collecting usage data and it is on these data that the statistics presented in the following section are based.

Finally, during the course of the collection, users were sent an email questionnaire requesting some demographic data and also some Internet usage data. These data were combined with those collected by Qustodio. The questionnaire is shown in section 5.3 in the annex of the report.

³ It could be less than that if the two applications were running in parallel.

2.5. Results

This section presents two sets of results. The first one contains results on the shares of users and of usage time that correspond to each type of activity. They are shown in sections 2.5.1 - 2.5.3. The second, smaller set, in section 2.5.4, presents data on the use of Internet collected with the help of the questionnaire shown in the annex of section 5.3. These data are juxtaposed with relevant statistics produced by the regular ICT survey.

Data on users and usage time

The data of the form shown in Box 1, earlier, were processed by a PHP script and were converted in a tabular format with one row per “child” and with the date, the usage time and shares data in columns. This form greatly facilitates their processing by database systems or statistical packages.

The data list types of websites and, separately, the apps used by each user. Since some of the apps also represent activity online (e.g. the Youtube app on a smartphone or tablet) they had to be allocated to the categories of online activity too. This categorisation was made by members of the project team, with an examination of each individual app listed in the raw data produced by the tool.

Some apps were too “obscure” to identify even after searches in Google. Therefore we have added category “Not clear” to the activities, because for these app it is not even clear whether they represent online or offline activity. Moreover, Qustodio itself sometimes reports shares allocated simply to “Internet”. This category is shown as “Internet – unspecified” in the statistics presented here. Finally, we use category “Internet – other” to group together categories of online activities with very small usage times recorded for very few users.

Note: The shares of time allocated to each app were added for all apps falling into a single category. This creates problems of double counting because two apps may have been used in parallel; this cannot be observed in the recorded data. Therefore, the aggregated usage times and shares of time are larger than or equal to the actual times.

2.5.1. Share of users that have engaged in each type of online activity

Table 2 shows the shares of users that engaged, even once, on each type of online activity over the monitoring period.

The most popular activities with users are social networks (95.8%), shopping websites (93.8%), entertainment or technology websites (89.6% each) and emailing (77.1%). All these activities are even more popular among women, except for shopping websites which is slightly more popular with men.

Differences between the two sexes are not large. However, some types of activity, namely usage of online storage facilities (cloud), visits to government websites, listening to web radio, attendance of online courses, online not-networked games are carried out only by men.

More differences are observed between age classes. Persons 25 years old or younger engage much more in gambling or gaming than persons aged between 25 and 44. The latter on the other hand engage more in reading news or visiting sports-related websites. Comparisons with persons older than 44 years cannot be made: the reliability of data for them is very small however because the sample contains only four such persons.

There are some pronounced differences between levels of education too. A larger share of persons with postgraduate education engage in visits to education sites or in participation to online forums. Only such persons visit government websites while on the other hand they do not visit gambling websites. Persons without tertiary education engage in larger percentages in gaming.

Finally, for most activities a larger share of people engage in them during working days of the week rather than on non-working days. For the most part the differences are not great.

Table 2. Share of users (%) that engage in each type of online activity¹.

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ²	ISCED 2-3	ISCED 5A	ISCED 5A-6 ³			
Cloud services	5.6			3.3			4.5			2.1	2.1
Education unspecified	33.3	23.3	21.4	30.0	25.0	26.3	18.2	57.1	24.4	16.7	27.1
Email	61.1	86.7	78.6	83.3	25.0	84.2	72.7	71.4	61.0	70.8	77.1
Employment	5.6	6.7	7.1	6.7		10.5		14.3	7.3		6.3
Entertainment	83.3	93.3	100.0	83.3	100.0	100.0	77.3	100.0	85.4	87.5	89.6
Finding information	11.1	3.3	7.1	6.7		5.3	9.1		4.9	6.3	6.3
Forums	33.3	33.3	28.6	30.0	75.0	31.6	27.3	57.1	22.0	29.2	33.3
Gambling	27.8	23.3	35.7	16.7	50.0	26.3	31.8		24.4	18.8	25.0
Games unspecified	38.9	46.7	57.1	33.3	75.0	68.4	31.8	14.3	26.8	39.6	43.8
Government	5.6			3.3				14.3		2.1	2.1
Internet other	33.3	33.3	42.9	26.7	50.0	42.1	27.3	28.6	26.8	18.8	33.3
Internet unspecified	33.3	20.0	28.6	26.7		21.1	31.8	14.3	29.3	25.0	25.0
Listening to web radio	5.6			3.3			4.5			2.1	2.1
Networked Games	16.7	6.7	14.3	6.7	25.0	21.1	4.5		9.8	8.3	10.4
Not clear	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.6 ⁴	97.9	100.0
Attending online courses	5.6			3.3			4.5			2.1	2.1
Online not networked games	11.1		7.1	3.3		5.3	4.5		2.4	2.1	4.2

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ²	ISCED 2-3	ISCED 5A	ISCED 5A-6 ³			
Pornography	27.8	13.3	14.3	20.0	25.0	10.5	18.2	42.9	14.6	12.5	18.8
Reading News	50.0	43.3	35.7	50.0	50.0	42.1	45.5	57.1	31.7	35.4	45.8
Shopping	94.4	93.3	92.9	93.3	100.0	100.0	90.9	85.7	78.0	89.6	93.8
Social networks	88.9	100.0	100.0	93.3	100.0	100.0	90.9	100.0	92.7	95.8	95.8
Sports	38.9	20.0	7.1	33.3	50.0	26.3	27.3	28.6	22.0	20.8	27.1
Technology	88.9	90.0	92.9	86.7	100.0	100.0	81.8	85.7	78.0	85.4	89.6
Telephony	50.0	43.3	50.0	46.7	25.0	36.8	59.1	28.6	41.5	43.8	45.8
Travel	5.6	20.0	21.4	10.0	25.0	15.8	13.6	14.3	7.3	10.4	14.6
Viewing / listening to online films / music	33.3	3.3	21.4	13.3		10.5	18.2	14.3	9.8	14.6	14.6

¹ All figures are rounded to one decimal digit. Therefore, 0.0 denotes values less than 0.05%.

² The sample contains only four persons aged 45+; therefore the results for them are unreliable.

³ ISCED 5A has been split into post-graduate degrees (e.g. Masters) in this column and tertiary non-postgraduate degrees in the column to its left.

⁴ Although all users have at some point engaged in “not clear” activities, some have done so only during working days and others only during the weekend, therefore the shares per type of day are less than 100%.

2.5.2. Share of time online allocated to different activities

This and the following section show statistics on the time spent on average per activity. The statistics have been computed for two user groups:

- All users
- Active users: in these statistics, average usage time is computed only over those users that have carried out the relevant online activity even once over the period of data collection. In other words, users which never carried out this type of activity during the monitoring period are not counted in the averages.

Table 3 shows the share of time that each user on average allocates to different activities. Then, Table 4 shows the similar shares only for active users.

Social networks are the most popular activity for the average user, occupying 11.1% of his time. They are followed by visits to entertainment websites (6.7% of online time) and visits to technology websites (4.9%) of time.

Focusing on the active users of each activity, we see that watching films or listening to music online is the activity that draws most the attention of its users (14.1% of their time). Social networks and visits to entertainment websites have shares (11.3% and 7.3%) close to those for the average user, indicating that their use is spread over the whole user population.

Table 3. Share of online time (%) that users devote on average to each type of online activity¹. All users taken into account.

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ²	ISCED 2-3	ISCED 5A	ISCED 5A-6 ³			
Cloud services	0.0			0.0			0.0		0.0	0.0	0.0
Education unspecified	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.3	0.1	0.1	0.1
Email	2.3	3.4	1.0	3.5	2.9	3.2	2.7	2.9	1.0	4.0	2.9
Employment	0.0	0.0	0.0	0.0		0.0		0.0	0.0	0.0	0.0
Entertainment	7.1	6.4	8.6	5.3	9.9	6.4	6.6	7.6	7.0	6.6	6.7
Finding information	0.2	0.1	0.0	0.2		0.0	0.3		0.1	0.2	0.1
Forums	0.8	0.3	0.5	0.6	0.1	0.2	0.3	1.3	0.9	0.3	0.5
Gambling	2.8	0.2	1.6	1.5	0.4	0.2	2.8		1.7	1.1	1.3
Games unspecified	0.3	1.4	3.3	0.2	0.8	2.1	0.2	0.1	1.0	0.9	0.9
Government	0.2			0.1				0.4	0.0	0.1	0.1
Internet other	0.2	6.2	1.0	5.1	1.2	9.1	0.5	0.2	3.1	4.0	3.7
Internet unspecified	2.7	0.6	0.9	2.0		0.3	2.9	0.4	1.2	1.6	1.5
Listening to web radio	0.0			0.0			0.0		0.0	0.0	0.0
Networked Games	1.3	0.3	2.9	0.0	1.1	2.0	0.0		0.5	0.9	0.8
Not clear	13.0	12.0	12.2	12.5	12.4	8.6	17.0	9.6	13.4	11.9	12.4
Attending online courses	0.0			0.0			0.0		0.0	0.0	0.0
Online not networked games	0.1		0.0	0.1		0.0	0.1		0.2	0.0	0.1

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ²	ISCED 2-3	ISCED 5A	ISCED 5A-6 ³			
Pornography	0.2	0.0	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.1	0.1
Reading News	2.6	0.7	0.1	2.3	0.2	0.7	2.4	1.2	1.5	1.6	1.6
Shopping	4.3	3.4	3.9	3.6	4.2	3.3	4.7	2.5	4.4	3.4	3.8
Social networks	9.5	12.3	15.8	10.2	9.0	9.0	13.9	8.9	12.2	10.5	11.1
Sports	0.9	0.9	0.9	1.0	0.2	1.4	0.4	1.0	1.5	0.5	0.9
Technology	4.7	5.0	3.5	5.4	4.2	4.0	4.9	6.4	4.4	5.1	4.9
Telephony	4.8	1.0	3.5	2.8	0.8	1.0	2.7	5.4	3.1	2.3	2.6
Travel	0.0	0.1	0.2	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.1
Viewing / listening to online films / music	4.9	0.0	3.6	2.2		1.8	1.9	3.2	2.2	2.0	2.1

¹ All figures are rounded to one decimal digit. Therefore, 0.0 denotes values less than 0.05%.

² The sample contains only four persons aged 45+; therefore the results for them are unreliable.

³ ISCED 5A has been split into post-graduate degrees (e.g. Masters) in this column and tertiary non-postgraduate degrees in the column to its left.

Table 4. Share of online time (%) that users devote on average to each type of online activity¹. Only active² users of each activity taken into account.

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ³	ISCED 2-3	ISCED 5A	ISCED 5A-6 ⁴			
Cloud services	0.0			0.0			0.0			0.0	0.0
Education unspecified	0.4	0.3	0.1	0.4	1.9	0.3	0.4	0.4	0.4	0.7	0.4
Email	4.8	3.7	1.6	4.5	4.8	3.3	4.8	4.8	1.9	5.6	4.0
Employment	0.1	0.3	0.1	0.3		0.2		0.2	0.5		0.2
Entertainment	7.8	7.0	8.6	6.1	9.9	6.4	8.2	7.6	7.8	7.3	7.3
Finding information	1.8	2.4	0.4	2.7		0.4	2.7		2.3	2.2	2.1
Forums	1.4	0.7	1.2	1.6	0.1	0.7	0.8	1.8	2.9	0.7	1.1
Gambling	6.0	0.7	3.2	7.5	0.5	0.7	5.9		4.4	5.1	3.7
Games unspecified	1.1	4.5	7.3	1.0	2.0	5.5	0.8	0.9	4.0	3.5	3.1
Government	1.2			1.2				1.2		2.1	1.2
Internet other	0.7	11.0	1.9	14.1	1.5	12.2	2.0	0.5	7.2	13.8	7.9
Internet unspecified	11.5	3.7	2.6	10.8		2.0	10.9	8.8	6.7	8.5	7.9
Listening to web radio	0.1			0.1			0.1			0.1	0.1
Networked Games	10.8	2.0	16.6	0.6	1.9	5.6	0.0		3.1	7.0	5.0
Not clear	13.0	12.0	12.2	12.5	12.4	8.6	17.0	9.6	13.7	12.1	12.4
Attending online courses	0.2			0.2			0.2			0.2	0.2
Online not networked games	1.9		0.3	3.2		0.3	3.2		8.2	0.7	1.9

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ³	ISCED 2-3	ISCED 5A	ISCED 5A-6 ⁴			
Pornography	0.5	0.3	0.7	0.4	0.7	0.3	0.4	0.6	0.9	0.5	0.4
Reading News	4.6	1.6	0.2	3.8	1.0	1.7	4.9	1.8	3.9	4.0	3.1
Shopping	4.3	3.4	4.1	3.6	4.2	3.3	4.8	2.6	5.6	3.6	3.8
Social networks	9.9	12.3	15.8	10.5	9.0	9.0	14.6	8.9	12.6	10.7	11.3
Sports	1.7	5.4	9.3	2.6	0.9	6.3	1.1	2.6	5.8	2.3	2.8
Technology	4.9	5.3	3.7	5.8	4.2	4.0	5.4	6.7	5.0	5.6	5.1
Telephony	7.2	2.4	5.6	4.9	4.3	3.3	4.0	9.6	7.2	4.7	5.0
Travel	0.0	0.5	3.6	0.4	0.0	0.2	0.5	0.7	2.5	0.3	0.3
Viewing / listening to online films / music	14.2	11.0	19.2	12.5		19.9	16.7	9.3	18.4	14.4	14.1

¹ All figures are rounded to one decimal digit. Therefore, 0.0 denotes values less than 0.05%.

² Please refer to the beginning of section 2.5.2 for the definition of “active user”.

³ The sample contains only four persons aged 45+; therefore the results for them are unreliable.

⁴ ISCED 5A has been split into post-graduate degrees (e.g. Masters) in this column and tertiary non-postgraduate degrees in the column to its left.

2.5.3. Average amount of time per day, user and type of online activity

Table 5 shows the average amount of time that each user on average allocates to different activities.

The amounts of time devoted per day by the average user to activities correlate, as expected, with the shares of online time shown earlier. The most popular activity is participation to social networks with 22.1 minutes per day on average. Other very popular activities are visits to entertainment websites (13.3 minutes) and visits to technology websites (9.7 minutes). Social networks attract the largest amount of time from users of all categories.

Similar averages only for active users, defined in the manner explained in section 2.5.2 are shown in Table 6. Viewing or listening to online media and participation in social networks are now most popular with 36.9 and 36.1 minutes per day.

Table 5. Amount of time (minutes) that users devote on average per day to each type of online activity¹. All users taken into account.

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ²	ISCED 2-3	ISCED 5A	ISCED 5A-6 ³			
Cloud services	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Education unspecified	0.1	0.1	0.0	0.2	0.0	0.0	0.1	0.1	0.1	0.1	0.2
Email	2.0	3.9	0.4	4.5	0.9	2.4	2.3	1.2	0.7	5.1	5.8
Employment	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Entertainment	6.0	7.3	3.5	6.7	3.3	4.8	5.6	3.1	4.9	8.4	13.3
Finding information	0.1	0.1	0.0	0.3	0.0	0.0	0.3	0.0	0.1	0.2	0.3
Forums	0.7	0.3	0.2	0.8	0.0	0.2	0.3	0.5	0.6	0.4	1.0
Gambling	2.4	0.2	0.6	1.9	0.1	0.2	2.4	0.0	1.2	1.4	2.6
Games unspecified	0.3	1.6	1.3	0.3	0.3	1.6	0.2	0.0	0.7	1.1	1.8
Government	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.2	0.2
Internet other	0.2	7.1	0.4	6.5	0.4	6.8	0.5	0.1	2.2	5.1	7.3
Internet unspecified	2.3	0.7	0.4	2.6	0.0	0.2	2.5	0.2	0.9	2.1	2.9
Listening to web radio	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Networked Games	1.1	0.4	1.2	0.0	0.4	1.5	0.0	0.0	0.4	1.2	1.5
Not clear	11.0	13.7	5.0	15.9	4.1	6.4	14.6	3.9	9.5	15.3	24.7

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ²	ISCED 2-3	ISCED 5A	ISCED 5A-6 ³			
Attending online courses	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Online not networked games	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.1
Pornography	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.2
Reading News	2.2	0.9	0.0	3.0	0.1	0.5	2.1	0.5	1.1	2.1	3.1
Shopping	3.6	3.8	1.6	4.6	1.4	2.5	4.1	1.0	3.1	4.4	7.5
Social networks	8.0	14.1	6.4	13.0	3.0	6.7	12.0	3.6	8.6	13.5	22.1
Sports	0.8	1.0	0.4	1.3	0.1	1.0	0.3	0.4	1.1	0.7	1.7
Technology	4.0	5.7	1.4	6.9	1.4	3.0	4.2	2.6	3.1	6.6	9.7
Telephony	4.0	1.1	1.4	3.5	0.3	0.7	2.4	2.2	2.2	3.0	5.2
Travel	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.1	0.1
Viewing / listening to online films / music	4.1	0.1	1.5	2.8	0.0	1.3	1.6	1.3	1.6	2.6	4.2

¹ All figures are rounded to one decimal digit. Therefore, 0.0 denotes values less than 0.05 minutes.

² The sample contains only four persons aged 45+; therefore the results for them are unreliable.

³ ISCED 5A has been split into post-graduate degrees (e.g. Masters) in this column and tertiary non-postgraduate degrees in the column to its left.

Table 6. Amount of time (minutes) that users devote on average per day to each type of online activity¹. Only active² users of each activity taken into account.

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ³	ISCED 2-3	ISCED 5A	ISCED 5A-6 ⁴			
Cloud services	0.1			0.1			0.1			0.1	0.1
Education unspecified	1.4	0.7	0.2	1.3	1.4	0.4	1.0	1.7	1.0	1.9	1.0
Email	11.7	12.3	3.2	13.4	45.1	11.7	10.9	17.3	5.3	17.6	12.1
Employment	0.1	0.5	0.1	0.5		0.4		0.3	0.7		0.3
Entertainment	26.7	22.4	20.7	21.1	46.4	21.0	24.9	30.0	25.5	24.6	24.2
Finding information	3.6	7.0	0.7	6.9		0.7	6.9		4.6	5.8	4.8
Forums	6.3	2.7	4.1	6.4	0.4	2.9	2.9	8.3	11.7	2.9	4.4
Gambling	26.4	2.9	9.5	32.0	2.8	3.0	23.6		19.7	17.1	15.9
Games unspecified	2.8	10.4	15.4	2.7	5.4	11.1	2.6	1.2	11.2	8.1	7.5
Government	8.1			8.1				8.1		13.4	8.1
Internet other	1.8	49.2	4.8	48.9	10.4	60.6	4.2	2.0	27.6	61.9	29.0
Internet unspecified	21.3	6.4	5.6	17.6		3.5	20.0	11.9	10.4	16.4	14.0
Listening to web radio	0.1			0.1			0.1			0.2	0.1
Networked Games	23.4	12.7	35.1	0.9	17.7	24.1	0.1		12.3	30.6	19.2
Not clear	41.4	37.5	29.4	39.8	58.1	28.2	47.8	37.9	41.3	39.6	39.2
Attending online courses	0.3			0.3			0.3			0.5	0.3
Online not networked games	4.5		0.8	7.3		0.8	7.3		16.7	1.7	4.5

	Sex		Age			Education level			Non-working day	Working day	Total
	Males	Females	<25	25-44	45+ ³	ISCED 2-3	ISCED 5A	ISCED 5A-6 ⁴			
Pornography	2.0	0.9	0.6	1.7	2.0	1.1	1.3	2.4	3.7	1.4	1.6
Reading News	15.6	5.3	0.5	14.1	2.5	4.7	14.8	9.1	13.9	13.1	10.1
Shopping	13.8	11.3	10.2	11.9	19.6	10.8	13.9	11.3	16.0	12.2	12.4
Social networks	32.5	38.5	38.1	34.1	42.1	29.5	41.7	34.8	39.3	35.0	36.1
Sports	6.3	15.1	30.9	9.6	2.0	15.9	3.6	12.9	20.4	7.1	9.4
Technology	16.2	17.7	9.3	19.8	19.5	13.3	16.4	29.2	17.5	19.3	17.0
Telephony	25.3	5.8	13.9	14.7	16.7	7.0	11.1	57.5	19.5	13.9	14.6
Travel	0.1	1.6	2.2	1.2	0.2	1.1	1.7	0.9	2.7	1.8	1.3
Viewing / listening to online films / music	37.8	13.9	38.8	36.0		41.9	26.6	61.0	71.1	38.9	36.9

¹ All figures are rounded to one decimal digit. Therefore, 0.0 denotes values less than 0.05 minutes.

² Please refer to the beginning of section 2.5.2 for the definition of “active user”.

³ The sample contains only four persons aged 45+; therefore the results for them are unreliable.

⁴ ISCED 5A has been split into post-graduate degrees (e.g. Masters) in this column and tertiary non-postgraduate degrees in the column to its left.

2.5.4. Questionnaire-based statistics on Internet use

The first two tables present statistics on the shares of users that access the Internet via devices other than PCs, smartphones or tablets. The share computed from ICT survey data is only 1% for the totality of such devices. The pilot survey's sample shows much higher percentages, possibly due to its having been drawn from an Internet panel.

Table 7. Devices other than PC, smartphone or tablet, used for Internet access.

Device	Number of users	Share of users (%) *
Gaming (games console)	8	17
Reader electronic book (e-book reader)	6	13
Music player, movies or multimedia (media player)	17	35
TV connected to the Internet (Smart TV)	7	17

*The respondents had the opportunity to mark more than one option.

Table 8. European statistics on devices used in Greece for connection to the Internet (ICT survey).

Indicator	2011	2012	2013
Share of individuals (%) using another handled device (e.g. PDA, MP3 player, e-book reader, handled games console, excluding tablet computer) to access Internet	:	1*	:

* Unreliable data

(Eurostat source: isoc_cimobi_dev)

The next two tables show the frequency of use of computers in Greece. All pilot sample members use computers daily or almost daily, while according to the ICT survey, around three quarters of individuals did so during the last three years. The highest frequency of use among the sample's members again shows that being drawn from an online panel they differ from the average Greek person in terms of computer use.

Table 9. Frequency of computer use within the last 3 months (mid-September to mid-December).

Frequency	Number of users	Share of users(%)
Every day or almost every day	48	100
At least once a week but not every day	-	-
Less than once a week	-	-

Table 10. European statistics on frequency of computer use in Greece during the last three months (ICT survey).

Share of individuals (%) using computers ...	2011	2012	2013
daily	73	75	78
at least once a week (but not every day)	19	16	15
once a week or less frequently	8	9	7

(Eurostat source: isoc_ci_cfp_fu)

The shares of individuals that used the websites of public authorities in order to send completed forms or to obtain applications, certificates, etc. (Table 11), are comparable with those computed by the ICT survey (Table 12). However, the use of e-government sites for downloading of information is much more prevalent among the members of the pilot sample.

Table 11. Reasons of Internet use while dealing with public services during 2013.

Reason	Number of users	Share of users (%) [*]
To send completed forms	17	35
To download information	35	73
To obtain applications, certificates, etc.	20	42
No transactions with public services throw Internet	10	21

^{*}The respondents had the opportunity to mark more than one option. Therefore the sum of shares may exceed 100%.

Table 12. European statistics on the e-Government activities of individuals via websites in Greece (ICT survey).

Share of individuals (%) [*] who used Internet to ...	2011	2012	2013
send completed forms	24	32	32
download official forms	28	31	31
obtain applications, certificates, etc.	42	51	52

(Eurostat source: isoc_ciegi_ac)

^{*}The respondents had the opportunity to mark more than one option. Therefore the sum of shares may exceed 100%.

The members of the sample (Table 13) have made their last purchase over the Internet more recently than Greek Internet users in general (Table 14). 63% made it during the last three months, while the corresponding percentage according to the last three ICT surveys ranges between 24% and 28%.

Table 13. Time period of the most recent online purchase of goods or services for personal use.

Time	Number of persons	Share of users (%)
Last 3 months	30	63
Between 3 and 12 months ago	5	10
More than a year ago	3	6
No transactions	10	21

Table 14. European statistics on time period of the most recent online purchase of goods or services for personal use in Greece (ICT survey).

Share of individuals (%) who used the Internet during the last year, which ...	2011	2012	2013
made their last online purchase in the last 3 months	24	28	28
made their last online purchase between 3 and 12 months ago	9	8	12

(Eurostat source: isoc_ec_ibuy)

The same pattern of more intensive Internet use by the pilot sample rather than the population of Greek Internet users in general is evident in the online purchases made by the two groups. At least 68% of the sample purchased some kind of item online during 2013 (Table 15). The corresponding share from the ICT survey is not available (see footnote of Table 16); the goods category most purchased were clothes and sports goods (not shown), purchased by 36% of users.

Table 15. Items purchased via Internet for personal use during 2013 (excluding orders or purchases through e-mail).

Items	Number of users who made online purchases during 2013	Share (%) [*]
Movies, music	4	11
Electronic books, magazines, newspapers or training material (e-learning)	7	18
Software, including computer games	9	24
Other	26	68

^{*} The respondents had the opportunity to mark more than one option. Therefore the sum of shares may exceed 100%.

Table 16. European statistics on items purchased via Internet for personal use in Greece (ICT survey).

Share of individuals who made online purchases (%) [*] , which purchased ¹ ...	2011	2012	2013
films/music	10	10	9
books/magazines/e-learning material	20	20	17
computer software	19	15	14

(Eurostat source: isoc_ec_ibuy)

¹ Category 'other' is not shown because its share is neither reported nor can it be computed due to its comprising types of goods purchased by overlapping groups of users.

Finally, among users that purchased goods online, between 17% and 23% of them had purchased items downloaded from the Internet instead of receiving them via the post (Table 17). The corresponding shares from the ICT survey (Table 18, next page), which refer to 2011, the latest year for which this indicator is available, are up to 5%.

Table 17. Items purchased via Internet for personal use and acquired by downloading through websites during 2013.

	Number of users who had online purchases during 2013 downloaded from the Internet	Share (%) [*] of users who made online purchases in 2013
Movies, music	8	23
Electronic books, magazines, newspapers or training material (e-learning)	9	26
Software, including computer games	9	26
Other	6	17

^{*}The respondents had the opportunity to mark more than one option. Therefore the percentages are not summed in 100%.

2.6. Conclusions

Although it has not been possible to replicate all activities that an NSI would undertake, the results of the pilot study of individuals have shown the potential of the automated recording of data.

The types of online activities of individuals can be discerned at great detail and therefore rich classifications can emerge for statistical use. Moreover, the classifications can change to fit evolving statistical needs. Even historical data can be converted easily to the new classifications.

Table 18. European statistics on items purchased via and downloaded from the Internet for personal use in Greece (ICT survey).

Share of individuals who made online purchases (%) [*] , which downloaded the purchased ¹ ...	2011 ²
films/music	Not. av.
books/magazines/e-learning material	2
computer software	5

(Eurostat source: isoc_ec_ibuy)

1 Category 'other' is not shown because its share is neither reported nor can it be computed due to its comprising types of goods purchased by overlapping groups of users.

2 Data for 2012 and 2013 not available at the time of writing this report.

The variations of usage time can be observed and reported to the desired degree of temporal detail. One can easily imagine charts showing the evolution of usage time or of the share of users engaging in a specific activity for any category of users recorded and over any period of time. Similarly the variation can be shown by day of the week (i.e. "average Monday", "average Tuesday", etc) or by hour of the day.

The data are also recorded with great accuracy since there is no intervention of the individuals' cognitive processes. Reduced recall of past activities, which is a common problem in questionnaire-based surveys, does not affect the measurements.

Moreover, the measurements are obtained with great speed, irrespective of the size of the sample. The initial set-up of the software can be implemented in parallel for all or almost all sample members. Subsequently the software operates independently on each device and therefore the procedure is easily scalable to larger samples.

The speed of data collection also allows the repetition of the survey more frequently than traditional surveys. A quarterly data collection is feasible.

In addition, the installation of the software to users' devices makes possible the retention of the selected sample as a panel, which will provide measurements for the accurate estimation of changes in Internet usage.

Finally, data can be combined and jointly analysed with data collected with regular questionnaires. In this report we have only utilised the demographic information from such questionnaires with the automatically collected data. Other data could have been used in exactly the same way.

On the other hand the method has several disadvantages. The most serious is the lack of trust from individuals towards the producer of statistics. The pilot study managed to acquire the consent of 3.7% of the online panel, which was approached by the company that had created the panel. This rate of cooperation is comparable with the rates reported by recent, similar studies⁴. In fact, the rate of 3.7% was

⁴ 5.8% of the chosen sample according to 'Bouwman, H., Heerschap, N., de Reuver, M. (2012) Mobile handset study 2012. The Hague: Statistics Netherlands' (p.10); 3.8% of the sample according to 'European Commission (2012) Internet as a data source. Luxembourg: Publications Office of the European Union.' (p. 148).

achieved with the use of a small financial incentive. The possibility that such an incentive may have to be used should not be ruled out by NSIs.

The chosen software cannot work on devices with the iOS operating system, i.e. iphone and ipad. This excludes a substantial share of the target population from the survey. Due to the design of iOS, this problem afflicts several tools that could be used for data collection. Care is therefore needed in the selection of the software tool; the development of bespoke solutions might be necessary.

An additional problem in the pilot study was the lack of transparency of the measurement process implemented by the tool. As shown earlier, it was not clear how usage time is defined by the makers of the software and why there were discrepancies between the reported durations of usage time (in minutes) and durations as shares of total usage time. An NSI must not accept such lack of transparency; it should have complete knowledge of what each measurement means. Time and resource constraints of the pilot study did not allow us to resolve this issue.

Finally, the software reports usage times per category of site; it does not report the times at which usage started and ended. To compute usage times for aggregates of categories, the producer of statistics must add the separate usage times. If the categories of sites have been used concurrently, which is very likely, the aggregated times will overestimate the true usage times. This is another point that shows the need for complete knowledge and control of the measurement process by the NSI.

Overall, the use of activity monitoring software shows great promise as a data collection tool and the ESS should carry out additional investigations of the statistical methodology and practical arrangements needed for its incorporation in regular statistical production.

3. Pilot survey of functionalities offered by the websites of business enterprises

3.1. Scope of the pilot

The aim of the pilot collection was to replicate the current survey on ICT usage and e-commerce in enterprises as far as possible, with the following differences:

- the pilot targets functionalities / technologies available to the websites of enterprises and not ICT usage or e-commerce in general
- data collection in the pilot is automated via software that retrieves and analyses the content of websites.

As in the individuals' pilot the limited time and resources devoted to the pilot surveys led to compromises and focus was again put on the most important aspects of the survey, those that differentiate it from the regular ICT survey.

3.1.1. Target population of the pilot survey

The target population of the pilot is the set of all Greek enterprises with a website, independently of economic activity or size in terms of number of persons employed. Therefore, it is both wider and narrower than the target of the Greek ICT survey, which covers almost all but not all economic activities, excludes enterprises with less than 10 employees and does not have possession of a website as a requirement for inclusion. The choice of this target population was dictated by the unavailability of a suitable sampling frame for the ICT survey's target population as will be explained in section 3.2.

3.1.2. Statistical indicators produced

All indicators that have been produced in the pilot survey are of the sort "Percentage of enterprises whose website ..." and they refer to whether the site provides specific types of information, uses particular types of technologies or offers certain facilities to its users.

An enterprise's website has been defined as the set of pages whose addresses start with the same single URL that characterizes the enterprise. For example, the website of Agilis SA is the set of pages whose addresses start with www.agilis-sa.gr.

The indicators measured in the pilot survey are the following:

1. Percentage of enterprises whose website provides a contact URL: the indicator refers to whether the site gives a contact URL among the contact information presented to users.
2. Percentage of enterprises whose website provides a contact email address.
3. Percentage of enterprises whose website provides a contact telephone number.
4. Percentage of enterprises whose website provides a contact postal address.
5. Percentage of enterprises whose website offer pages in the national language. The national language in the case of the pilot is Greek.

6. Percentage of enterprises whose website offer pages in English.
7. Percentage of enterprises whose website presents the date of its last update. The date does not need to be present on all pages. Presence in at least one page suffices.
8. Percentage of enterprises whose website presents the site's privacy policy.
9. Percentage of enterprises whose websites provides user registration facility.
10. Percentage of enterprises whose website presents its site map to users.
11. Percentage of enterprises whose website uses web analytic tools. The indicator refers to the deployment in the website of tools that analyse the number, provenance and behaviour of visitors to it. Such tools need not be – and usually are not – visible to the visitors.
12. Percentage of enterprises whose website advertises open positions or provides forms for applying for a job online.
13. Percentage of enterprises whose website provides links to multimedia content.
14. Percentage of enterprises whose website provides links to social networks or blogs.
15. Percentage of enterprises whose website provides links to wikis and wiki-sharing tools.

Certain of these indicators are comparable with indicators included in the model questionnaire of the 2013 ICT survey⁵. More specifically, the questionnaire contains the following questions:

- In January 2013, did the Website or Home Page have any of the following?
 - A privacy policy statement, a privacy seal or certification related to website safety – *comparable to indicator 8*
 - Advertisement of open job positions or online job application - *comparable to indicator 12; the question was optional in the 2013 survey*
- In January 2013, did your enterprise use any of the following social media?
 - Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer, etc) – *partly comparable to indicator 14, since the indicator does not cover internal networks such as Yammer*
 - Enterprise's blog or microblogs (e.g. Twitter, Present.ly, etc) – *comparable to indicator 14*
 - Multimedia content sharing websites (e.g. YouTube, Flickr, Picassa, SlideShare, etc) – *comparable to indicator 13*

⁵ Eurostat (2013) *Methodological manual for statistics on the Information Society – survey year 2013* (v. 3). Luxembourg: Eurostat. The questionnaire is annex 3.1 of the document.

- Wiki based knowledge sharing tools– *partly comparable to indicator 15 since the indicator does not cover internal wiki-based sharing tools*

3.2. Sampling procedure

The original intention was to draw a sample of enterprises from the business register of ELSTAT reproducing the stratified sampling scheme followed by the latter. Such a sample cannot be drawn from the register by third parties; it has to be requested and prepared from ELSTAT. During a meeting with the responsible ELSTAT members of staff on 20/9/2013 it turned out that the provision of the sample would require at least one month from the moment a formal request would be submitted to the authority.

The project team therefore decided to resort to other means of drawing a sample, even a non-random one. It was felt that the actual selection of the sample, carried out in the same manner as it is done in the regular survey, does not offer any input to the testing of the automated data collection method. The novel features of the method are found in the way it measures data; they can be tested on all kinds of samples.

Moreover, at the moment of analysing the pilot's results there were no published statistics on those indicators which are also measured in the ICT survey to compare them with. As reported in deliverable D2 of the project, the accuracy of the collected data was assessed with the help of human operators who visited the websites of a random subsample and tried to collect the same data by inspecting the sites' content. Therefore the lack of a random sample for the pilot would not cause issues in the assessment of accuracy either.

Private business registers, offered at a price by private vendors in Greece, were too costly for the resources of the project and in the end we resorted to a convenience sample. It was drawn from a list of enterprises compiled by project partner EELLAK, which contains contact details of Greek enterprises that have received in the past European funding for research. The total list contains 1777 enterprises.

A random sample of 281 enterprises was drawn from this list. The size of the sample was constrained by the fact that the software tool used in the collection of the data needed customisation effort which increases with the addition of more enterprises.

3.3. Software tool used in the pilot

The technique used for the automatic collection of data was *web crawling*. It amounts to visiting web addresses (URLs) and copying their content to a local repository for later processing. Web crawling is commonly used by Web search engines in order to facilitate indexing which is crucial for web searching. In the pilot, we applied what is commonly called *web scraping*, a data mining process which focuses on collecting specific parts of information from a web site and not all its content.

There are several utilities for web crawling but they are mostly adapted to static websites. Dynamic Web page creation has revolutionized the Web, but it has also hidden its content. For instance, it is not possible to view the source of a Twitter account's profile page. The page contains only javascript code. Almost everything on a Twitter page is built dynamically through JavaScript, and the crawlers cannot see any of it. There are crawlers that can deal with such pages but they usually require manual customization for every site they need to visit.

We opted to using **Google’s Custom Search Engine (CSE)**⁶, instead of any specific utility. It provides an interface to the user in order to specify a list of sites and a list of keywords to search for in these sites. The approach is explained in the following section.

We selected CSE for various reasons. Firstly, CSE relies on the crawling and indexing of the Web by Google. In other words, crawling has already been carried out by Google tools and the retrieved context has been suitably indexed for searching. A second reason was that the CSE employs probabilistic matching, i.e. it returns results even when a part of a keyword is identified in a site. This allows “catching” sites where the terms that we search for appear in a derivative form of the keywords we are using. Moreover, CSE allows the use of a simplified, easily accessible, regular expressions language, same as the special words and notation one uses in the regular search bar of Google. It also permits simultaneous searches in several languages (although only English were used in this pilot). Finally it allows the user to exclude specific types of content. In our case, for example, we needed to exclude the content of documents (Word, Powerpoint, PDF) provided by the websites; this is easily achieved with a suitable operator of the expressions language of CSE (e.g. *-filetype:ppt*).

3.4. Implementation of the pilot

The collection of the data relies on the use of keywords. Each of the indicators listed in section 3.1.2 is viewed as resulting from answering “Yes” to a question asking whether the website has / provides / uses / offers the mentioned type of content or facility.

Instead of asking questions we specified a number of keywords relevant to each indicator. Appearance of even one of these in at least one page of a website was considered as a “Yes” to the corresponding fictional question. Therefore we only needed to provide suitable keywords and the addresses of the websites of the sample to the CSE; it would then search among the content that Google has already indexed.

The selection of suitable keywords was not an one-off operation. Initial “trial” sets of keywords were used and their results were reviewed by human operators and cross-checked versus the findings of manual searches in the websites. Additional keywords and stems of keywords were then proposed and tried again. The final list of keywords used for each indicator is shown in Table 19 below.

Table 19. List of keywords used for collecting the data of the pilot survey.

Indicator	Relevant keywords
Website provides a contact URL	url, Website
Website provides a contact email address	e-mail, Email, E-mail, email, eMail, E
Website provides a contact telephone number	telephone, telephone number, Phone, Tel., Fax, Tel/Fax, T:, tel, TELEPHONE

⁶ <https://www.google.com/cse/all>.

Indicator	Relevant keywords
Website provides a contact postal address	address, Postal Address, Post code, P.O. box,
Website offers pages in the national language	Greek, EL
Website offers pages in English	Language, English, EN
Website presents the date of its last update	Last Update, Last Updated, Dated
Website presents the site's privacy policy	privacy policy, terms of use, Privacy Statement, Conditions of use, Terms and Conditions, Terms & Conditions, Privacy, Legal, DISCLAIMER, Disclaimer, Copyright
Website offers user registration facility	Signin, login, Login, register, Create an Account, openID, registration, Subscribe
Website presents its site map to users	sitemap, site map, SITEMAP, Sitemap, Site Map
Website uses web analytic tools	analytics, google analytics
Website advertises open positions or provides forms for applying for a job online	jobs, vacancies
Website provides links to multimedia content	Mpeg
Website provides links to social networks or blogs	widgets, Facebook, LinkedIn, Yammer, Twitter, Follow us, Share this page, Like us, T, F, BLOGS, Follow
Website provides links to wikis and wiki-sharing tools	Wikis

The CSE returns a list of URLs (pages, within each website) where any of these keywords has been found. Therefore, if for example site www.agilis-sa.gr contains in four of its pages the keyword

“telephone” and in three more (possibly overlapping) it contains the keyword “tel”, the results will list seven URLs with the keyword found in each one attached to them. Post-processing with a text parser grouped such findings into a single “hit” per indicator and website. The results of descriptive analysis of the numbers of hits produced the statistics shown in section 3.5.

3.5. Results

As it has been described in section 3.2, the sample of pilot survey was a convenience sample drawn from a very small list of enterprises. For this reason no stratification was carried out. The results however are presented, wherever appropriate, broken down by economic activity and by region where the enterprise is located.

Due to the small size of the sample and the way in which it was selected the accuracy of the results is probably small; moreover, it cannot be quantified with, for example, variance estimates.

NACE rev. 2 at single digit level has been used for the classification of enterprises by economic activity. Due to the small size of the sample certain single-digit activities were present in very small numbers or not present at all in it. Because of this some single-digit classes have been merged in the presentation of the results. The need however to respect the thematic affinity of the classes being merged forced us to keep two classes, “Accommodation and food service activities” (I) and “Real estate activities” (L) separate, although they are not present in the sample. The final list of activity classes and the number of enterprises per class are shown in Table 20 below.

For the classification of enterprises by location, level 2 of the Nomenclature of territorial units for statistics (NUTS 2) has been used. All such regions of Greece except for one, “Notio Aigaio” (EL 42), were represented in the sample. The number of enterprises by region is shown in Table 21.

Table 20. Distribution of the enterprise sample of the pilot survey by economic activity (NACE rev. 2).

Code	Economic activity	Number of enterprises
A - B	Agriculture, forestry and fishing & Mining and quarrying	9
C	Manufacturing	62
D - E	Electricity, gas, steam and air conditioning supply & Water supply; sewerage; waste management and remediation activities	7
F	Construction	5
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	45
H	Transporting and storage	4
I	Accommodation and food service activities	-
J	Information and communication	59
K	Financial and insurance activities	4

Code	Economic activity	Number of enterprises
L	Real estate activities	-
M	Professional, scientific and technical activities	54
N	Administrative and support service activities	4
O - S	Public administration and defence; compulsory social security, Education, Human health and social work activities, Arts, entertainment and recreation & Other services activities	28
	Total	281

Note: the economic activity of each enterprise has been deduced by members of the team based on the content of their websites.

Table 21. Distribution of the enterprise sample of the pilot survey by region (NUTS 2).

Region (NUTS 2 code)	Number of enterprises
Anatoliki Makedonia, Thraki (EL 11)	3
Kentriki Makedonia (EL 12)	52
Dytiki Makedonia (EL 13)	3
Thessalia (EL 14)	8
Ipeiros (EL 21)	5
Ionia Nisia (EL 22)	1
Dytiki Ellada (EL 23)	7
Stereia Ellada (EL 24)	3
Peloponnisos (EL 25)	20
Attiki (EL 30)	169
Voreio Aigaio (EL 41)	2
Notio Aigaio (EL 42)	
Kriti (EL 43)	8
Total	281

Note: the location of each enterprise has been deduced by members of the team based on the contact details in their websites.

3.5.1. Availability of contact information on the websites

This section presents the findings of the automated software about the type and extent of availability of contact information on the enterprises' websites. Telephone number is the most widely available type of information (86.1% of enterprises provide it), followed by email address. The same pattern appears in all economic activity classes except for "Agriculture, forestry and fishing; Mining and quarrying".

Telephone number is the most common type of contact information in all regions except for Thessalia (EL 14), Voreio Aigaio (EL 41) and Kriti (EL 43). In several regions URL is the second most frequently provided contact information, surpassing email address.

The smallest percentages of availability of contact information appear in the few enterprises with economic activity "Administrative and support service activities" (N) and in the enterprises of Thessalia (EL 14) and Voreio Aigaio (EL 41).

"URL" may seem odd as a type of information, given that all data are retrieved from websites. In other words URLs are available in all cases. Nevertheless the indicator shows the percentage of enterprises that provide URLs as part of the content of their websites and does not refer to URLs in the address bar of the browser.

Table 22. Percentage (%) of enterprises that present particular types of contact information on their website; by economic activity (NACE rev. 2).

Economic activity	URL	e-mail address	Telephone number	Postal address	Number of enterprises
A - B	55.6	100.0	88.9	77.8	9
C	74.2	77.4	83.9	67.7	62
D - E	57.1	57.1	85.7	28.6	7
F	80.0	80.0	100.0	60.0	5
G	82.2	88.9	91.1	77.8	45
H	75.0	75.0	75.0	50.0	4
I	-	-	-	-	0
J	76.3	74.6	93.2	66.1	59
K	100.0	100.0	100.0	100.0	4
L	-	-	-	-	0
M	68.5	74.1	74.1	63.0	54

N	50.0	50.0	75.0	75.0	4
O - S	78.6	85.7	89.3	67.9	28
Total	74.4	79.0	86.1	67.6	281

Table 23. Percentage (%) of enterprises that present particular types of contact information on their website; by region (NUTS 2).

Region	URL	e-mail address	Telephone number	Postal address	Number of enterprises
Anatoliki Makedonia, Thraki (EL 11)	100.0	100.0	100.0	100.0	3
Kentriki Makedonia (EL 12)	69.2	73.1	78.8	53.8	52
Dytiki Makedonia (EL 13)	33.3	66.7	100.0	33.3	3
Thessalia (EL 14)	75.0	75.0	62.5	75.0	8
Ipeiros (EL 21)	60.0	80.0	80.0	60.0	5
Ionia Nisia (EL 22)	100.0	100.0	100.0	100.0	1
Dytiki Ellada (EL 23)	85.7	71.4	100.0	85.7	7
Stereia Ellada (EL 24)	100.0	66.7	100.0	66.7	3
Peloponnisos (EL 25)	70.0	75.0	90.0	60.0	20
Attiki (EL 30)	75.1	81.1	88.2	72.2	169
Voreio Aigaio (EL 41)	100.0	50.0	50.0	50.0	2
Notio Aigaio (EL 42)	-	-	-	-	0
Kriti (EL 43)	87.5	100.0	87.5	62.5	8
Total	74.4	79.0	86.1	67.6	281

3.5.2. Language options of websites

The second group of indicators quantify the availability of enterprise websites in Greek and in English. As it was expected the share of sites available in Greek exceeds that of websites available in English although not by much. This shows that the majority of websites are available in both languages.

It seems strange that there are enterprises which do not provide their site in Greek. This result is an artefact of the way the indicator is measured. There is no linguistic analysis of the content's language. The availability of a page (hence of the site) in a given language is detected by the existence of textual links (e.g. "English") towards pages in that language. This way of measurement is also probably the reason for results such as for "Wholesale and retail trade; repair of motor vehicles and motorcycles" (G) or Kentriki Makedonia (EL 12) and Sterea Ellada (EL 24) where more enterprises seem to offer their websites in English than in Greek.

Table 24. Percentage (%) of enterprises that present their website in particular languages; by economic activity (NACE rev. 2).

Economic activity	Greek (national)	English	Number of enterprises
A-B	66.7	66.7	9
C	83.9	74.2	62
D-E	85.7	71.4	7
F	100.0	60.0	5
G	77.8	86.7	45
H	75.0	50.0	4
I	-	-	0
J	81.4	76.3	59
K	100.0	100.0	4
L	-	-	0
M	81.5	70.4	54
N	100.0	50.0	4
O-S	78.6	78.6	28
Total	81.5	75.4	281

Table 25. Percentage (%) of enterprises that present their website in particular languages; by region (NUTS 2).

Region	Greek (national)	English	Number of enterprises
Anatoliki Makedonia, Thraki (EL 11)	100.0	100.0	3
Kentriki Makedonia (EL 12)	71.2	76.9	52
Dytiki Makedonia (EL 13)	100.0	100.0	3
Thessalia (EL 14)	75.0	75.0	8
Ipeiros (EL 21)	80.0	80.0	5
Ionia Nisia (EL 22)	100.0	100.0	1
Dytiki Ellada (EL 23)	85.7	85.7	7
Stereia Ellada (EL 24)	66.7	100.0	3
Peloponnisos (EL 25)	75.0	55.0	20
Attiki (EL 30)	84.0	74.6	169
Voreio Aigaio (EL 41)	100.0	100.0	2
Notio Aigaio (EL 42)			
Kriti (EL 43)	100.0	87.5	8
Total	81.5	75.4	281

3.5.3. Website facilities

The two following tables present the availability of certain facilities or types of information on enterprise websites. The most prevalent facility is a site map (36.7%), followed by a statement of the privacy policy implemented in the website (27.4%). On the other hand no user registration facility was detected in any sample unit. Usage of web analytics tool (a facility for the website owners, usually not visible to the visitors to the site) has been detected to almost 15% of websites.

In general, large percentages are observed only in activity classes or regions with small numbers of sample units, where accuracy is even less than in the rest of the sample.

Table 26. Percentage (%) of enterprises that have specific facilities on their website; by economic activity (NACE rev. 2).

Economic activity	Last update date	Privacy policy statement	Registration facility	Site map	Use of web analytics tools	Number of enterprises
A-B	0.0	11.1	0.0	22.2	0.0	9
C	0.0	33.9	0.0	35.5	8.1	62
D-E	0.0	14.3	0.0	28.6	0.0	7
F	0.0	20.0	0.0	40.0	20.0	5
G	4.4	28.9	0.0	42.2	15.6	45
H	25.0	25.0	0.0	75.0	25.0	4
I	-	-	-	-	-	0
J	6.8	32.2	0.0	39.0	32.2	59
K	50.0	100.0	0.0	75.0	50.0	4
L	-	-	-	-	-	0
M	1.9	18.5	0.0	29.6	3.7	54
N	0.0	25.0	0.0	50.0	25.0	4
O-S	10.7	17.9	0.0	32.1	7.1	28
Total	4.6	27.4	0.0	36.7	14.2	281

Table 27. Percentage (%) of enterprises that have specific facilities on their website; by region (NUTS 2).

Region	Last update date	Privacy policy statement	Registration facility	Site map	Use of web analytics tools	Number of enterprises
Anatoliki Makedonia, Thraki (EL 11)	0.0	66.7	0.0	33.3	0.0	3
Kentriki Makedonia (EL 12)	5.8	19.2	0.0	26.9	7.7	52

Region	Last update date	Privacy policy statement	Registration facility	Site map	Use of web analytics tools	Number of enterprises
Dytiki Makedonia (EL 13)	33.3	0.0	0.0	33.3	0.0	3
Thessalia (EL 14)	12.5	0.0	0.0	37.5	12.5	8
Ipeiros (EL 21)	0.0	40.0	0.0	0.0	40.0	5
Ionia Nisia (EL 22)	0.0	0.0	0.0	0.0	0.0	1
Dytiki Ellada (EL 23)	14.3	28.6	0.0	71.4	14.3	7
Stereia Ellada (EL 24)	0.0	0.0	0.0	33.3	33.3	3
Peloponnisos (EL 25)	0.0	15.0	0.0	15.0	10.0	20
Attiki (EL 30)	4.1	31.4	0.0	41.4	16.6	169
Voreio Aigaio (EL 41)	0.0	50.0	0.0	50.0	0.0	2
Notio Aigaio (EL 42)	-	-	-	-	-	0
Kriti (EL 43)	0.0	50.0	0.0	50.0	12.5	8
Total	4.6	27.4	0.0	36.7	14.2	281

3.5.4. Other content of the websites

Finally, the availability of certain other types of content was measured. Slightly more than one third of the enterprises advertise vacancies or provide forms for online applications for work on their sites. This share does not vary greatly by economic activity, except for two classes with very few members in the sample. Attiki (EL 30) where more than half of the sample is located has a rate of availability of this information that is 8 percentage units higher than the sample's average. Wikis are the second most common type of information, available by slightly more than 20% of the sample's enterprises. On the other hand very few enterprises provide multimedia content on their sites and almost no enterprise offers links to social networks or blogs.

Table 28. Percentage (%) of enterprises offering additional content on their website; by economic activity (NACE rev. 2).

Economic activity	Open job positions or online job application	Multimedia content	Social networks or blogs	Wikis and wiki- based sharing tools	Number of enterprises
A-B	22.2	0.0	0.0	22.2	9
C	30.6	0.0	0.0	21.0	62
D-E	28.6	0.0	0.0	14.3	7
F	20.0	20.0	0.0	0.0	5
G	40.0	2.2	2.2	22.2	45
H	25.0	25.0	0.0	25.0	4
I	-	-	-	-	0
J	45.8	10.2	0.0	27.1	59
K	75.0	0.0	0.0	100.0	4
L	-	-	-	-	0
M	29.6	1.9	0.0	5.6	54
N	50.0	0.0	0.0	0.0	4
O-S	28.6	3.6	0.0	39.3	28
Total	35.2	3.9	0.4	21.7	281

Table 29. Percentage (%) of enterprises offering additional content on their website; by region (NUTS 2).

Region	Open job positions or online job application	Multimedia content	Social networks or blogs	Wikis and wiki-based sharing tools	Number of enterprises
Anatoliki Makedonia, Thraci (EL 11)	33.3	0.0	0.0	33.3	3
Kentriki Makedonia (EL 12)	26.9	0.0	0.0	15.4	52
Dytiki Makedonia (EL 13)	0.0	33.3	0.0	33.3	3
Thessalia (EL 14)	12.5	12.5	0.0	12.5	8
Ipeiros (EL 21)	40.0	0.0	0.0	20.0	5
Ionia Nisia (EL 22)	0.0	0.0	0.0	0.0	1
Dytiki Ellada (EL 23)	57.1	14.3	0.0	42.9	7
Stereia Ellada (EL 24)	0.0	0.0	0.0	33.3	3
Peloponnisos (EL 25)	5.0	5.0	0.0	10.0	20
Attiki (EL 30)	42.6	4.1	0.6	24.3	169
Voreio Aigaio (EL 41)	0.0	0.0	0.0	50.0	2
Notio Aigaio (EL 42)	-	-	-	-	0
Kriti (EL 43)	50.0	0.0	0.0	12.5	8
Total	35.2	3.9	0.4	21.7	281

3.6. Assessment of the accuracy and specificity of the collected data

The use of keywords may generate spurious results. For example the word “telephone” will be used in a page listing contact information but it may also be used in a different context, e.g. the company apologising in its site “... for our helpdesk telephones not been operational yesterday morning”. Keywords not thought of may be used in other websites and their presence will go undetected.

In order to test the accuracy and specificity of the results of the data collection tool we selected at random 61 enterprises (approximately 22% of the sample) and we examined manually and recorded whether the

indicators apply to them. The results of the manual search have been compared with the outcome that has been produced by the data collection tool. They are summarised in the following table.

Table 30. Characterisation of the results of the CSE over a sample of 61 enterprises.

Indicator	True negative (%)	True positive (%)	False negative (%)	False positive (%)	True positives as share of reported positives (%)	True positives as share of positives (%)
Website provides a contact URL	24.6	14.8	8.2	52.5	22.0	64.3
Website provides a contact email address	3.3	77.0	9.8	9.8	88.7	88.7
Website provides a contact telephone number	1.6	75.4	21.3	1.6	97.9	78.0
Website provides a contact postal address	4.9	55.7	36.1	3.3	94.4	60.7
Website offers pages in the national language	6.6	65.6	13.1	14.8	81.6	83.4
Website offers pages in English	11.5	59.0	24.6	4.9	92.3	70.6
Website presents the date of its last update	100.0	0.0	0.0	0.0	-	-
Website presents the site's privacy policy	70.5	16.4	11.5	1.6	91.1	58.8
Website presents its site map to users	57.4	34.4	4.9	3.3	91.2	87.5
Website uses web analytic tools	90.2	1.6	0.0	8.2	16.3	100.0
Website advertises open positions or provides forms for applying for a job online	57.4	21.3	11.5	9.8	68.5	64.9
Website provides links to social networks or blogs	70.5	0.0	29.5	0.0	-	0.0
Website provides links to wikis and wiki-sharing tools	85.2	0.0	0.0	14.8	0.0	-
Website provides links to multimedia content	72.1	0.0	27.9	0.0	-	0.0

where,

- “True negative” (TN) is when both CSE and manual search have concluded that the indicator does not exist;
- “True positive” (TP) is when both CSE and manual search have concluded that the indicator does exist;
- “False negative” (FN) is when CSE does not identify the indicator whereas manual search revealed that the indicator does exist;
- “False positive” (FP) is when CSE identifies the indicator whereas manual search revealed that the indicator does not exist.
- “True positives as share of reported positives” is the ratio $TP / (FP+TP)$. If the ratio has a small value, this means that it returns a lot of spurious findings (the false positives) and therefore has little specificity. On the other hand the higher the ratio the higher the specificity.
- “True positives as share of positives” is the ratio $TP / (TP+FN)$. If the ratio has a small value, this means that a lot of sites, which have the characteristic of interest are not detected. In other words, the indicator is not sensitive to the characteristic’s presence. On the other hand the higher the ratio the higher the sensitivity.

The performance of the CSE is variable between the indicators. The two most relevant columns in Table 30 reveal the following:

- Column before last: the rate of correct spotting of an indicator by the CSE ranges from 0% for links to wiki-based sharing tools up to larger than 90% for contact telephone numbers or postal addresses and for availability of pages in English.
- Last column: the rate of true occurrences of an indicator that have been correctly spotted by the CSE ranges from 0% in the case of links to social networks, blogs or multimedia content up to 100% for the case of usage of web analytic tools. The latter however is unreliable because it the human operators practically relied on keywords too, i.e. the sites mentioning usage of web analytics, in order to discern whether such tools are used. Other large rates appear in the availability of contact email address, site map and pages in the national language. Overall however, the performance is not as accurate as would be wished for statistical purposes.

The most accurate indicators (high rates in the two last columns) are those about the availability of contact email address and telephone number, the availability of a site map and the availability of the site’s pages in English.

The high rates of false negatives and false positives require some explaining:

- The false detection of contact URLs is due to the use of very generic keywords: ‘URL’ and ‘website’ which appear frequently in contexts other than that of contact information. This indicator is perhaps superfluous in a setting of data retrieved from websites.
- The main causes of the high rates of non-detection of contact telephone or contact postal addresses are two: a) the use of Greek keywords by the websites and b) the use of icons (e.g. an image of a telephone or of an envelope) instead of keywords. These two reasons will be mentioned again below.
- Many Greek enterprises offer their pages only in English. The many wrong detections of pages in Greek are caused by the too generic keywords employed: ‘EL’, which may appear, for example, as part of a postal address and ‘Greek’ which may refer to Greek products and not to the Greek language.
- The non-detection of pages in Greek or in English is caused by the same two reasons mentioned earlier: the use of icons or of keywords in Greek.
- The false detection of job listings or job application forms is caused by the too generic keywords used. They are ‘vacancies’ and ‘job’. They can appear even if no jobs are listed; for example a link ‘vacancies’ may be always available although the respective page may be empty (‘Currently no vacancies’).
- The non-detection of job listings or job application forms is caused by the small number of keywords and by non-use of keywords in Greek.
- The false detection of wikis is stranger. It is not clear why some sites have been indexed as containing the keyword ‘Wikis’. Manual inspection of the sites and of their HTML source code has failed to detect the keyword.
- Links to social networks, blogs, etc. are not detected because of the frequent use of icons (e.g. the logos of Facebook or Twitter) instead of verbal references to them.
- The same reason causes the high rate of non-detection of links to multimedia content. Many sites provide icons or thumbnails of videos or other multimedia content instead of verbal links. The lack of a sufficient number of keywords is another reason of non-detection.

The results indicate that the selection of keywords is a crucial activity in this type of data collection with a considerable impact on the quality of statistics. Due to the fact that keywords can always appear in contexts not relevant to the indicators, there are probably limits to how accurate such a data collection can be. Finally, keywords are not appropriate for target characteristics that manifest themselves without keywords: the language of a page becomes evident with linguistic analysis of its content, some links are usually presented as icon, while the deployment of particular tools (e.g. web analytics) or technologies by the site is better detected with detection of the technologies themselves.

3.7. Conclusions

The automatic collection of data from the web sites of enterprises has merits but the results of the particular approach chosen in the pilot study are not very encouraging.

Some of the positive features of this mode of data collection are similar to those of monitoring software (section **Error! Reference source not found.**).

After an initial set-up period, devoted to the specification of keywords and other site features to be detected, the collection of data is a lot faster than traditional survey data collection. It is also scalable to large sample sizes. This permits the implementation of data collection at higher frequencies and to larger samples than traditional surveys.

Furthermore, the data collection that relies on Google's search infrastructure and indexing is non-intrusive. Google has already processed the data and the NSI is querying Google's results and not the sites.

The speed, automation and possible non-intrusiveness of the approach mean that a panel sample of enterprises can be set-up by the producer of statistics. To move things a little further, even a 'census' of enterprise sites could be established over the long term, for indicators, which can be measured accurately enough. Financial costs and time requirements of such a census should of course also be taken into account in any decision-making.

The disadvantages of the specific data collection mode used in the pilot outweigh its merits. The most serious is that the data returned by the search engine contain many spurious findings while on the other hand several occurrences of the site characteristics in which we were interested went un-noticed. The results suffer from lack of both sensitivity and specificity. This is a deficiency of keywords. There seems to exist a limit to how specific keywords can be to the targeted site features: most features are associated with terminology, which also applies to other, unrelated, issues.

An obvious improvement of the approach's sensitivity is to also include keywords in the national language of each country. A second direction for potential improvement is to download the HTML source code of web sites and extract keywords from it as well. This would permit detection of filenames (e.g. 'envelope.gif' for an icon showing a postal envelope and accompanying the display of postal contact information) or reserved words (e.g. 'mailto') indicative of features of the sites. This approach requires the use of additional crawlers besides Google's search engine.

Detection capabilities could possibly improve if linguistic analysis of a site's content identified directly the language it is written in, instead of relying on imprecise keywords. Moreover, key icons (e.g. the logos of Facebook or Twitter) could be detected with some kind of image analysis or image search.

Besides site features that are manifested through keywords that cannot be specific enough there are other features which are not connected to verbal aspects of the sites. For example, video thumbnails may be the links to Youtube videos, without any keywords. Furthermore, web analytics may be deployed on a site invisibly to its visitors. Such features require the utilisation of tools that detect technologies rather than keywords.

Based on the results of the pilot study it can be inferred that the developed methodology for collecting data from enterprise web sites does not produce statistics of high enough quality. A more extended appraisal of the method, which will encompass aspects of multilingualism, extraction of source code and detection of technologies, is needed for a more informed decision about its usefulness.

4. General conclusions

The two pilot surveys gave contrasting results. The one among individuals gave promising results despite its problems. Monitoring of activities online (or offline if required) can give very rich, detailed information, adaptable to changing statistical needs. The reluctance of users to be monitored is a major obstacle. Limits in processing and storage capacity can also emerge in large scale or long-term applications. With suitable sample design for the selection of individuals and devices it seems that statistical issues will not be serious.

On the other hand, the survey among enterprises gave inaccurate results while also missing information that could have been obtained with a questionnaire. The detection of site features cannot rely only on keywords: linguistic analysis, image search and detection of technologies could be useful additions with considerable impact on the accuracy of results. The type of indicators that can be measured by visiting websites and analysing their content or technologies needs careful consideration and the tools to be used need careful tuning.

5. Annexes

5.1. Annex 1 – Information note sent to potential members of the sample of the pilot survey of individuals

Trial production of official statistics with automatic data collection using Internet on computers, phones and tablets

The Information and Communications Technologies (ICT) is an important factor in economic development. The national statistical offices of the European Union member countries produce official statistics regarding their use by the public.

The company [Agilis S.A. Statistics and Informatics](#) and the non-profit [Free Software Company/ Open Source Software \(FSC / OSS\)](#) conduct a research project on behalf of Eurostat to consider whether some of the necessary data can be collected by automatic recording of users activity of the Internet, instead of using questionnaires as it occurs now.

Within the project's framework, enterprises would like to organize a pilot data collection from members of our panel. The collection method planned is the following:

- The members of the panel, who are wishing to participate in the collection, have to state this to NAME OF MARKET RESEARCH COMPANY and answer a short questionnaire in order to identify one or two most commonly used devices of each user.
- The participants will receive a message from NAME OF MARKET RESEARCH COMPANY, which will include codes and guides for the installation of the software on one or two devices that are going to be used in the pilot survey.
- The software records the user's activity for two weeks. The data are received by the NAME OF MARKET RESEARCH COMPANY and are protected from any impermissible access by any third party.
- At the end of the two weeks, the software will be shut down and users will receive from NAME OF MARKET RESEARCH COMPANY instructions to uninstall the software from their devices.
- Users may be asked to send additionally the history of their browser (browser history) for these two weeks. This is an alternative data source of using the Internet⁷.
- Users will respond to a short questionnaire about the use of Internet and point out their opinions for this measurement mode⁸.
- The NAME OF MARKET RESEARCH COMPANY provides completely anonymized data to the other two companies, in order to calculate aggregate statistical indicators.

⁷ In the end it was decided not to request this piece of information from the sample members.

⁸ The members' opinions on this mode of measurement are discussed in section 7 of deliverable D2 of the project.

The software that will be used is “Qustodio” (<http://www.qustodio.com>). “Qustodio” is a software of “Parental Control” and records sites, applications that are used by a user and also the start and end time of each visit. The sites are categorized into groups and for those groups the related statistical indicators are being computed. The advantages offered by the use of this tool instead of using researchers to collect data in person, is firstly the fact that the data are collected in a quicker way and secondly the avoidance of errors due to oversight.

The statistical indicators that will be calculated are the following:

- Duration of use of specific type of sites (minimum, maximum, average) per visit per day
- Number of visits in a specific type of sites (minimum, maximum, average) per day and in total, during the survey.
- Percentage of users who have visited specific type of sites during the study.

The types of sites are the following⁹:

- Social Networks
- Communication via Internet (e.g. Skype)
- News
- Search information about goods or services
- Radio via Internet (web radio)
- Online games
- Monitoring / listening of movies, photographs and music.
- Health
- Educational
- Travel

The statistics that are going to be published in research reports of the project will concern only certain values of the indicators for the total of the participants or for large subsets of those (by sex, age group etc.). Nothing personal, nominal information will be provided to both companies.

For more information, please contact:

A contact name of NAME MARKET RESEARCH COMPANY.

⁹ The list does not mention all types of sites; it was reduced in order to save space and was provided for indicative purposes.

5.2. Annex 2 – Screening questionnaire sent to interested potential members of the sample of the pilot survey of individuals

Screening questionnaire

1. Do you use a personal computer (desktop computer or laptop) to access the Internet?
[Please do not include any kind of personal computer supplied by your employer, in your answer, even if you are bringing it at home]

- a. Yes
 b. No (→ question 3)
 c. No answer (→ question 3)

2. **(Only for those who have responded Yes in question 1)** Do you exclusively use this computer or do you at least have a personal user's account on it?

- a. Yes, I use it exclusively.
 b. Yes, I have a personal user's account even though I share this computer with other users.
 c. No. None of the above.
 d. No answer.

3. Do you have a mobile phone?

- a. Yes
 b. No (→ question 5)
 c. No answer (→ question 5)

4. **(Only for those who have responded Yes in question 3)** What is the operating system of your mobile phone?

- a. Android

--

b. iOS (iPhone)

c. Other. Please specify.....

d. Don't know. Please specify the brand and the mobile model name.....

e. No answer.

5. Do you own a tablet computer, personal or in your household, which you use often?
[Please do not include any tablet supplied by your employer, in your answer, even if you are bringing it at home]

a. Yes

b. No (→ question 7)

c. No answer (→ question 7)

6. **(Only for those who have responded Yes in question 5) What is the operating system of your tablet?**

a. Android

b. iOS (iPhone)

c. Other. Please specify.....

d. Don't know. Please specify the brand and the tablet's model name.....

e. No answer.

7. Which of the following mobile devices do you use more often for accessing the Internet?

a. Mobile phone

b. Tablet

c. Neither of them

--

d. No answer

--

5.3. Annex 3 – Additional data questionnaire sent to members of the sample of the pilot survey of individuals

Assessment of your participation in the pilot, automatic data collection of Internet use

1. Did your participation in the pilot, automatic data collection of Internet use create concerns, fears or other kinds of non- technical problems?
 - a. Yes
 - b. No (→ question 3)

2. **(Only for those who have responded a in question 1)** Please describe these problems.

3. Did you encounter technical problems during the installation of the software that automatically collects data on any of your devices? **[Answer for all the devices on which you had any problems and describe them briefly]**
 - a. Yes, in the computer. Describe briefly _____
 - b. Yes, in the mobile phone. Describe briefly _____
 - c. Yes, in the tablet. Describe briefly _____
 - d. No, it did not create any problem in any of my devices

4. Did the use of the automatic data collection software create technical problems when using your devices? **[Answer for all the devices on which you had any problems and describe them briefly, e.g. increased battery consumption, slow running, crashes etc.]**
 - a. Yes, in the computer. Describe briefly _____
 - b. Yes, in the mobile phone. Describe briefly _____
 - c. Yes, in the tablet. Describe briefly _____
 - d. No, it did not create any problem in any device

5. Suppose that an official statistical service has chosen you at random as part of a sample for collecting data automatically by using Internet, in order to produce official statistics. Would you accept to participate?
 - a. Yes, definitely (→ go to questions Demographic data, data on computer and Internet use)
 - b. Yes, under certain conditions (→ Q6)
 - c. No, I would not accept (→ Question 7)

6. **(Only for those who answered b to 5)** Please tell us the conditions under which you would accept to participate. (→ go to questions Demographic data, data on computer and Internet use)
7. **(Only for those who answered c to 5)** Please tell us why you would not accept to participate.

Demographic data, data on computer and Internet use

1. Age, according to the most recent birthday: _____
2. Gender
 - a. Man
 - b. Woman
3. In what country were you born?
 - a. Greece
 - b. Another country of the European Union
 - c. Another country outside the European Union
4. Citizenship
 - a. Greek
 - b. Another country of the European Union
 - c. Another country outside the European Union
5. Completed level of education
 - a. There have not attended /completed any level of education
 - b. Primary school
 - c. Secondary school - lower technical schools
 - d. High school
 - e. Technical schools
 - f. University
 - g. Universities, Military Schools, Open University
 - h. Masters (Msc., MBA, MA, MLitt, MPHIL)
 - i. Doctorate (PhD)
6. Main occupation
 - a. Employee

- b. Self-employed
 - c. Unemployed
 - d. Pupil, student
 - e. Another case (housewife, soldier, retired, rentier, unable to work etc.)
7. Region of residence: _____
8. Which of the following¹⁰, do you use to access the Internet? **[You can mark more than one]**
- a. Music player, movies or multimedia (media player)
 - b. Reader electronic book (e-book reader)
 - c. TV connected to the Internet (Smart TV)
 - d. Gaming (games console)
9. How often did you use a computer, on average, during the last 3 months?
- a. Every day or almost every day
 - b. At least once a week but not every day
 - c. Less often than once a week
10. During 2013, for which of the following reasons did you use Internet, as part of your dealings with public services for personal matters? **[You can mark more than one]**
- a. To download information
 - b. To obtain application forms, certificates, etc.
 - c. To send completed forms
 - d. I didn't have transactions with public services or I had but I didn't use Internet
11. When was the most recent purchase or order goods or services via Internet (but not via e-mail), for your personal use?
- a. The quarter October - December, 2013
 - b. During the period January - September 2013
 - c. Before 2013
 - d. I have never bought/ ordered (→ end of interview)
12. **(Only for those who answered a, b, or c to 11)** Which of the following items did you purchase via Internet, for your personal use during 2013 (excluding orders or purchases through e-mail)? **[You can mark more than one]**
- a. Movies, music
 - b. Electronic books, magazines, newspapers or training material (e-learning)
 - c. Software, including computer games
 - d. Other (→ end of interview)

¹⁰ Questions about usage of PCs, smartphones and tablets had been asked with the screening questionnaire.

13. **(Only for those who answered a, b, or c to 12)** Which of the following items purchased via Internet for personal use, you downloaded from websites instead of receiving regular mail? **[You can mark more than one]**
- a. Movies, music
 - b. Electronic books, magazines, newspapers or training material (e-learning)
 - c. Software, including computer games

5.4. Annex 4 – Information note sent to owners of website enterprises, potential members of the sample of the pilot survey

Trial production of official statistics with automatic data collection of non-confidential data coming from the enterprises' websites

The Information and Communications Technologies (ICT) is an important factor in economic development. The national statistics offices of the European Union member countries produce official statistics regarding the use of the ICT by the enterprises.

The company Agilis S.A. of Statistics and Informatics and the non-profit Free Software Company/ Open Source Software (FSC / OSS) conduct a research project on behalf of Eurostat to consider whether some of the necessary data can be collected automatically from the enterprises' websites, instead of using questionnaires as it occurs now.

The information that the enterprises usually provide on their websites, includes data related to the following statistical indicators:

- Does the site provide the enterprises' contact e-mail address?
- Does the site provide the enterprises' contact phone number?
- Does the site provide the enterprises' address details?
- Is the website available in Greek?
- Is the website available in English?
- Does the website publish the date of its last update?
- Does the website publish the privacy policy which is related with the security of the website?
- Does the website provide a registration facility?
- Does the website publish a site map?
- Does the website use online tools for web analytics?
- Does the enterprise announce new job opportunities through its website?
- Is there an online job application form?
- Does the website provide any links to multimedia content?
- Does the website provide any links to social networks or blogs (Facebook, LinkedIn, Yammer, Twitter, etc.) ?
- Does the website provide any links to wikis?

There are two ways of finding and collecting the necessary information:

- a) through the Google web search engine tool by using the appropriate key words. The main purpose of this is to take full advantage of the way Google has organized the data of your website and to highlight the appropriate information.
- b) by making questions directly to your website's code as it is appearing on the

Internet. These questions aim to highlight the automated information which is available at your website code, i.e. Is there a sitemap? Do you use some kind of a web analytics tool? ect.

The two options described above ensure that the data are being searched **only among non-confidential information** which is available for everyone who visits the enterprise's website. The advantages offered by the use of this tool instead of using researchers to collect data in person, is firstly the fact that the data are collected in a quicker way and secondly the avoidance of errors due to oversight.

For more information you can contact:

Photis Stavropoulos

Tel: 2111003310 (internal 147)

photis.stavropoulos@agilis-sa.gr

Akadimias 96-100, 10677, Athens

12.5. D6 – Cookbook for the implementation of new methods and indicators at national level

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’

D6 Cookbook for the implementation of new methods and indicators at national level

March 2014

Document Service Data

Type of Document	Deliverable		
Reference:	D6 – Cookbook for the implementation of new methods and indicators at national level		
Version:	2	Status:	Draft
Created by:	Photis Stavropoulos	Date:	12/3/2014
Distribution:	European Commission – Eurostat/G6, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	31/12/2013	Initial release
2	12/3/2014	Completion of the blank sections of the initial release

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

European Commission – Eurostat/G6	1
Contract No. 50721.2013.002-2013.169.....	1
‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’	1
Introduction.....	5
Part 1 - Statistics on the use of Internet by individuals	6
1 Statistical product.....	6
1.1 Statistical unit.....	6
1.2 Target population	6
1.3 Periodicity	6
1.4 Observation variables	6
1.5 Summary measures, aggregated variables, indicators and tabulation	7
1.6 Explanatory notes	8
2 Production methodology	9
2.1 Timetable – Survey period	9
2.2 Frame population	9
2.3 Sampling design	9
2.3.1 Stratification.....	10
2.3.2 Sample size.....	10
2.3.3 Weighting – Grossing up methods	11
2.4 Survey type.....	11
2.4.1 Data collection method.....	11
2.4.2 Independent versus embedded survey.....	12
2.4.3 Mandatory versus voluntary survey.....	13
2.4.4 Coping with refusals of selected individuals to be included in the sample.....	13
2.4.5 Quality control systems.....	13
2.5 Data processing	15
2.5.1 Data validation	15
2.5.2 Non-response treatment	15
2.5.3 Unit non-response.....	16
2.5.4 Item non-response	18
2.6 Data analysis.....	19
2.6.1 Post-processing	19
2.6.2 Computation of indicators	19
2.6.3 Estimation of the accuracy of the indicators.....	19
2.7 Confidentiality and privacy issues	19
3 Annexes	21
3.1 Software tools	21
3.2 Model questionnaire.....	22
3.3 Transmission format	23
Part 2 - Statistics on the facilities of business websites.....	26
1 Statistical product.....	26

1.1	Statistical unit	26
1.2	Target population	26
1.3	Periodicity	27
1.4	Observation variables	27
1.5	Summary measures, aggregated variables, indicators and tabulation	28
1.6	Explanatory notes	28
2	Production methodology	32
2.1	Timetable – Survey period	32
2.2	Frame population	32
2.2.1	Updating the Business Register with website information	32
2.3	Sampling design	33
2.3.1	Stratification.....	33
2.3.2	Sample size.....	33
2.3.3	Weighting – Grossing up methods	35
2.4	Survey type	36
2.4.1	Data collection method	36
2.4.2	Independent versus embedded survey.....	37
2.4.3	Mandatory survey versus voluntary survey	37
2.4.4	Contact person of the survey	38
2.4.5	Coping with refusals of selected enterprises to be included in the sample.....	38
2.4.6	Quality control systems.....	38
2.5	Data processing	39
2.5.1	Misclassification treatment.....	39
2.5.2	Non-response treatment	40
2.5.3	Unit non-response.....	41
2.6	Data analysis	43
2.6.1	Post-processing	43
2.6.2	Computation of indicators	43
2.6.3	Estimation of the accuracy of the indicators.....	43
2.7	Confidentiality and privacy issues	44
3	Annexes	45
3.1	Software tools	45
3.1.1	Web crawlers	45
3.1.2	Google’s Custom Search Engine	46
3.2	Example of mapping between target functionalities and keywords	49
3.3	Transmission format	50

Introduction

One strand of project “Internet as a data source” deals with the investigation of a user-centric and a web site-centric method of automatic collection of data about individuals and enterprises.

In brief, the user-centric method consists in installing monitoring software on computing devices of individuals (computers, smartphones, tablets) with access to the Internet and recording Internet and application usage data. The site-centric method consists in crawling business web sites and identifying, by analysis of the text displayed in them, functionalities that they offer to users.

The present deliverable is a guide for the application of these methods for the production of official statistics. Its audience are the producers of official statistics. The guide borrows its structure and some of its content from Eurostat’s “Methodological manual for statistics on the Information Society”¹. It contains two parts, part I presenting the user-centric method and part II the site-centric one.

¹ Eurostat (2013) *Methodological manual for statistics on the Information society*, v. 3. Luxembourg: Eurostat.

Part 1 - Statistics on the use of Internet by individuals

1 Statistical product

This chapter describes the statistical information to be produced, which is separate from the production methodology.

The elements that make up the statistical product, at an input level, are the statistical unit, the target population and the observation variables, and at the output level, the periodicity and the summary measures, aggregate variables and tabulation. Covering all the elements of the statistical product, the statistical concepts and the nomenclatures are also needed to assure harmonization and comparability of statistics.

1.1 Statistical unit

The statistics on the use of Internet by individuals have the individual as the statistical unit. This is the unit that we want to observe or analyse.

1.2 Target population

The *target population* of the statistics on the use of Internet by individuals consists of all individuals, aged 16 or over who use computers or mobile devices with access to the Internet.

The *frame population* is an operationalization of the target population, taking the form of a list of elements of the target population. Although a target population can be easily defined, in practice a list of all its elements is needed for its complete or partial (in case a sample is used) observation, and that can be very difficult to obtain. That list should be complete and include every element of the target population only once. However, most of the time it will suffer from both under-coverage and over-coverage. The frame population will be further explained in chapter 2.2.

1.3 Periodicity

The periodicity is quarterly, meaning that the data are collected and compiled once per quarter.

A quarterly survey has been made possible by the employment of the data collection method presented in section 2.4. The quarterly frequency is appropriate in view of the need for relevant and recent information on a “fast moving” study domain like the information society.

1.4 Observation variables

The survey will collect data on two groups of observation variables, distinguished by their nature and the mode of data collection used for each one.

The first group comprises *demographic background variables*, useful for defining demographic sub-groups of the target population and producing statistics about them. All of them are qualitative variables, i.e. they collect categorical information. Data on them will be collected

from population registers or with a questionnaire. A proposed model questionnaire may be found in the annexes, in section 3.2.

The second group consists of variables that measure *usage time*. Data on them will be collected with the help of software tools, which are installed on the devices of the users and monitor their activity. They are a loosely defined set because the data record usage time, within each calendar day, for each individual web site or application that the user has used. Therefore, the number of variables is equal to the number of sites or applications visited or used each day. Section 1.6 gives more detail about the definition of usage time.

NSIs may also wish to collect data on variables about the use of ICTs or the Internet similar to those collected in the Community survey on ICT usage in households and by individuals. The present guide does not deal with such variables; they are covered in detail by Eurostat's Methodological Manual for Information Society Statistics.

1.5 Summary measures, aggregated variables, indicators and tabulation

Two aggregated variables and three indicators will be produced from the collected data on usage time.

The aggregated variables are:

- Number of users of a given web site or application.
- Total amount of time that users spend on a given web site or application.

The indicators are:

- The share of users that have used a given web site or application.
- The average amount of time per user and calendar day spent on using a given web site or application.
- The share of total usage time that users devote to a given web site or application.

The indicators are computed by dividing the aggregated variables with suitable totals:

- Shares of users are computed by dividing numbers of users of each site or application with the total number of users.
- The average amount of time per site or application is computed by dividing the total time spent on it by the total number of users multiplied by the number of calendar days of the reference period.
- The share of total usage time per site or application is computed by dividing the total time spent on it by the total usage time of all sites and applications over the whole reference period.

The reference period of the statistics is the observation period, i.e. the period of time during which the monitoring software was active on the users' devices.

The aggregated variables and indicators can also be computed for specific sub-populations defined by the background demographic variables (e.g. by sex) or by usage time variables (e.g. average time per day on a given application measured only on the actual users of the application). Moreover, the aggregated variables and indicators can be computed separately for working days, weekends, holidays, etc.

Section 2.6 provides more details about the computation of the aggregated variables and indicators.

1.6 Explanatory notes

Usage time is the time users spend actively on their device visiting websites or using specific applications. The definition of ‘active usage’ of an application or website is therefore crucial metadata of the automatic recording of activity. An application may be on but not used all the time; the same applies to open web browser windows connected to a specific page.

An option would be to consider as active only windows which are in front of the user, i.e. not covered by other open windows. Thresholds can also be set for the amount of time that no action is taken on screen before the computer is considered as idle. Particular cases however may make the distinction between idle and active time difficult: for example, how do you characterise a media player that plays a music album hidden behind other windows?

A pilot survey undertaken in the context of a feasibility study for DG CONNECT², for example, adopted the following definition of a visit to a web domain: *“a series of page views from one user within one domain with a maximum length of 30 minutes between page views”*. Another pilot survey for Statistics Netherlands³ defined browsing sessions as follows: *“when a user visits a web domain within 30 seconds after closing a previous web domain, we assume that the user moves from one website to another within the same browsing session”*.

The NSI will therefore need to define carefully what constitutes active usage and which user actions signify its start and end, so that they can be recorded by the monitoring software. If the NSI chooses to use off-the-self monitoring software it should be sure that it knows what it measures and how it measures it, so that the properties of the produced statistics can be assessed properly.

The software presented in the annexes, in section 3.1, which was used by the authors of this cookbook in a pilot survey⁴, stops counting time after 5 minutes of idle time. Moreover, it separates usage time into “web time”, “social activity time” and “apps time”:

- Total time denotes all time that the computer is active, even if not on the Internet.
- Web time records only the time spent visiting websites. It is computed as one minute per connection. If for example an open tab makes connections, each connection counts as one minute.

² European Commission (2012) *Internet as a data source*. Luxembourg: Publications Office of the European Union.

³ Bouwman, H., Heerschap, N., de Reuver, M. (2012) *Mobile handset study 2012*. The Hague: Statistics Netherlands.

⁴ See deliverable D3 of the project “Internet as a data source”.

- Social activity corresponds to Facebook activities only, e.g. chatting with friends and is not a subset of web activity, although visits to the Facebook page also count as web activity.
- Apps time is the time spent using applications, even offline.

2 Production methodology

2.1 Timetable – Survey period

It is recommended that the monitoring of users' activity takes one calendar month. The time of activation for each sample member should ideally be spread uniformly over the whole quarter. If however this entails high costs for the NSI monitoring may be concentrated in a shorter period.

2.2 Frame population

This issue was already discussed in chapter 1.2 on the *target population*. The *frame population* (of *sampling population*) is the frame from which the sample will be drawn. Ideally, this list of units should be equivalent to the target population as both overcoverage and undercoverage can induce bias and affect the reliability of the survey results.

Any of the sampling frames used for social surveys that cover the same population as the present survey can be used. The sampling frame of the Community survey on ICT usage in households and by individuals is an obvious candidate. The frame population however will be a subset of that survey's frame population because it must include only users of computers or devices with access to the Internet.

The types of devices and the operating systems supported by the monitoring software restrict the frame population further:

- For practical reasons the population should be restricted to users of computer, smartphones and tablets. Users of "featurephones" or of other devices with access to the Internet (e.g. game consoles, smart TVs, etc.) should not be included as monitoring software is probably not available for them.
- Not all operating systems are covered by the available monitoring software. It may not economically feasible or worthwhile to include operating systems with very small user base. As a minimum Windows and OSX (Mac) should be covered on computers and iOS and Android on smartphones and tablets.

2.3 Sampling design

In essence, the present survey is a regular social survey with a novel model of data collection. Therefore, the sampling design adopted for it can be an adaptation of the design used in some of the other social surveys, e.g. in the ICT survey. The last stage of sampling, where an individual is selected at random, must be replaced by the selection of an individual who uses a

computer, smartphone or tablet with connection to the Internet. The final sample units should be the individuals but each participating country should design its sample selection according to what is most efficient to that country.

The survey should be based on a probability sample from which results representative of the population and its demographic breakdowns, defined in the questionnaire of background variables, could be derived.

No precision requirements have been set for the results of the survey. The sampling design and the resulting sample size should be appropriate for obtaining sufficiently accurate, reliable and representative results on the survey characteristics and breakdowns. The desired accuracy of the results should be decided at national level, taking into account the proposed quarterly periodicity of the survey and the costs for its implementation.

2.3.1 Stratification

The recommendation is to use a stratified sample of individuals or households with the aim to form groups of units characterised, in relation to the variables subject of the survey, by maximum homogeneity within the groups and maximum heterogeneity between the groups. Achieving this goal in statistical terms means precision of estimates, or a reduction in sampling errors on a part with the sample quantity.

Each country should use the stratification variables according to what is most efficient to that country with particular attention to the demographic size of the localities.

The experiences from the current ICT survey regarding the effectiveness of stratification variables will be useful input for the survey in question.

2.3.2 Sample size

Calculation of sample sizes should take into account that this is a survey with multiple objectives. It has to ensure representative results for all the estimates produced. In particular, calculation of sample size should take into account that statistics have to be tabulated by age, sex, education level, employment situation, geographical location, etc.

As budgets are limited, the design of samples requires trade-offs along various dimensions. Larger samples make it possible to analyse sub-groups in depth but increase survey costs.

On the basis of the previous considerations, it is suggested to adopt a mixed view, based on both cost and organisational criteria and on an evaluation of the sample errors of the main estimates on a national level and with reference to each of the territorial domains and to each of the breakdown variables of interest.

The calculation of sample sizes should be based on precision requirements. On this basis countries should decide on sample design and calculate the sample sizes in order to receive estimates with sufficient quality and within possible budgetary constraints.

2.3.3 Weighting – Grossing up methods

Weighting factors are to be calculated taking into account in particular the probability of selection and external data relating to the distribution of the population being surveyed, where such external data are held to be sufficiently reliable.

As the sampling design used will probably differ strongly across countries, it is difficult to present ‘fit-all’ guidelines. Moreover, the weighting procedures / grossing up methods are usually determined by the sampling design used. The discussion is more of a theoretical nature and goes beyond the scope of this manual.

Unit non-response will affect this survey in the same manner as any other social survey. It will possibly be much more extensive due to the data collection mode used. Where more advanced methods for dealing with unit non-response are not feasible, it is advised to correct for unit non-response by adjusting the grossing up weights. Ideally, auxiliary information such as socio-economic differences between respondents and non-respondents should be taken into account.

If the selection of a sample from the regular sampling frame of the NSI is not feasible, e.g. because the rate of refusals is very high, the NSI may have to resort to other frames, which are not representative of the target population. For example a panel of users may be the frame. In this case, weighting may not be possible due to lack of appropriate weights valid for the population.

2.4 Survey type

2.4.1 Data collection method

As mentioned in section 1.4 the survey will combine automatic data collection for the ICT usage variables with questionnaire-based collection for the background variables.

Automatic collection is implemented with monitoring software that is installed on the devices of the sample members. The description of a possible software tool is given in the annex, in section 3.1.

NSIs should consider carefully the choice of software. Except from off-the-self tools, the option of a custom-built tool should be examined. Factors that should be taken into account include the following:

- Users should have the ability to switch the software off and on at will, so that they are in control.
- The software tools should be easy to install. Nevertheless, the NSI should be ready to provide support.
- The software tools should not be intrusive to the users of the devices. For example, they should not ask frequently for usernames or password or raise alerts when users visit specific types of website or use specific types of applications.
- The software tools must be applicable across as many types of device and operating systems as possible. This increases the comparability of the measurements.

Windows and OSX on computers and iOS and Android on smartphones and tablets should be covered with the same tool if possible.

- The software tools must not hinder the operation of the devices and must not be too resource-hungry.
- The concepts implemented in the way the software tools measure the variables of interest should match the concepts of the statistics that will be produced; alternatively the tools should be customisable enough so that the NSI can adapt them to its needs. It is therefore crucial that the way the tools work is transparent to the NSI.
- Customisability is further a requirement for technical reasons too. For example, the NSI may need itself (or contractors hired by it) to adapt the tools to different devices and operating systems.
- The technical skills available to, or affordable for the NSI should be adequate for the maintenance and deployment of the software tools.
- The cost of the tools (purchase or development and operation and maintenance costs) should be within the reach of the NSI and justifiable by the quality of the produced statistics.
- The degree of control of the NSI over the collected data. Some software tools that are offered as a service over the Internet (for example the tool presented in section 3.1) store the collected data on servers of the company that sells the software. The company can have access to the data. This is not desirable for NSIs who need full and exclusive access to the data so that they can ensure their protection from unauthorised access.

Face-to-face interviews, telephone interviews and web or postal surveys are all possible techniques of collecting the background, questionnaire-based data. Since the members of the sample will have access to devices connected to the Internet a web survey, with frequent reminders by email is the best option. The socio-demographic characteristics which can be found in registers need not be collected in the survey.

2.4.2 Independent versus embedded survey

For practical reasons, an important number, but not the majority, of countries have embedded the current ICT usage survey into an existing social survey.

The present survey however may raise concerns over privacy, due to its mode of data collection. These concerns could lead to refusals which could spill over to the ‘host’ survey too. Therefore, an independent survey is the safest option. A pilot survey, on the other hand, could be undertaken in order to examine whether embedding the survey in a current survey would affect participation to the latter.

2.4.3 Mandatory versus voluntary survey

There is no legal basis for the survey and the data collection mode will be novel to most individuals. The survey should therefore be voluntary, in its first rounds at least.

2.4.4 Coping with refusals of selected individuals to be included in the sample

Refusal of selected individuals to provide data is a common feature of social surveys. It is expected that the rate of refusal in the present survey will be higher than usual for two reasons at least: a) the lack of legal obligation to provide data, b) the use of monitoring software for automatic data collection, which creates fears for computer viruses, spyware and breach of privacy.

NSIs are accustomed to dealing with refusals in social surveys and the means they usually employ should be employed. In addition it is required to stress very strongly that the collected data will be treated like other statistical data and will be protected from un-authorised access.

Moreover, the NSI should explain in layman's terms the measures it takes to protect the data, the uses that will be made of them, the types and number of personnel that will access them and the length of time over which it will retain them. All these explanations should be included in a letter that will be given to the selected sample members.

A financial incentive could also be considered. A pilot survey carried out under contract for Eurostat offered vouchers which cost €25 per sample member. The use of such incentives is also reported in the two studies mentioned in section 1.6. The gains in sample size should be carefully counter-weighted with the potential bias that will be introduced by the use of the incentive.

2.4.5 Quality control systems

Quality control systems are of course country-specific as most statistical institutes have standard procedures and guidelines for plausibility checks or logic tests of datasets.

Such controls can be executed on-line, at the moment of the data capture by the interviewer or the data entry in the statistical institute, or after the data entry process (a program checks the data and prints the errors to be checked or corrected). The present survey may employ interviewers for the background demographic variables, if they are not compiled from population registers or collected with a web questionnaire. Most of the data will be recorded automatically by software. This affects the types of problems that may occur. These potential problems are discussed briefly below.

- **Measurement error**

There are a number of sources of measurement error: survey instruments (monitoring software, questionnaire), the respondent, the information system, the mode of data collection for the questionnaire-based variables, the interviewer.

The measurements of usage time made by the software tool may contain inaccuracies. For example, the software may be applying a threshold and record all usage times below it as equal to the threshold. The tool presented in section 3.1 for example measures each usage instance as

one minute long if it less than that. Moreover, the activities of the users may be mis-classified by type. The properties of the software must be known in detail to the producer of the statistics and possible inaccuracies must be quantified to the extent possible.

Measurement errors in the demographic variables are possible, arguably to the same extent as in any other social survey. Validation checks embedded in software in computer assisted data collection and interviewer monitoring and follow-up should be employed to the extent justified by the available resources and expected error rates.

- Invalid response

Relatively unimportant in the ICT usage survey as most answers are limited to Yes or No. However, it is possible that several items were ticked in questions where only one answer is expected.

- Relationship error

Comparing the answers across the survey can reveal inconsistencies between the answers. It is possible that e.g. an individual aged 18 indicates higher educational level, which is most probably an absolute error.

- Compulsory question left unanswered

Such an error may appear in the demographic variables only. The use of computer-assisted interviewing should avoid it. With traditional interviews or self-administered mail surveys, this error is more important.

In terms of quality of the survey *as such*, the methodology and outcomes of the survey can be benchmarked against other surveys:

- Representativeness

It can be useful to do an *ex-post* check of the representativeness of the sample, e.g. does the sample have a representative age distribution, is there some variability in the occupational and educational codes?

- Year-to-year comparison at aggregate level

Comparing the results for the current year with the previous survey can also reveal quality problems where the change is outside the range of the expected changes. For example, the share of time users spend in particular activities may decrease sharply, which could be caused by the sample members switching off the monitoring software when they carry out these activities. In such cases, it is of course possible that the problem stems from the previous survey exercise. For this purpose, it can be interesting to produce some simple tabulations of the survey results.

- Coherence or consistency with other surveys

The results can be compared with results from related survey or studies. However, in case inconsistent results are observed, it is not always easy to identify which survey gave the ‘wrong’ results.

2.5 Data processing

This chapter mainly discusses data validation and the treatment of non-response. Although the grossing-up methods can be considered as a part of the *data processing*, this topic has been discussed in section 2.3.3.

2.5.1 Data validation

The measurements of usage time as such, i.e. the recorded times, contain no errors by being automatic. The bias caused by the fact that periods of time less than one minute long are recorded as one minute long can only be estimated by simulated usage sessions. The results can be used in order to estimate correction factors.

The assignment of usage times to types of activity, e.g. type of the web sites visited, is prone to error. Software tools use lists of web sites and applications which map them to categories (e.g. Google.com is a ‘search’ web site, while IMDB.com is an ‘entertainment’ web site). Tools can be trusted to map the most commonly used sites and applications to correct categories. Less well known sites or applications may be mapped to the wrong category or, more possibly to an ‘other’ category. NSI personnel can review the listing used by the tool or the list of the sites and applications detected in the sample’s activities. Their categories they are mapped to can be changed if deemed incorrect or can be refined if deemed too aggregated.

Data validation of the demographic variables can be carried out with the means that each NSI employs in social surveys.

2.5.2 Non-response treatment

Introduction

An important source of non-sampling error in surveys is the effect of non-response on the survey results. Non-response can be defined as the failure to obtain complete measurements on the (eligible) survey sample. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response.

The latter case occurs when the interviewer was either unable to contact the respondent, no member of the household was able to provide the information, the respondent refused to participate in the survey or not enough information was collected (i.e. the response is too incomplete to be useful). This type of non-response is called unit non-response (see 2.5.3): the sample unit does not provide any of the data required by the survey. Unit non-response is generally handled by adjusting the weight of the households and/or individuals that responded to the survey to compensate for those that did not respond.

Partial non-response or item non-response (see 2.5.4) occurs when the respondents did not answer all questions because they did not understand or misinterpreted a question, refused to answer a question or could not recall the requested information. Item non-response is generally dealt with by imputation.

The present survey is mostly carried out without questionnaires. Partial non-response can take the form of hiding specific types of activity by switching off the software when engaged in them. For example, an individual may switch the software off before visiting gambling web sites. This is akin to not answering a question of the sort “Do you visit gambling web sites?”

Effect of non-response on the quality of the data

Non-response – unit as well as item non-response – can seriously affect the quality of the data collected in a survey. Firstly, the characteristics (or answering pattern) of the non-respondents can be different from those collected among the sample units who did provide eligible answers. If such difference is systematic, serious bias can be introduced in the survey results. Secondly, the reduction of the sample size (overall or for certain variables) will increase the variance of the estimates. Thirdly, non-response can have an impact on the total cost of a survey exercise. Not only because a larger initial sample may be necessary, but also because of higher unit costs of the last few percentages of respondents (due to multiple visits). Finally, non-response can be an indicator of poor overall quality of the survey and thus create an image or confidence problem.

Minimising non-response

As prevention is always better than cure, attention should be given to avoiding non-response rather than treating non-response. The use and structure of advance letters, the provision of assistance in the usage of the software, the use of incentives, the dissemination of previous results or the mandatory nature of the survey can all have an impact on the number of non-contacts or refusals.

As this issue is common to all surveys, it will not be discussed in detail in this manual.

2.5.3 Unit non-response

Introduction

Unit non-response is defined as households/persons that are included in the sample but that have not participated to the survey and for which information consequently is missing for all the variables.

Types of non-respondents include:

- Non-contact
- Refusals
- Inability to respond
- Rejected interviews
- Ineligible: out-of-scope
- Other ineligible

- Other non-response

Unit non-response can introduce bias in the survey results especially in situations in which the non-responding units are not representative of those that responded. Non-response increases both the sampling error, by decreasing the sample size, and non-sampling errors.

Weighting adjustment for unit non-response

The principal method for unit non-response adjustment is weighting. Most strategies for weighting for non-response involve dividing the respondents into a set of comprehensive and mutually exclusive groups, referred to as weighting classes. A weight is then applied to each class.

Weighting classes

In order to implement non-response adjustments, it is required to create weighting classes. It is desirable to divide the sample in "response homogeneity groups/classes". The response rates should be as homogeneous as possible within these classes and different between the classes. Data used to form these classes must be available to both non-respondents and respondents. Usually it is possible to get information about demographical (age, gender, ethnicity), geographical (urban/rural, zip code) or socio-economical (employment, income) variables from administrative data.

More advanced methods for creating weighting classes are methods like classification based on a categorical search algorithm or a logistic regression model using auxiliary variables to estimate the probability of response (cooperation in the present case).

Sample-Based Weighting Adjustment

In sample-based weighting adjustment the weight adjustment applied in each class is equal to the reciprocal of the ratio of selected sample size to respondents within the class (the inverse of the response rate within the class). This non-response adjustment factor should be multiplied with the initial base weight.

A simple example:

	Population (I)	Sample size (II)	Respondents (III)	Respondent with characteristic (IV)	Non-response adjustment Factor (V = II / III)	Initial Base Weight (I / II = VI)	Adjusted Base Weight (V*VI=VII)	Adjusted population estimate (=VIII)
Male	8 820 000	2 100	1 600	1 000	1.31	4 200	5 502	5 502 000
Female	9 020 000	2 200	1 750	1 200	1.26	4 100	5 166	6 199 200
Total	17 840 000	4 300	3 350	2 200				11 701 200

Alternative forms of sample-based weighting are that the weights are not inverse response rates but estimated coefficients of a regression model (where survey response is the left-side variable). In this case, the weights are reciprocals of the response rates estimated by the regression model.

Population-Based Weighting Adjustment

Population-based weighting adjustment requires population estimates and class membership of respondents. If there is no data available about the non-respondents, population-based adjustment still is possible since this uses external control counts for the population and not data from the sample. The method is used to correct simultaneously for both non-coverage and non-respondents. The method is used similarly to the sample-based method.

In population-based adjustment (post-stratification adjustment) the classes are created based on variables, which are known both for respondents and for the population. Weights are then applied in proportion to the ratio of population to achieved sample, so that the sums of the adjusted weights are equal to population totals for certain classes of the population.

A two-step procedure of first adjusting for non-response (sample-based adjusting) and then adjusting to known population counts is a common method that is used. However, this procedure is the same as a population-based weighting adjustment if the weighting classes in the sample-based and the population-based weighting adjustment are the same.

If the strata used in the stratification are used as classes in the weighting adjustment, there is no need for the weighting adjustment. The adjusted weighting procedure is then equal to the final grossing up/weighting procedure.

2.5.4 Item non-response

Introduction

As already mentioned above, there are several reasons for the data being unavailable. These include switching the software off for the automatically collected data, and the refusal to provide an answer, the inability to provide an answer, inadequate quality of the provided answer (e.g. implausible, incomplete, inconsistent with answers to other questions, etc.) for the questionnaire-based variables. It can be caused by either the respondent (e.g. refusal) or the interviewer (e.g. failure to record the answer adequately) but also by the survey design itself (e.g. ambiguous routing or filtering).

In case a particular individual shows too many errors, or if too many data are missing, it can be assumed that the household/individual in question has not co-operated satisfactorily in the survey. Here, the best solution is probably to remove the household/individual from the database and adjust the weighting coefficients for the other households accordingly. It is however difficult to define a threshold for the amount of missing data that will render an individual's data useless. Moreover, in order to define such a threshold it is also necessary to be able to distinguish, in the data, periods of time when the device was off from periods of time when the monitoring software was off. If the distinction cannot be made then there is no clear identification of missing data at all and no question of threshold.

2.6 Data analysis

2.6.1 Post-processing

The data may not be in the right format for computation of the indicators of interest as they are recorded by the software tool. The specific software that will be presented in section 3.1 for example, measures as web activity only the visits to web sites. The use of the ‘YouTube’ app on a mobile device for example, is not measured as web activity.

Therefore NSI staff must pore over the collected data and map applications to categories of web activity similar to the categorisation of web sites. This is achieved more efficiently if a parser extracts the names of all applications used which appear in the data, the personnel maps them to categories and the categories are inserted automatically to all occurrences of each application names. The same can be done to websites which the monitoring software has failed to assign to a particular category.

2.6.2 Computation of indicators

Section 1.5 defines the indicators that can be produced from the usage time data. The indicators are computed as estimated numbers of individuals or estimated total usage time, which are then suitably subdivided into population sub-groups, e.g. number of male users or total usage time by persons with level of education ISCED 5 or 6, and divided by appropriate estimated or known population totals.

The estimates are weighted sums of the data items. The weights can be based on selection probabilities (being their reciprocal) or can also incorporate adjustments for unit non-response or result from calibration of the data versus known population totals. The issue has been discussed in section 2.3.3.

2.6.3 Estimation of the accuracy of the indicators

The accuracy of an indicator is estimated by its standard error (the square root of the variance). The estimation of the sampling variance should take into account the sampling design (e.g. the stratification).

The indicators should be treated by the NSI just like other indicators from social surveys. Each individual has provided data (the usage times and demographic variables) and has a corresponding weight assigned to him / her.

2.7 Confidentiality and privacy issues

The automatic recording of Internet and application usage data is bound to raise concerns from the sample members and by civic society in general. Moreover, the recorded data are personal information and there is national and European legal context to which their collection and processing should abide.

Clearly, the implementation of the survey and the production of statistics should be well inside the limits of the law. The NSIs cannot risk damaging their credential or jeopardizing the public’s trust in them and its willingness to participate in social surveys in the future.

As a first step, the survey and the processes should be transparent to the sample members. The NSI should inform the potential sample members of: a) the compulsory or optional nature of the survey and its legal basis, if any, b) the name and position of the person or body in charge of the survey, c) the purpose of the survey, d) the categories of data collected and processed, e) the statistics that will be produced, f) the fact that the data will be kept confidential and used exclusively for statistical purposes, g) the guarantees to ensure the confidentiality and the protection of personal data, h) the categories of persons or bodies to whom the personal data may be communicated, i) the way in which consent can be refused or withdrawn and, in the case of compulsory surveys, the possible sanctions this would entail, g) where applicable, the conditions of the exercise of the rights of access and rectification. The individuals should also be informed about the possibility of obtaining further information on request.

The selected sample members should give their explicit consent (what is termed “opt-in”) to be included in the survey. The indication by which they signify their agreement must leave no room for ambiguity regarding their intent.

Care should be taken in that explicit consent is not sufficient for some types of personal data. These types include data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, health or sex life. The recording of the particular, e.g. health-related, web sites visited may be considered as revealing such particular characteristics of individuals. Therefore, the NSIs’ legal department must examine carefully whether the recording of such sites abides to the specific national laws or whether the data that will be recorded must be tweaked appropriately.

Moreover, the software should provide to users the ability to switch it off and on with ease. This gives individuals a means of temporarily withdrawing and reaffirming consent and increases their trust in the NSI.

3 Annexes

3.1 Software tools

The software selected for monitoring and recording the users' activities is the online parental control service Qustodio⁵. We first explain its regular usage for monitoring children's activities online as this helps explain its deployment for recording usage times.

A parent signs in to the service and defines one "child" for each combination of child, device and user account on the device. For example, if a household has a desktop and a laptop PC and the two children have one account each on each device, their father will define four "children" in the service. This enables individualised monitoring. The parent uses separate passwords per "child" so as to activate monitoring. A second form of usage, suitable for schools that want to control activity over the school's computers, is to have a single administrator user name and password. The administrator then defines "children" for all the devices used in the premises of the school.

For each child Qustodio provides a separate web page, with aggregated statistics. The aggregation period can be modified, spanning one, seven, 15 or 30 days. A sample of results is shown in Box 1.

Box 1. Sample of Qustodio results for a single "child" and a single 24-hour period.

1181_pc.htm ← This is the "child"

62.5% Using Microsoft Office Word Using Microsoft Office Word

18.8% Surf on Search Portal websites Surf on Search Portal websites

9.4% Surf on Social Network websites Surf on Social Network websites

3.1% Surf on Shopping websites Surf on Shopping websites

3.1% Surf on Webmail websites Surf on Webmail websites

3.1% Using DUPLEX Using DUPLEX

0:59 Total usage time during specified period

0:08 Hours of Web activity

0:00 Hours of Social activity

0:56 Hours of Apps usage

A drawback, which unfortunately affects most monitoring software tools, is that Qustodio does not run on iPhones or iPad.

⁵ www.qustodio.com

Two options are available for the installation of the software to the devices of the sample members:

1. Installation by the sample members themselves: detailed instructions must be sent to each of the sample units indicating the device on which they have to install Qustodio and the way of installing it. They must be instructed to choose a specific “child” name so that the NSI can assign appropriately the results of monitoring. The approach requires a screening phase, during the recruitment of the sample, during which the contacted sample units list the computers and mobile devices with access to the Internet which they use regularly.
2. Generation of accounts on the users’ behalf: the NSI can generate accounts and email the sample units with the device, child name along and instructions (less than in the first option) about the installation.

The NSI must offer support to the members of the sample. A help desk must be set up the contact details of which must be given to all members of the sample. It must be expected that requests for assistance will be quite numerous.

3.2 Model questionnaire

Background demographic variables

1. Date of birth (dd/mm/yyyy): _____
2. Gender
 - a. Male
 - b. Female
3. In what country were you born?
 - a. [Country of the survey]
 - b. Other country of the European Union
 - c. Other country, outside the European Union
4. Citizenship
 - a. [Citizenship of the country of the survey]
 - b. Citizenship of another country of the European Union
 - c. Citizenship of another country, outside the European Union
5. Completed level of education
 - a. Has not attended / completed primary education (ISCED 0)
 - b. Primary education (ISCED 1)
 - c. Lower secondary education (ISCED 2)
 - d. Upper secondary education (ISCED 3)
 - e. Post-secondary non-tertiary education (ISCED 4)
 - f. Short-cycle tertiary education (ISCED 5)
 - g. Bachelor’s or equivalent level (ISCED 6)
 - h. Master’s or equivalent level (ISCED 7)
 - i. Doctoral or equivalent level (ISCED 8)
6. Main occupation
 - a. Employee
 - b. Self-employed
 - c. Unemployed
 - d. Pupil, student, not in labour force

- e. Other case not in labour force (housewife, soldier, retired, rentier, unable to work etc.)
7. Region of residence: _____

3.3 Transmission format

The NSIs will deliver to Eurostat the flat, comma-separated file, with the collected micro-data that result from the post-processing of the collected data. Each record of the file will contain data about one calendar day, one individual member of the sample and one device of this individual.

The fields of the file and the format of each one will be the following:

- Survey identifier: a string identifying the survey, to be agreed between Eurostat and the NSIs, common to all records.
- Reference period: a character string of format YYYYQA (where A=1, 2, 3, 4 denotes the quarter), which shows the reference period of the data. It is common to all records.
- User ID: a unique identification number for each individual. Format: 8-digit number.
- Device: the type of device. It can take the following values:
 - 111: a desktop or laptop computer running Windows
 - 112: a desktop or laptop computer running OSX
 - 113: a desktop or laptop computer running Linux
 - 211: Windows smartphone
 - 212: iPhone
 - 213: Android smartphone
 - 311: Windows tablet
 - 312: iPad
 - 313: Android tablet
 - 888: other mobile device (e.g. Blackberry smartphone)
 - 999: unknown / not recorded
- Gender. It can take the following values:
 - 1: male
 - 2: female
 - 9: unknown / not recorded
- Age: the age of the individual, in integer year, at last birthday. Format: three digit number, no decimal digits.
 - 999: age unknown / not recorded (← meaning the date of birth was not recorded)
- Country of birth. It can take the following values:
 - 1: the country of the NSI
 - 2: other EU country
 - 3: non-EU country
 - 9: unknown / not recorded
- Citizenship. It can take the following values:
 - 1: the country of the NSI
 - 2: other EU country

- 3: non-EU country
 - 9: unknown / not recorded
- Level of educational attainment
 - 0: ISCED 0
 - 1: ISCED 1
 - 2: ISCED 2
 - 3: ISCED 3
 - 4: ISCED 4
 - 5: ISCED 5
 - 6: ISCED 6
 - 7: ISCED 7
 - 8: ISCED 8
 - 9: unknown / not recorded
- Employment status
 - 1: Employee
 - 2: Self-employed
 - 3: Unemployed
 - 7: Pupil, student, not in labour force
 - 8: Other case not in labour force (housewife, soldier, retired, rentier, unable to work etc.)
 - 9: unknown / not recorded
- Region of residence: the NUTS level 2 code of the region of residence of the individual.
- Date: the date to which the data refer. Format: dd/mm/yyyy.
 - 99/99/9999: unknown / not recorded
- Day of the week: the day to which the date refers. It can take the following values:
 - 1: Monday
 - 2: Tuesday
 - 3: Wednesday
 - 4: Thursday
 - 5: Friday
 - 6: Saturday
 - 7: Sunday
 - 9: date missing
- Design based weight: the weight assigned to the individual based on the sample design only, as if no frame imperfections and no non-response exist. The format is 10-digit number with 2 decimal digits.
- Final weight: the weight assigned to the individual after adjustments for frame imperfections non-response and other possible calibration of weights. The format is 10-digit number with 2 decimal digits.

{The following fields record time, measured in minutes, spent on specific activities, those indicated by each field's name. Since they refer to one calendar day they can take any integer value between 0 and 24*60=1440. 0 and 1440 are allowed. Since activities may be carried out in parallel, the sum of the values across on record can be greater than 1440.}

- Total time: total usage time, i.e. time that the computer was active, even if not on the Internet.
- Web time: time spent visiting websites.

- Social time: time spent on Facebook activities.
- Apps time: time spent using applications, even offline. *For the usage time fields listed until here please refer also to section 1.6.*
- Using cloud storage facilities
- Doing an online course (in any subject)
- Education activity, other: time spent on online activities / web sites related to education, but not to doing an online course, e.g. searching for information about courses.
- Email
- Employment: time spent on online activities / web sites related to employment.
- Entertainment
- Finding information about goods or services
- Forums
- Gambling
- Games, unspecified
- Government: time spent on government web sites.
- Listening to web radio
- Networked games
- Participating in social networks
- Playing online, but not networked games
- Adult content
- Reading news
- Shopping
- Sports
- Technology
- Telephoning / video calling (via webcam) over the internet
- Using services related to travel or travel related accommodation
- Viewing / listening to online images, videos, music
- Internet, other: time spent on online activities / web sites that cannot be classified in one of the other categories.
- Offline: time spent on offline activities.
- Not clear: time spent using applications for which it cannot be distinguished whether they involve online activity or not.

Part 2 - Statistics on the facilities of business websites

1 Statistical product

This chapter describes the statistical information to be produced, which is separate from the production methodology.

The elements that make up the statistical product, at an input level, are the statistical unit, the target population and the observation variables, and at the output level, the periodicity and the summary measures, aggregate variables and tabulation. Covering all the elements of the statistical product, the statistical concepts and the nomenclatures are also needed to assure harmonization and comparability of statistics.

1.1 Statistical unit

The statistics on the facilities of business websites have the enterprise as the statistical unit. This is the unit that we want to observe or analyse. Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community defines it as follows:

"The enterprise is the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. An enterprise carries out one or more activities at one or more locations. An enterprise may be a sole legal unit."

The enterprise thus defined is an economic entity which can therefore, under certain circumstances, correspond to a grouping of several legal units. Some legal units, in fact, perform activities exclusively for other legal units and their existence can only be explained by administrative factors (e. g. tax reasons), without them being of any economic significance. A large proportion of the legal units with no persons employed also belong to this category. In many cases, the activities of these legal units should be seen as ancillary activities of the parent legal unit they serve, to which they belong and to which they must be attached to form an enterprise used for economic analysis.

1.2 Target population

The *target population* of the statistics on the facilities of business websites is the group of enterprises which have a company web site and are delimited by the following attributes:

- **Economic activity**

Enterprises classified in the following categories of NACE Rev. 2:

- Section C – "Manufacturing";
- Section D, E – "Electricity, gas and steam, water supply, sewerage and waste management";
- Section F – "Construction";
- Section G – "Wholesale and retail trade; repair of motor vehicles, motorcycles;
- Section H – "Transportation and storage";

- Section I – "Accommodation and food service activities";
- Section J – "Information and communication";
- Section L – "Real estate activities";
- Division 69-74 – "Professional, scientific and technical activities";
- Section N – "Administrative and support activities";
- Group 95.1 – "Repair of computers"; and
- Classes/groups 64.19 + 64.92 + 65.1 + 65.2 + 66.12 + 66.19 – "Financial and insurance activities".

The enterprises are classified in one of these categories according to their **principal** activity.

▪ **Enterprise size**

Enterprises with 10 or more persons employed;

Please note that the number of persons **employed** is defined in Commission Regulation (EC) No 250/2009 of 11 March 2009 (p.38-39, Code: 16 11 0; Number of persons employed) and should not be confused with the number of **employees** or with **FTE's**.

([http://eur-](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:086:0001:0169:en:PDF)

[lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:086:0001:0169:en:PDF](http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:086:0001:0169:en:PDF))

▪ **Geographic scope**

Enterprises located in any part of the territory of the Country.

The *frame population* is an operationalization of the target population, taking the form of a list of elements of the target population. Although a target population can be easily defined, in practice a list of all its elements is needed for its complete or partial (in case a sample is used) observation, and that can be very difficult to obtain. That list should be complete and include every element of the target population only once. However, most of the time, it will suffer from both under-coverage and over-coverage. The frame population will be further explained in the chapter 2.2.

1.3 Periodicity

The periodicity is quarterly, meaning the data are collected and compiled once per quarter.

A quarterly survey has been made possible by the employment of the data collection method presented in section 2.4. The quarterly frequency is appropriate in view of the need for relevant and recent information on a “fast moving” study domain like the information society.

1.4 Observation variables

The compiled raw data are lists of URLs. They are the addresses of pages inside a business web site (e.g. “[www.\[enterprise-name\].com/services](http://www.[enterprise-name].com/services)”) in which a certain target functionality has been detected. The list of URLs is organised into groups, one per target functionality.

These lists are straightforwardly converted into data on binary variables (“Yes / No”), that refer to the business web site as a whole (e.g. to [www.\[enterprise-name\].com](http://www.[enterprise-name].com)). One variable is defined for each target functionality and takes value “Yes” if the functionality has been detected in at least one page of the web site. In other words, if there is a group of URLs for the

functionality, the corresponding target variable takes value “Yes”; if there is not a single URL for it, the variable takes value “No”.

1.5 Summary measures, aggregated variables, indicators and tabulation

One aggregated variable and one indicator, for each target functionality, will be produced from the collected data on web site functionalities:

- The aggregated variable is the number of business web sites that offer a given functionality
- The indicator is the share of business web sites that offer this functionality

The indicator is computed by dividing the aggregated variable with the total number of business web sites.

The number of indicators therefore depends on the size of the set of functionalities that it is of interest to detect.

The reference period of the statistics is the observation period, i.e. the period of time during which the crawler software visited the web sites and recorded their contents.

The aggregated variables and indicators can also be computed for specific sub-populations defined by background characteristics of the enterprises, such as economic activity, size, location, etc.

Section 2.5 provides more details about the computation of the aggregated variable and the indicator.

1.6 Explanatory notes

The term “functionality” is used in this guide in order to denote types of information provided by the site, features of the site or facilities that it offers to visitors. For example, three “functionalities” are:

- the availability of enterprise contact information on the site [information];
- the availability of the site in English, besides the national language [feature];
- the availability of user registration facility so that visitors can then have personalised content [facility].

A possible set of functionalities, proposed in project “Internet as a data source”, with their definitions, are given in the following table.

Table 1. List of web site functionalities proposed as “targets” of the survey.

Functionality	Definition	Comments
Contact information -	The site lists a URL (web address) among the contact information that it provides to visitors; this may or may	

D6 – Cookbook for the implementation of new methods and indicators at national level

Functionality	Definition	Comments
URL	not be the same as the main URL of the site	
Contact information - Email address	The site lists an email address among the contact information that it provides to visitors	
Contact information - Telephone number	The site lists a telephone number among the contact information that it provides to visitors	
Contact information - Postal address	The site lists a postal address among the contact information that it provides to visitors	
Availability of the web site in the national language	At least one of the pages of the web site is provided in the national language.	
Availability of the web site in English	At least one of the pages of the web site is provided in English.	
Availability of "last updated" date	The site lists the date on which it was last updated.	
Availability of privacy policy	The site displays (or provides a link to a document containing) the privacy policy of the site. This is a description of the use of personal information - particularly personal information collected via the website - by the website owner. It also describes measures taken to guarantee secure handling of financial information.	Relevant indicator produced from the regular ICT survey too.
Availability of registration facility	The web site has facility for users to sign up and then sign in.	
Availability of personalised content for regular/repeated visitors	The web site has the ability to recognise the user from previous visits (login/password) and adapt the content of the pages accordingly.	Relevant indicator produced from the regular ICT survey too.
Availability of site map	A site map is a list of pages of the web site accessible to crawlers or users. It can be either a document in any form, or a web page that lists the pages, typically organized in hierarchical fashion.	

D6 – Cookbook for the implementation of new methods and indicators at national level

Functionality	Definition	Comments
Display of the number of visitors	At least one page of the web site displays the number of visitors since a - listed too - given point in time.	
Availability of product catalogues	The web site provides lists of products or services offered by the enterprise to its clients. They might include also the characteristics of these products or services. The information may be static or dynamic (extracted online from a database and as such always updated).	Relevant indicator produced from the regular ICT survey too.
Availability of price lists	The web site provides provides a product catalogue which includes prices.	Not common for certain types of enterprises, e.g. in the services sector. Relevant indicator produced from the regular ICT survey too.
Possibility for site visitors to customise or design the products	The web site provides an interactive interface where users can choose from several possible characteristics of the products (colour etc.) or services and see online in the site the impact, for instance, on the price. The interface might also include the possibility for the user to visualise the appearance of the product with the options that were selected. The carrying out of simulations or any calculations (e.g. what-if calculations) for products like loans in the financial sector, belongs here as well.	Relevant indicator produced from the regular ICT survey too.
Availability of online ordering or reservation or booking facility	The web site provides a facility which allows the user to order products or services with no additional contact offline or via e-mail required (for the ordering). A shopping cart and checkout facility is such an example. It includes also the facility for reservation of hotel rooms or the booking of flights. It does not include a link in the website which directs the user to an e-mail application which requires the user to send the order via e-mail. Payment may or may not be included in the ordering facility, e.g. payment may be made on reception of the product or by other	Relevant indicator produced from the regular ICT survey too.

Functionality	Definition	Comments
	<p>means other than electronic payment.</p> <p>Carrying out a transaction via online banking in general does not qualify as online ordering; specific cases however, e.g. when buying shares (with a commission to be paid to the bank), qualify as online orders in the banking sector.</p>	
Availability of online order tracking facility	The web site provides facility that aims to keep the customer informed on the progress of the ordering and delivery process.	Relevant indicator produced from the regular ICT survey too.
Listing of open job positions or availability of online job application facility	This item includes both cases where just simple information on job vacancies is provided in the web site as well as those where the site provides also an online facility for candidates to apply for the jobs.	Relevant indicator produced from the regular ICT survey too.
Number of open job positions in the enterprise, listed in the web site	The number of job opening listed in the web site.	
Availability of links to multimedia content (audio, videos, etc)	The web site provides links to multimedia content hosted in the servers of the enterprise.	
Availability of links to content in multimedia sharing sites (YouTube, Flickr, etc)	The web site provides links to multimedia content hosted in multimedia sharing sites.	
Availability of links to social networks or blogs (Facebook, LinkedIn, Yammer, Twitter, etc)	The web site provides links to social networks or blogs.	
Availability of links to wikis and wiki-based sharing tools	The web site provides links to wikis and wiki-based sharing tools.	

It must be stressed that the collected data need not include the content of the web site's pages; e.g. they need not include the contact information listed in the site. They must include the information of whether the content of interest exists or not; e.g. whether the site provides contact information.

Eight of these indicators, as stated in the comments of the table, have corresponding indicators from the regular ICT survey. The statistics about them compiled with the present survey can be cross-checked versus those of the regular survey. Moreover, estimated variances from the regular ICT survey can be used to specify the allocation of the present survey's sample to strata (see section 2.3.2).

2 Production methodology

2.1 Timetable – Survey period

It is recommended that the crawling of all selected web sites takes place in as short a period of time as possible so that the data are comparable. The continuous increase in computing power and improvement of available software mean that the extraction and storage of data are getting faster too. One week should be sufficient for a few thousands of enterprises and their sites.

Moreover, the tools may be deployed off regular working hours, when visits to the web sites by regular users can be expected to be very few or nil.

2.2 Frame population

The survey will be carried out in the form of a sample survey. The *frame population* (or *sampling population*) is the list of enterprises from which the sample will be drawn. Ideally, this list of enterprises should be equivalent to the target population as both over-coverage and under-coverage can induce bias and affect the reliability of the survey results.

The sample for the survey should be drawn from the business register in the different Member-States as defined in Council Regulation (EEC) No 2186/93. Part of this register is the activity code at the four-digit (class) level of NACE Rev. 2 and the size measured by the number of persons employed of the enterprises.

2.2.1 Updating the Business Register with website information

An important issue is whether the register contains the addresses of the enterprises' web sites. These are the only practical contact information needed for the type of survey described in this guide. This information must also be of good quality. There should be high degree of trust in the fact that enterprises listed without URLs in the register do not have a web site. Nevertheless, many business registers do not contain yet URLs. NSIs should try to add them by asking enterprises, in the context of the ICT or other business surveys, to provide their URL. They should moreover include URLs in their regular register maintenance procedures.

During the time required for enriching the register with URL information, the NSIs could take a sub-sample of the most recent ICT survey as the sample of the present survey. This will also allow them to cross-check the statistics compiled from crawler data with the corresponding regular ICT statistics.

2.3 Sampling design

In essence, the present survey is a regular business survey with a novel model of data collection. Therefore, the sampling design adopted for it can be an adaptation of the design used in some of the other business surveys, e.g. in the ICT survey. It should therefore be based on a probability sample from which results representative of the population could be derived.

No precision requirements have been set for the results of the survey. The sampling design and the resulting sample size should be appropriate for obtaining accurate, reliable and representative results on the survey characteristics and breakdowns. The desired accuracy of the results should be decided at national level, taking into account the proposed quarterly periodicity of the survey and the costs for its implementation.

2.3.1 Stratification

The recommendation is to use a stratified sample of enterprises with the aim to form groups of units characterised, in relation to the variables collected in the survey, by maximum homogeneity within the group and maximum heterogeneity between the groups.

The economic activity (in terms of NACE) and size (in terms of the number of persons employed) of the enterprise should be used for the stratification of the sample. This information, according to the Council Regulation (EEC) N° 2186/93 on business registers for statistical purposes, is present in the sampling frame and can, therefore, be used to stratify the sample *à priori*.

The purpose of the stratification by main economic activity and size class is to assure *à priori*, accurate results for breakdowns according to them. In fact, if the sample is not stratified by these variables, the number of enterprises which casually end up in some NACE category, size class, or region might be too small to produce accurate results.

For the definition of the categories and level of detail of the stratification variables, the desired level of dissemination concerning NACE-aggregates and size-classes has to be taken into account. The stratification of the frame population has to be at least as detailed as this level of dissemination.

2.3.2 Sample size

Calculation of sample sizes should take into account that this is a survey with multiple objectives. It has to ensure representative results for all the estimates produced. In particular, calculation of sample size should take into account that each statistic has to be tabulated by NACE category and size class.

As budgets are limited, the design of samples requires trade-offs along various dimensions. Larger samples make it possible to analyse sub-groups in depth but increase survey costs. Depending on the type of crawler that will be used, the marginal increase of cost with each additional sample unit will range from negligible (when a generic crawler is used) to very high (when the crawler requires customization for each different web site).

On the basis of the previous considerations, it is suggested to adopt a mixed view, based on both cost and organisational criteria and on an evaluation of the sample errors of the main estimates on a national level and with reference to each of the territorial domains and to each of the breakdown variables of interest.

The calculation of sample sizes should be based on precision requirements. On this basis countries should decide on sample design and calculate the sample sizes in order to receive estimates with sufficient accuracy and within possible budgetary constraints.

In practice, the sample size is usually calculated by applying the desirable overall reliability of the estimate to a target-variable. This target variable can be one of special relevance for the survey or one that correlates well with the majority of the variables to be collected. The resulting sample size is set by the dispersion of this target-variable. However, some times for several reasons, e.g. so as not to exceed a given administrative burden of enterprises, a maximum number of enterprises to be surveyed is defined. This number of enterprises is allocated to the different strata in such a way that the reliability of the estimates is optimized. An efficient way to allocate a specified number of enterprises to the different strata is the so-called Neyman-allocation, meaning that the number of enterprises is allocated to the relevant strata in proportion to the variance of a specified target-variable in these strata.

$$n_h = n \times \frac{N_h \times S_h}{\sum N_h \times S_h}$$

Where: n_h is the number of units in the sample in stratum h ;

n is total sample size;

N_h total number of units in the frame population for stratum h ;

S_h true standard deviation in stratum h for the relevant variable.

Estimates of the variance of the target variable might come from the survey from a previous year. If the survey contemplated is the first one of its kind, the variance estimates of corresponding indicators from the regular ICT survey (see section 1.6) may be used.

Additional to the outcome of the Neyman-allocation, a minimum number of enterprises in each stratum can be specified. For larger enterprises one can decide to include them integrally in the survey. However, for qualitative data like those collected in the present survey this is not crucial. More advanced sampling techniques may be used as long as it is possible to calculate the indicators specified in this guide.

By specifying a maximum number of enterprises in the sample it is useful to anticipate - based on experience with a previous survey or another comparable survey - a response rate. If experience shows that only 50 percent of the enterprises addressed actually participate in a survey, the sample size should be adapted to this response rate, meaning that it should be doubled.

2.3.3 Weighting – Grossing up methods

The grossing up method, or weighting procedure, to be adopted for the production of figures for the total target population is determined in the first place by the sampling design used. The weighting factors are calculated taking into account in particular the probability of selection of each unit in the sample.

In this chapter, the explanation of weighting will assume the selection of a stratified random sample, which is the method recommended in this guide. The formulas have to be modified if a different sampling design is used.

In the second place, the grossing up method is determined by the type of variables collected and the statistics produced with those variables. In the present survey all variables are binary, as stated in section 1.4. and results will be published as percentages of the number of enterprises. To produce these results the observations are **weighted by the number of enterprises** in the stratum to which they belong.

The regular ICT survey also employs weighting by the number of persons employed because the majority of the labour force works in bigger enterprises, where ICT usage is qualitatively and quantitatively different from the others. However, the sophistication of business web sites is not a ‘privilege’ of large enterprises; on the contrary several innovative small firms have very modern web sites and in fact use them to carry out many of their business activities, like marketing or sales. Therefore this type of weighting will not be presented in this guide.

Basic weighting by number of enterprises

Assuming that a stratified random sampling is used, the estimator of a total in the population based on the sample is:

$$Y = \sum Y_h, \quad (1)$$

$$Y_h = \frac{N_h}{n_h} \sum_{i \in h} y_{hi}, \quad (2)$$

where:

Y is the estimated total value of variable y for the total population

Y_h is the estimated value of variable y for the total population in stratum h ;

☐ total number of units in the frame population for stratum h ;

n_h is the number of units in the sample in stratum h ;

☐ is the value of variable y of enterprise i in stratum h . If its value is “YES” then it assumes the value 1 in the formula. If its value is “NO” it assumes value 0. This way the total of this variable is the number of enterprises whose web site has the target functionality to which this variable refers.

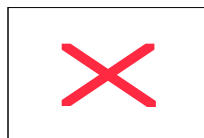
To compute the percentages, these totals are divided by the total number of enterprises.

For the total population:



(3)

For each stratum:



(4)

In the grossing up, each enterprise i in strata h has the following weight

$$w_{hi} = \frac{N_h}{n_h}, \quad (5)$$

which gives how many enterprises in the population this sampled unit represents.

2.4 Survey type

2.4.1 Data collection method

As mentioned in section 1.4 the survey will rely on crawler software that collects data automatically from the web sites of the enterprises of the sample.

The software can detect two types of information:

- The text that forms the content of the web site and the text that comprises the computer-code of the web site.
- The technologies implemented by the web site.

Not all tools detect both types of information.

The NSI must create a mapping between the target functionalities and possible “keywords” in the text which show the presence of the functionalities. Such a mapping is shown in the annex, in section 3.2. Similarly the NSI defines a mapping between functionalities and technologies adopted by the web site.

The crawler then visits all pages of the web sites and either analyses the content and technologies on the spot or extracts the content and stores it locally (in the NSI’s servers) together with information about the technologies detected.

At some point this information is analysed and a list of URLs is created for each target functionality: it contains those URLs where at least one keyword and / or technology associated with this functionality has been detected.

The description of a possible software tool is given in the annex, in section 3.1.

NSIs should consider carefully the choice of software. Except from off-the-self tools, the option of a custom-built tool should be examined. Factors that should be taken into account include the following:

- There are several different technologies used in the design and operation of a web site. One basic distinction is between “static” and “dynamic” web sites. In the latter type content is “built” dynamically and most crawler software tools cannot see any of it. The tool(s) adopted by the NSI should be able to work with as many as possible types of site with as little required customization for each type as is feasible.
- The information extraction by the software tools should not be so intensive as to limit in any perceptible way the ability of the web sites to function properly. Moreover, the tools may be deployed off regular working hours, when visits to the web sites by regular users can be expected to be very few or nil.
- The tools should be customisable enough so that the NSI can adapt them and make the collected data match, in terms of definitions, its needs.
- Customisability is also a requirement for technical reasons. For example, the NSI may need itself (or contractors hired by it) to adapt the tools to different types of web sites.
- The technical skills available to, or affordable for the NSI should be adequate for the maintenance and deployment of the software tools.
- The cost of the tools (purchase or development and operation and maintenance costs) should be within the reach of the NSI and justifiable by the quality of the produced statistics.
- The data collected by the tools must be stored in local servers of the NSI and not on some cloud-based servers provided by a third party. In this way the NSIs can ensure their protection from un-authorised access.

2.4.2 Independent versus embedded survey

The survey may raise concerns about the protection of confidential business data, due to its mode of data collection. These concerns could lead to refusals which could spill over to the ‘host’ survey too. Therefore, an independent survey is the safest option. A pilot survey, on the other hand, could be undertaken in order to examine whether embedding the survey in a current survey would affect participation to the latter.

2.4.3 Mandatory survey versus voluntary survey

There is no legal basis for the survey and the data collection mode will be novel to most site owners. The survey should therefore be voluntary, in its first rounds at least.

2.4.4 Contact person of the survey

In most cases the IT manager is the appropriate person to contact about the survey. However, not all small enterprises have an IT manager; in these cases either the owner or the general administrator should be contacted. If a contact person of the company is listed in the business register it is useful to contact this person about the survey.

2.4.5 Coping with refusals of selected enterprises to be included in the sample

The refusal of selected enterprises to allow access to their web sites is an eventuality that cannot be ignored. It is expected that the rate of refusal in the present survey will be higher than usual for two reasons at least: a) the lack of legal obligation to provide data, b) the use of crawler software for automatic data collection, which might create fears for denial-of-service attacks and breach of privacy.

NSIs are accustomed to dealing with refusals in business surveys and the means they usually employ should be employed. In addition it is required to stress very strongly that the collected data will be treated like other statistical data and will be protected from un-authorised access.

Moreover, the NSI should explain in layman's terms the measures it takes to protect the data, the uses that will be made of them, the types and number of personnel that will access them and the length of time over which it will retain them. All these explanations should be included in a letter that will be given to the selected enterprises.

The provision of incentives in kind could also be considered. Each enterprise for example could be offered a custom-made report which will contrast its situation with the statistics of the whole target population and of the population of enterprises in its stratum. The topic of the survey however may be of little interest for a substantial subset of the sample. The offered report therefore could refer to a spectrum of business statistics beyond ICT ones; in this case it would not be comparative, unless there are relevant data about the selected enterprise, but it can be a sector report.

2.4.6 Quality control systems

Quality control systems are of course country-specific as most statistical institutes have standard procedures and guidelines for plausibility checks or logic tests of datasets.

Some of the most common errors or problems are briefly discussed below.

- **Measurement error**

One potential source of measurement error exists in the survey: the crawler software. As it was explained in section 2.4.1 the collection of data relies on the detection of keywords and / or of certain technologies in the web sites content and computer code respectively.

The sensitivity and specificity of the mapping between functionalities and keywords / technologies defines the possible extent of measurement errors. There will be cases where the functionality is present but none of the keywords or technologies associated with it are detected; this means the mapping is not sensitive enough for this functionality and needs additional elements. There will also be cases where some keywords or technologies are present

but not the functionality: this means the mapping is not specific enough and alternative elements are needed.

The way to detect such errors is to have human operators inspect “manually” a subset of the samples web sites and record discrepancies between their findings and the tool’s findings concerning the functionalities. These discrepancies can be used to estimate adjustment factors for the survey’s results and, more importantly, can guide the improvement of the mapping for future applications.

- Representativeness

It can be useful to do an *ex-post* check of the representativeness of the sample, e.g. does the sample have a representative size class distribution, is there some variability in the economic activities?

- Year-to-year comparison at aggregate level

Comparing the results for the current year with the previous survey can also reveal quality problems where the growth is outside the range of the expected changes. For example the share of web sites with a particular functionality may decrease sharply, which may be caused by errors in the keyword and technology mapping. In such cases, it is of course possible that the problem stems from the previous survey exercise. For this purpose, it can be interesting to produce some simple tabulation of the survey results.

- Coherence or consistency with other surveys

The results can be compared with results from related survey or studies. However, in case inconsistent results are observed, it is not always easy to identify which survey gave the ‘wrong’ results.

2.5 Data processing

This chapter mainly discusses the treatment of misclassification and of non-response. Although the grossing-up methods can be considered as a part of the data processing, this topic is discussed above has been discussed in section 2.3.3. Moreover, all data being collected automatically, the only errors that can occur are due to the not sensitive or specific enough mapping between functionalities and keywords / technologies. The checks that can be implemented were described in section 2.4.6, right above.

2.5.1 Misclassification treatment

Misclassification occurs when an enterprise is included in the survey, because according to register data used for stratification it belongs to a size class and sector of activity covered by the survey but in reality it should not have been included. In other cases misclassification that enterprises should have been classified in a size class or NACE category different than those in which they belong according to the register. The misclassification will then possibly mean that the enterprises should belong to a different stratum than the one used for stratification.

Such a situation can arise due to frame population imperfections. Frame imperfections can occur when there is a time lag between the actual situation for an enterprise and the information

available in the registers. It often takes a certain period of time to update register information after a change in the number of employed persons or a change of sector of activity has occurred.

This time lag in updating register information implies that there is a difference between the target population (i.e. the population that the survey intends to cover) and the frame population (i.e. the population that the survey actually covers based on information available in registers).

Recommendation in case of misclassification of enterprises

The first issue is to detect misclassification. Since no data relevant to economic activity or size class are collected possible misclassification must be detected before data collection. This will require to contact the appropriate contact person by telephone and, together with informing them about the survey, verify the economic activity and size indicated in the register.

If it turns out that the enterprise should belong to a different stratum, and since the sample size allocation cannot be repeated on that time, new strata should be built and the weights used in computations should be changed accordingly.

If it turns out that the enterprise is outside the target population (e.g. it is too small) it should not simply be excluded from the sample. Such an approach could be hazardous as correction then only is made for enterprises that fall beyond the cut-off limit (e.g. 10 persons employed) and not for enterprises that fell beyond the limit according to register information and that during the reference period of the survey exceeded it. A more appropriate approach is in those cases to assume that enterprises where the number of persons employed has decreased below the cut-off limit offset the enterprises that have increased in number of employed persons and that exceed the cut-off limit. Enterprises falling below the cut-off limit are then treated as respondents and not as over-coverage.

2.5.2 Non-response treatment

Introduction

An important source of non-sampling error in surveys is the effect of non-response on the survey results. Non-response can be defined as the failure to obtain complete measurements on the (eligible) survey sample. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response.

In the present survey there are no questions. Therefore, partial non-response can only occur if, “unknown” to the crawler software, the web site employs some technical protection means which forbid the software from detecting some types of functionality or obtaining a subset of its content.

Total non-response occurs when the enterprise refused to participate in the survey or when the technical means employed forbid the collection of any information by the crawler. This type of non-response is called unit non-response (see section 2.5.3): the sample unit does not provide any of the data required by the survey. Unit non-response is generally handled by adjusting the weight of the enterprises from which data were collected to compensate for the rest.

Effect of non-response on the quality of the data

Non-response (unit as well as item non-response) can seriously affect the quality of the data collected in a survey. Firstly, the characteristics (or answering pattern) of the non-respondents can be different from those collected among the sample units from which data were collected. If such difference is systematic, serious bias can be introduced in the survey results. Secondly, the reduction of the sample size (overall or for certain variables) will increase the variance of the estimates. Thirdly, non-response can have an impact on the total cost of a survey exercise. Not only because a larger initial sample may be necessary, but also because of higher unit costs of the last few percentages of respondents (due to sending of reminders or repeated telephone calls). Finally, non-response can be an indicator of poor overall quality of the survey and thus create an image or confidence problem.

Minimising non-response

As prevention is always better than cure, attention should be given to avoiding non-response rather than treating non-response. The number (and timing) of reminder letters or call backs, the use and structure of advance letters, the dissemination of previous results or the mandatory nature of the survey can all have an impact on the number of non-contacts or refusals.

As this issue is common to all surveys, it will not be discussed in detail in this manual.

2.5.3 Unit non-response

Introduction

Unit non-response is defined as enterprises that are included in the sample but that have not participated in the survey and for which information consequently is missing for all the variables.

Weighting adjustment for unit non-response

The principal method for unit non-response adjustment is weighting. Most strategies for weighting for non-response involve dividing the responding enterprises into a set of comprehensive and mutually exclusive groups, referred to as weighting classes. A weight is then applied to each class.

Weighting classes

In order to implement non-response adjustments, it is required to create weighting classes. It is desirable to divide the sample in "response homogeneity groups/classes". The response rates should be as homogeneous as possible within these classes and different between the classes. Data used to form these classes must be available to both non-respondents and respondents. Usually it is possible to get information on size, economic activity, legal status, location, and other variables in the business registers.

More advanced methods for creating weighting classes are methods like classification based on a categorical search algorithm or a logistic regression model using auxiliary variables to estimate the probability of response (cooperation in the present case).

Sample-Based Weighting Adjustment

In sample-based weighting adjustment the weight adjustment applied in each class is equal to the reciprocal of the ratio of selected sample size to respondents within the class (the inverse of the response rate within the class). The grossing-up factor should then be multiplied by the non-response adjustment factor.

A simple example:

Size Class	Population (I)	Sample size (II)	Respondents (III)	Respondent with characteristic (IV)	Non-response adjustment Factor ($V = II / III$)	Initial Grossing-up factor ($VI = I / II$)	Adjusted Grossing-up factor ($VII = V * VI$)
Small	35 141	878	764	595	1.15	40.0	46.0
Medium	5 362	882	821	795	1.07	6.1	6.5
Big	761	761	624	543	1.22	1.0	1.2
Total	41 264	2 521	2 209	1 933			

Alternative forms of sample-based weighting are that the weights are not inverse response rates but estimated coefficients of a regression model (where survey response is the left-side variable). In this case, the weights are reciprocals of the response rates estimated by the regression model.

Population-Based Weighting Adjustment

Population-based weighting adjustment requires population estimates and class membership of respondents. If there is no data available about the non-respondents, population-based adjustment still is possible since this uses external control counts for the population and not data from the sample. The method is used to correct simultaneously for both non-coverage and non-respondents. The method is used similarly to the sample-based method.

In population-based adjustment (post-stratification adjustment) the classes are created based on variables, which are known both for respondents and for the population. Weights are then applied in proportion to the ratio of population to achieved sample, so that the sums of the adjusted weights are equal to population totals for certain classes of the population.

A two-step procedure of first adjusting for non-response (sample-based adjusting) and then adjusting to known population counts is a common method that is used. However, this

procedure is the same as a population-based weighting adjustment if the weighting classes in the sample-based and the population-based weighting adjustment are the same.

If the strata used in the stratification are used as classes in the weighting adjustment, there is no need for the weighting adjustment. The adjusted weighting procedure is then equal to the final grossing up/weighting procedure.

2.6 Data analysis

2.6.1 Post-processing

The data produced by the crawler will not be in the right format for computation of the indicators of interest and will require post-processing. The required post-processing is specific to the crawler that will be adopted by each NSI.

The specific software that will be presented in section 3.1.2 returns lists of web sites which contain the keywords of interest. One list per indicator is returned and it contains all URLs of the web sites of the sample where keywords were detected. For example it will not contain only `www.[company].com` but also `www.[company].com/contact`, `www.[company].com/services`, etc., wherever the keywords were found. The computation of the indicator requires only one “appearance” of each enterprise where the functionality of interest is available, i.e. only one appearance of each enterprise no matter how many keywords were found on how many pages. Post-processing, described in section 3.1.2 returns in the end this unique list.

2.6.2 Computation of indicators

Section 1.5 defines the indicators that can be produced from keyword occurrence data. The indicators are computed as estimated numbers of enterprises, which are then suitably subdivided into population sub-groups, e.g. by size class, and divided by the corresponding estimated or known population totals. For example, the number of enterprises with between 50 and 249 employees whose web site provides a shopping cart is estimated and then divided by the total number of enterprises of this size.

The estimates are weighted sums of the data items. The weights can be based on selection probabilities (being their reciprocal) or can also incorporate adjustments for unit non-response or result from calibration of the data versus known population totals. The issue has been discussed in section 2.3.3.

2.6.3 Estimation of the accuracy of the indicators

The accuracy of an indicator is estimated by its standard error (the square root of the variance). The estimation of the sampling variance should take into account the sampling design (e.g. the stratification).

The indicators should be treated by the NSI just like other indicators from business surveys. Each enterprise has provided data on a number of binary variables, has a corresponding weight assigned to it and the corresponding population proportions have been estimated. The estimation of their variance, even in the face of non-response is straightforward.

2.7 Confidentiality and privacy issues

The crawling of web sites is bound to raise concerns from the sample members. Although the content of web sites is public, its automated bulk extraction is different to the eyes of the site owners. For example, scraping of complete lists of prices may not be allowed although any visitor with enough patience could inspect all prices. Although the actual prices are not required for the indicators presented in this guide the site owners may find this hard to believe.

Clearly, the implementation of the survey and the production of statistics should be well inside the limits of the law. The NSIs cannot risk damaging their credential or jeopardizing the enterprises' trust in them and their willingness to participate in business surveys in the future.

As a first step, the survey and the processes should be transparent to the sample members. The NSI should inform the potential sample members of: a) the compulsory or optional nature of the survey and its legal basis, if any, b) the name and position of the person or body in charge of the survey, c) the purpose of the survey, d) the categories of data collected and processed, e) the statistics that will be produced, f) the fact that the data will be kept confidential and used exclusively for statistical purposes, g) the guarantees to ensure the confidentiality and the protection of data, h) the categories of persons or bodies to whom the data may be communicated, i) the way in which consent can be refused or withdrawn and, in the case of compulsory surveys, the possible sanctions this would entail, g) where applicable, the conditions of the exercise of the rights of access and rectification. The site owners should also be informed about the possibility of obtaining further information on request.

The selected sample members should give their explicit consent (what is termed “opt-in”) to be included in the survey. The indication by which they signify their agreement must leave no room for ambiguity regarding their intent.

3 Annexes

3.1 Software tools

3.1.1 Web crawlers

A variety of web crawler software tools are available. The most popular are the following:

- Wget⁷: one of the oldest web crawlers. It is available for MS Windows and Unix/Linux and supports FTP and HTTPS besides the standard HTTP protocol. It is implemented in C.
- cURL⁸: a command line utility for receiving and sending file URL syntax. It utilizes the libCURL library for implementing numerous Internet protocols (HTTP, HTTPS, FTP, SFTP, IMAP, POP, etc). cURL is implemented in almost every operating system
- Heritrix⁹: currently the main crawler and indexer of the Internet archive¹⁰. It was developed jointly by the Internet Archive and the Nordic national libraries. It is implemented in JAVA.
- scrapy¹¹: a fast high-level screen scraping and web crawling framework, used to extract structured data from web sites. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.
- DataparkSearch¹²: a search engine designed to organize search within a website, group of websites, intranet or local system.
- Norconex¹³: a web crawler initially created for Enterprise Search integrators and developers. It was released as open source under GPL3 on June 2013.
- PHP-Crawler¹⁴: an open source crawling script based on PHP and MySQL. Created to implement as simple as possible local web site searches, it became popular for small web sites on shared hosting.
- Htrack¹⁵: a free, open source web crawler and offline browser available for MS Windows, Mac OSX and various Linux alternatives.

The aforementioned utilities can handle web sites with static content. Dynamic web page creation via Asynchronous Javascript and XML or AJAX has revolutionized the web, but it has

⁷ <http://www.gnu.org/software/wget>

⁸ curl.haxx.se

⁹ <http://crawler.archive.org/>

¹⁰ <http://www.archive.org>

¹¹ <http://scrapy.org/>

¹² <http://www.dataparksearch.org/>

¹³ <http://www.norconex.com/product/collector-http/>

¹⁴ <http://astellar.com/php-crawler/>

¹⁵ <http://www.httrack.com/>

also hidden its content. Although that might be of desired result for some web sites it was seriously affecting their visibility because they were not showing in web search results. For that reason sitemaps were developed, which provide important aid in web crawling and scraping activities.

Some web pages which do not reside on sitemaps require more advanced techniques in case everything in a web page is built via JavaScript with hash tags¹⁶. This situation appears to users as a fixed URL in their browser followed by a new hash tag for every different web page. In order to be able to crawl such dynamic content the AJAX web crawling technique should be adopted. This technique¹⁷ is based on the fact that when crawler finds an AJAX URL (that is, a URL containing a #! hash fragment) it will request the content of it from the remote site in a slightly modified form. The remote server will return the content in the form of an HTML snapshot, which is then processed by the crawler.

It is also possible to use custom crawlers which can cope with Javascript. Although such effort requires manual customization for every site which could easily consume projects resources we mention the following capabilities:

- The Selenium¹⁸ regression web project with the WWW::Selenium module.
- Ruby's Capybara¹⁹: an integration test library, which can also be used to write stand-alone web-crawlers. Given that it uses backends like Selenium or headless WebKit, it interprets javascript out-of-the-box.
- Spider²⁰: programmable spidering of web sites with node.js and jQuery.
- The Mechanize²¹ library: used for automating interaction with web sites. Mechanize automatically stores and sends cookies, follows redirects and can follow links and submit forms. Form fields can be populated and submitted. With WWW::Mechanize::Firefox it is possible to let Firefox handle the complex JavaScript issues and then extract the resulting HTML.

3.1.2 Google's Custom Search Engine

An alternative tool was used in a pilot survey carried out for Eurostat: Google's Custom Search Engine²² (GCSE). This approach relies on the fact that Google has already crawled sites and indexed their content. The effort and cost required for preparing one of the aforementioned crawlers and quite possibly to modify it to adapt it to individual sites (a major multiplier of effort) are avoided.

¹⁶ <http://en.wikipedia.org/wiki/Hashtag#Hashtags>

¹⁷ <http://coding.smashingmagazine.com/2011/09/27/searchable-dynamic-content-with-ajax-crawling/>

¹⁸ http://search.cpan.org/~lukec/Test-WWW-Selenium-1.23/util/create_www_selenium.pl

¹⁹ <https://github.com/jnicklas/capybara>

²⁰ <https://github.com/mikeal/spider>

²¹ <http://mechanize.rubyforge.org/Mechanize.html>

²² www.google.com/cse

The purpose of GCSE is to let web site administrators include a search engine in their site for use by their visitors. This can be adapted to the needs of an NSI. An ‘internal’ web site can be prepared, accessible only by authorised staff and a CSE can be incorporated that will be instructed to search for specific keywords on specific groups of sites (the sample).

The creation of the search engine is very straightforward. The use provides the list of URLs that the engine will be searching into. Therefore, the URLs of the selected sample of enterprises will be provided.

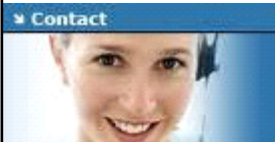
Subsequently the user can edit the search engine using the control panel provided by Google. The feature that is relevant for an NSI is the ability for image search: when searching whether a business site provides links to Facebook or Twitter, image search will detect the logos of these two networks, even if the keywords “Facebook” and “Twitter” do not appear in the site.

The collection of data proceeds **indicator by indicator**, i.e. target functionality by target functionality as follows:

1. A search query is typed in the search box; it is simply the set of keywords that correspond to the functionality. For example, to detect the functionality “Listing of open job positions or availability of online job application facility” (see Table 1) the query will be jobs vacancies if these are the two associated keywords.
2. The engine returns the search results as a list of URLs which contain at least one of the keywords (or their synonyms if this feature has been activated). The appearance of the list is the usual appearance of Google’s search results. A very small extract is given in Box 2. This list is in the form of an HTML file which must be stored for post-processing (see section 2.6.1). This list contains not only the main URL of the website (e.g. www.[company].com) but also all pages of the site that contain keywords (e.g. www.[company].com/contact, www.[company].com/services, etc.).


Box 2. Sample of GCSE search results.

[Shipping Companies - Patras Port Authority](#)
www.patrasport.gr/?section=1638&language=en_US



Central Agents: For Patras: PatraikaNautiliakaPraktoreiaS.A..Address: ΗρώωνΠολυτεχνείου 50. Post Code: 26441, Patra Phone: 2610-426000-10. Fax: 2610-...

[EBETAM A.E. - Contact](#)
www.ebetam.gr/?contact&lang=en



MIRTEC S.A. (Headquarters), A' Industrial Area, P.O.Box 13, GR-38500 Volos Tel : +30 24210 95340-2 Fax: +30 24210 95364, e-mail: volos.office@eb

3. Simple text scraping is applied to the list and only the URLs are retained. An identifier of the indicator is also appended in front of each URL, so that the data can be merged with the processed data for the other indicators. An example of the output of this scraping is given in Box 3.
4. A text-processing script, e.g. in Perl, is then ran on this output and maps its URL to the basic URL of the enterprise's web site. The basic URLs are those extracted initially from the sampling frame. This URL is appended to the end of each line of the file. All rows shown in Box 3, for example, would get enopsys.gr appended to their end.
5. Finally, these data are uploaded to a spreadsheet, database management or statistical software and can easily be processed to produce indicators. Next to each URL the location, economic activity, size and other characterizing information can be appended. In a spreadsheet, for example, a pivot table of URL by indicator would show the multiple appearances of each enterprise for a given indicator; converting all counts greater than zero into counts of 1 will provide the required 0 / 1 data for computation of the indicator.
6. A flat file can also be extracted for transmission to Eurostat. An example is given in section 3.3.

Box 3. Sample processed output of the GCSE results for indicator with ID N2b.

N2b : Link <http://enopsys.gr/en>

N2b : Link <http://enopsys.gr/en/photovoltaic-installations>

N2b : Link <http://enopsys.gr/en/company-profile>

N2b : Link <http://enopsys.gr/en/photovoltaic-installations/faq>

N2b : Link <http://enopsys.gr/el/weblinks/3-weblinks/6-green-energy>

N2b : Link <http://www.enopsys.gr/en/component/content/frontpage>

N2b : Link <http://enopsys.gr/en/news>

N2b : Link <http://enopsys.gr/en/photovoltaic-installations/projects>

N2b : Link <http://enopsys.gr/en/contacts/2-owners/2-gzarlaz-contact>

3.2 Example of mapping between target functionalities and keywords

As stated in section 2.4.1 the detection of target functionalities in business web sites relies on the detection of keywords or adopted technologies used as proxies of the functionalities. Such a mapping for a subset of the functionalities listed in Table 1 was used in a pilot survey. This mapping is shown in the following table.

Table 2. List of web site functionalities and matched keywords.

Functionality	Keywords
Contact information - URL	url, Website
Contact information - Email address	e-mail, Email, E-mail, email, eMail, E
Contact information - Telephone number	telephone, telephone number, Phone, Tel., Fax, Tel/Fax, T:, tel, TELEPHONE
Contact information - Postal address	address, Postal Address, Post code, P.O. box,
Availability of the web site in the national language	Language, Greek, EL

Functionality	Keywords
Availability of the web site in English	Language, English, EN
Availability of "last updated" date	Last Update, Last Updated Dated
Availability of privacy policy	privacy policy, terms of use, Privacy Statement, Conditions of use, Terms and Conditions, Terms & Conditions, Privacy, Legal, DISCLAIMER, Disclaimer, Copyright
Availability of registration facility	Signin, login, Login, register, Create an Account, openID, registration, Subscribe
Availability of site map	sitemap, site map, SITEMAP, Sitemap, Site Map
Listing of open job positions or availability of online job application facility	jobs, vacancies
Availability of links to multimedia content (audio, videos, etc)	mpeg,
Availability of links to social networks or blogs (Facebook, LinkedIn, Yammer, Twitter, etc)	widgets, Facebook, LinkedIn, Yammer, Twitter, Follow us, Share this page, Like us, T, F, BLOGS, Follow
Availability of links to wikis and wiki-based sharing tools	wikis

3.3 Transmission format

The NSIs will deliver to Eurostat the flat, comma-separated file, with the computed aggregates (estimated numbers of enterprises) that result from the post-processing and analysis of the collected data. Each record of the file will contain data about one indicator and one sub-group of the population of enterprises.

The fields of the file and the format of each one will be the following:

- Survey identifier: a string identifying the survey, to be agreed between Eurostat and the NSIs, common to all records.
- Reference period: a character string of format YYYYQA (where A=1, 2, 3, 4 denotes the quarter), which shows the reference period of the data. It is common to all records.

- Size: the size of the enterprise in terms of number of persons employed. Possible values:
 - B9: 0 – 9 persons
 - B10_49: 10 – 49 persons
 - B50_249: 50 – 249 persons
 - A250: 250 persons or more.
- Economic activity: the activity of the enterprise according to NACE rev. 2. The possible values are alphanumeric codes showing the NACE section (one-letter level) and division (2-digit level) of the enterprise's activity. Some codes correspond to ranges of divisions or sections. Possible values:
 - C10_18
 - C19_23
 - C24_25
 - C26_33
 - D_E
 - F
 - G
 - H
 - I
 - J
 - K6FINS: NACE 64.19, 64.92, 65.1, 65.2, 66.12, 66.19
 - L
 - M
 - N
 - ICT_T: an extra aggregate of NACE groups 26.1-26.4, 26.8, 46.5, 58.2, 61, 62, 63.1, 95.1
- Location: the NUTS level 2 code of the region where the enterprise is located.
- Indicator: a unique identification number for the indicator to which the record refers. Possible values:
 - CO1: Contact information - URL
 - CO2: Contact information - Email address
 - CO3: Contact information - Telephone number
 - CO4: Contact information - Postal address
 - LA1: Availability of the web site in the national language
 - LA2: Availability of the web site in English
 - IN1: Availability of "last updated" date
 - IN2: Availability of privacy policy
 - IN3: Availability of registration facility
 - IN4: Availability of personalised content for regular/repeated visitors
 - IN5: Availability of site map
 - IN6: Display of the number of visitors
 - PR1: Availability of product catalogues
 - PR2: Availability of price lists
 - PR3: Possibility for site visitors to customise or design the products
 - EC1: Availability of online ordering or reservation or booking facility
 - EC2: Availability of online order tracking facility
 - EM1: Listing of open job positions or availability of online job application facility

- EM2: Number of open job positions in the enterprise, listed in the web site
 - SN1: Availability of links to multimedia content (audio, videos, etc)
 - SN2: Availability of links to content in multimedia sharing sites (YouTube, Flickr, etc)
 - SN3: Availability of links to social networks or blogs (Facebook, LinkedIn, Yammer, Twitter, etc)
 - SN4: Availability of links to wikis and wiki-based sharing tools
- Value: the value of the indicator. Possible values are:
 - 0: denotes that the corresponding functionality is not available to the web site.
 - 1: denotes that the corresponding functionality is available to the web site.
 - 9: denotes that the corresponding functionality is not applicable to the web site.
- Count: the estimated number of enterprises, for this particular combination of size, economic activity and location, for which the indicator takes this particular value.
- Flag: any flags agreed between Eurostat and NSIs can be inserted here.
- Comment: Short notes can be appended here.

12.6. D4 – Feasibility analysis of selected data repositories for official statistics

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’

D4 - Feasibility analysis of selected data repositories for official statistics

April 2014

Document Service Data

Type of Document	Deliverable		
Reference:	D4 - Feasibility analysis of selected data repositories for official statistics		
Version:	6	Status:	Draft
Created by:	Marina Koumaki, Photis Stavropoulos, Alexandra Trampeli, Anais Santourian	Date:	10/4/2014
Distribution:	European Commission – Eurostat/G6, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	31/12/2013	Initial release
2	06/02/2014	Revised version (including two use cases of potential big data sources)
3	07/02/2014	Revised version (including three use cases of potential big data sources)
4	19/02/2014	Revised version (including five use cases of potential big data sources)
5	24/02/2014	Revised version: addition of introduction and general conclusions
6	10/4/2014	Revision based on comments received on 3/4/2014

Contact Information

Agilis S.A.

Statistics and Informatics

Acadimias 98 - 100 – Athens - 106 77 GR

Tel.: +30 2111003310-19

Fax: +30 2111003315

Email: contact@agilis-sa.gr

Web: www.agilis-sa.gr

TABLE OF CONTENTS

1	Introduction	4
2	Five use cases of potential big data sources	4
2.1	Vessel movement data from the Automatic Identification System (AIS)	4
2.1.1	AIS data: presentation of the source	4
2.1.2	AIS data: related official statistics	6
2.1.3	AIS data: feasibility of their use as input for official statistics	8
2.1.4	Main advantages	8
2.1.5	Access to data required as input for the derivation of Eurostat's variables	9
2.1.6	Computation of Eurostat's maritime transport statistics based on AIS data	12
2.1.7	AIS data: conditions for opening them to producers of official statistics	14
2.1.8	AIS data: conclusions	14
2.2	Real estate classified advertisements	15
2.2.1	Real estate classified advertisements: presentation of the source	15
2.2.2	Real estate classified advertisements: related official statistics	16
2.2.3	Real estate classified advertisements: feasibility of their use as input for official statistics	19
2.2.4	Main advantages	19
2.2.5	Issues that may preclude the use of Internet advertisement for the computation of housing indices	19
2.2.6	Real estate classified advertisements: conditions for opening them to producers of official statistics	21
2.2.7	Real estate classified advertisements: conclusions	21
2.3	Social media message data	21
2.3.1	Social media message data: presentation of the source	22
2.3.2	Social media message data: Related official statistics	26
2.3.3	Social media message data: feasibility of their use as input for official statistics	28
2.3.4	Main Advantages	29
2.3.5	Issues that may make difficult the use of social media data for the computation of subjective well-being indicators	29
2.3.6	Social media message data: Conclusions	30
2.4	Credit card transaction data (Visa Europe)	31
2.4.1	Credit card transaction data: presentation of the source	31
2.4.2	Visa Europe: EU Consumer Spending Barometer	32
2.4.3	Credit card transaction data: related official statistics	35
2.4.4	Credit card transaction data: feasibility of their use as input for official statistics	38
2.4.5	Credit card transaction data: conditions for opening them to producers of official statistics	38
2.4.6	Credit card transaction data: conclusions	39
2.5	Government financial transparency portal data	39
2.5.1	Financial transparency portal data: presentation of the source	39
2.5.2	Financial transparency portal data: related official statistics	40
2.5.3	Financial transparency portal data: feasibility of their use as input for official statistics	42
2.5.4	Financial transparency portal data: conditions for opening them to producers of official statistics	44
2.5.5	Financial transparency portal data: conclusions	44
3	General conclusions	44

1 Introduction

The aim of the project 'Internet as a data source' is to assess the feasibility of employing modern methodologies for producing high quality official statistics based on non-traditional data sources such as a) the monitoring of individual's activities online, b) the automatic collection of data from web sites, or c) the exploitation of Big Data.

The present report examines the potential of big data as a source of official statistics. Of particular interest are the so-called 'federated open data' which are (big) data from business or the public sector, generally not accessible by the public, but shared in an agreed and defined way with the producers of official statistics.

The present report examines five specific 'use cases', i.e. specific data repositories, most of them currently closed or partly open only, which could possibly be shared with producers of official statistics. Already open big data are also examined.

The report is organised as follows. Chapter 2 presents the potential of each of the five repositories. For each one the report presents the available data, the official statistics to which it can provide input, the way it could be employed in statistical production, its advantages and problems and the conditions under which National Statistical Institutes (NSIs) could have access to it. Chapter 3 presents the conclusions that emerge from the examination of the cases.

2 Five use cases of potential big data sources

2.1 Vessel movement data from the Automatic Identification System (AIS)

2.1.1 AIS data: presentation of the source

Among the numerous security regulations that came into effect after 2001 was the requirement for most commercial marine vessels to be fitted with Automatic Identification Systems (AIS). AIS is a primarily safety instrument required by the International Maritime Organization's (IMO) International Convention for the Safety of Life at Sea (SOLAS) that became fully operational in 2008. AIS provides a means for ships to electronically send data (about their position, destination, speed, etc.) with Vessel Traffic Services (VTS) stations as well as with other nearby ships.

AIS uses a positioning system, such as the Global Positioning System (GPS), in combination with other electronic navigation sensors and standardised Very High Frequency (VHF) transceiver to automatically exchange navigation information electronically. It is used by marine vessels in coordination with VTS for monitoring vessels' location and movements, managing vessel traffic and avoiding vessel collisions.

AIS messages are transmitted by ships using VHF signals. Vessel identifiers such as the vessel name and VHF call sign are programmed in during initial equipment installation. These are included in the signals transmitted by vessels along with location information originating from the ship's global navigation satellite system receiver. By transmitting a signal, vessels can be tracked by AIS base stations located along coastlines. When a vessel's position is out of the range of the terrestrial networks, signals are received via satellites that are fitted with special AIS receivers.

AIS is obligatory for vessels over 300 gross tonnage (GT) on international voyages, all passenger ships and vessels over 500 GT on domestic voyages. However, a very large number of vessels (over 70000) is fitted with AIS and the number is growing as smaller and cheaper devices are fitted even in small vessels (voluntarily).

There are two classes of AIS unit fitted to vessels, Class A and Class B. Class A units are a mandatory fit under the SOLAS convention to vessels above 300 gross tons or which carry more than 11 passengers in International waters. However, many other commercial vessels and some leisure crafts also fit Class A units. Class B units are currently not a mandatory fit. Class B units are designed for fitting in vessels which do not fall into the mandatory Class A fit category.

Navigation messages

AIS transceivers (of Class A) send data every 2-10 seconds depending on the vessel's speed-or every 3 minutes if at anchor. They include:

- The vessel's Maritime Mobile Service Identity (MMSI) – a unique nine digit identification number
- Navigation status – "at anchor", "under way using engine(s)", "not under command", etc.
- Rate of turn – right or left, (degrees per minute)
- Speed over ground – (knots)
- Position (longitude/latitude to 0.0001 minutes)
- Course over ground – (degrees, relative to true north to 0.1 minute)
- True heading – (degrees)
- True bearing at own position – (degrees)
- UTC Seconds – The seconds field of the UTC time when these data were generated. A complete timestamp is not present.

A different AIS message (also of Class A) that pertains to the vessel and the voyage is transmitted every 6 minutes:

- International Maritime Organisation's (IMO) ship identification number – a seven digit number that remains unchanged upon transfer of the ship's registration to another country
- Radio call sign – international radio call sign, up to seven characters, assigned to the vessel by its country of registry
- Vessel's Name
- Type of ship/cargo
- Length of vessel
- Location of positioning system's (e.g., GPS) antenna on board the vessel - in meters aft of bow and meters port of starboard
- Type of positioning system – such as GPS, DGPS or LORAN-C.
- Draught of ship – 0.1 meter to 25.5 meters
- Destination port
- Estimated time of arrival (ETA) at destination – UTC month/date hour: minute
- Optional: high precision time request, a vessel can request other vessels provide a high precision UTC time and date stamp

Class B transceivers, smaller and cheaper, have lower power and range (up to 15 km) and send shorter messages less frequently:

- Position message is sent every 30 sec to 3 min. and contains MMSI, time, speed over ground, course over ground, longitude, latitude, true heading
- Static (ship related) message is sent every 6 min including MMSI, boat name, ship type, radio call sign, length, and equipment vendor id.

2.1.2 AIS data: related official statistics

Maritime transport is the carriage of goods and passengers in sea-going vessels. European maritime transport statistics describe the movements in terms of type of cargo and passengers, the routes over which they are transported, the type, size and nationality of ships used to carry out that transportation.

European data collection on maritime transport provides a statistical description of the maritime component of the European transport activity in terms of its size and extent as well as its relation to other modes of transport.

From Eurostat's maritime transport statistics three domains reflecting vessel movements and carriage of goods across European ports appear to be relevant to data provided by AIS:

1. Vessel traffic (in number of vessels and in gross tonnage of vessels)
2. Maritime transport of goods (gross weight of goods)

Additionally, air emission statistics from maritime transport sector are of great relevance. Although, Eurostat does not yet compile official statistics on emissions from maritime transport, the emergence of detailed activity data from AIS provides an opportunity for the production of regular statistics on this domain.

Eurostat currently investigates the possibility of producing such statistics. Recently Eurostat has carried out a feasibility study in order to identify the methods used for emission estimation at national and international level in order to find out whether it would be feasible to compile European official statistics for this domain.

2.1.2.1 Vessel traffic

Eurostat's vessel traffic statistics (vessels calling at ports) provide data for two variables: (a) number of vessels in the ports in the European Union and (b) gross tonnage (GT) of vessels (which is a measure of the overall size of ship determined in accordance with the provisions of the International Convention on Tonnage Measurement of Ships (1969)).

These are disseminated broken down by type of vessel (e.g. container ship, liquid bulk tanker, etc. according to the International Classification of Ship Type (ICST)), size of vessel (in gross tonnage) and reporting country. They refer to the activity of ports of the reporting country during a quarter (quarterly data and are compiled on the basis of vessels arriving at the reporting port (inwards traffic). Annual results data are also compiled and disseminated.

The data are collected by the different data providers at port level. They cover ports handling more than one million tonnes of goods or recording more than 200 000 passenger movements annually (Main ports). However, data for some smaller ports may be included in the published results (since they are provided on a voluntary basis). Additionally, only movements of those vessels carrying goods and/or passengers for commercial activities (i.e. activities of loading or unloading cargo, embarking or disembarking passengers) are reported. Movements of vessels entering ports for other reasons, such as loading bunker fuel, sheltering from heavy weather or for repairing are excluded from the statistics.

2.1.2.2 Maritime transport of goods

Maritime transport of goods statistics covers data about the gross weight of goods handled (loaded and unloaded) in the port during a quarter (quarterly data). Annual data are also compiled and disseminated.

The "gross weight of goods" is defined as the tonnage of goods carried, including packaging but excluding the tare weight of containers or Ro-Ro units. In detail, the gross weight of each consignment is the weight of the actual goods together with the immediate packaging in which they are being transported from origin to destination, but excluding the tare weight of containers or Ro-Ro units (e.g. containers, swap bodies and pallets containing goods as well as road goods vehicles, wagons or barges carried on the vessel).

Data on gross weight of goods (in thousands of tonnes) are made available from Eurostat with different (combinations of) breakdowns, including the (a) reporting country, (b) direction (inwards vs. outwards), (c) type of traffic (national and international), (d) type of cargo, (e) loading status (loaded, empty, etc.), (f) type of vessel and (g) nationality of registration of vessels. Additionally, detailed data for each country are disseminated providing information about the gross weight of goods transported from the reporting country to "partner" ports from/to where goods are carried (i.e. the port of loading/unloading).

The data are collected by the different data providers at port level and cover the activity in Main ports. Additionally, only movements of those vessels carrying goods and/or passengers for commercial activities (i.e. activities of loading or unloading cargo, embarking or disembarking passengers) are reported. Movements of vessels entering ports for other reasons, such as loading bunker fuel, sheltering from heavy weather or for repairing are excluded from the statistics.

2.1.2.3 Maritime transport emissions

Emissions from maritime transport sector have been recognized as an increasingly significant factor of climate forcing and a growing concern for air quality. However, such statistics are not currently published by Eurostat.

The EU has been active in pursuing policies at the international level for reducing GHG emissions from shipping. The main policies in regulating emissions from maritime transport are still under development at regional and international level. Although a target has been set for 2050, the path towards that target has not been delineated. However, the main statistical requirements for policy monitoring are emissions estimates disaggregated based on ship activity data followed over time via a harmonised methodology.

More specifically, required statistics for EU emissions for policy monitoring should at least include data on emissions of Greenhouse Gases (GHGs) (in CO₂ tn equivalent) broken down by type of pollutant for ships calling to EU ports, type of ship, size of ship and flag state.

Despite the environmental orientation of transport policies, the current statistical system is not designed to assess the impact of transport to the environment or estimate GHG emissions from it. Transport and environment models require detailed transport activity data to calculate emissions, make projections and identify economic drivers affecting climate change.

Eurostat has already initiated activities for further monitoring environmental objectives. However, a number of tools for estimating GHG emissions have already been developed by Member States (MSs), international organisations and researchers.

There are two main approaches for the estimation of emissions from maritime transport: top-down and bottom-up.

A top-down approach calculates global emissions by quantifying fuel consumption, which then is transformed into emission estimates via emission factors. Fuel consumption is calculated using fuel sales from international bunkers (e.g. from the International Energy Agency (IEA)) and then estimates are computed using emission factors for each pollutant (CO₂, NO_x, SO_x, particulates - PM, etc). At this point top-down methods diverge:

- A full top-down approach will disaggregate global emissions at the desired regional level based on relevant statistics used as special proxies. These can include GDP, trade statistics, national emissions, national fleet size etc.
- A top-down approach with bottom-up geographical characterisation will use activity data in order to disaggregate global emissions.
- A bottom-up approach will start from detailed activity data and vessel characteristics and with some model assumptions on engine use and fuel consumption will compute emission estimates that are then aggregated at the desired level. Then a reconciliation of the total with emissions from a top-down approach is performed.

On the other hand, a full bottom-up approach estimates the emissions of a vessel at a specific instance and then aggregates the estimates to produce the desired statistics (e.g. over time and vessel fleet to provide total emissions).

2.1.3 AIS data: feasibility of their use as input for official statistics

The objective of this feasibility study is to investigate whether it would be computationally feasible for Eurostat to use as input data from the tracking of vessels based on AIS records for supplementing or replacing official statistics on maritime transport. Additionally, it aims to provide an evaluation about how methodological and practical restrictions can affect the overall quality of the statistics that can be produced.

For future and present needs of European statistics it appears that AIS data is a suitable and relevant source for complementing or replacing official maritime statistics. From the brief description of the source and maritime statistics produced by Eurostat it can be drawn the conclusion that the most relevant variables that can be compiled based on the data obtained from AIS are:

- Number of vessels
- Gross tonnage of vessels
- Gross weight of goods handled at European ports
- Air emissions from the maritime transport sector activity

2.1.4 Main advantages

AIS-based data contain detailed information about the position of the vessel and its route from the port of departure (or last known AIS position) to the port of destination, along with the information that pertains to the vessel. They completely cover ship activity of EU vessels, vessels sailing in or around EU waters and vessels sailing towards and/or from EU ports.

These data are transmitted continuously and in huge amounts, providing a comprehensive and detailed data set for individual vessels, which can be aggregated to a population's average characteristics providing accurate statistics in the desired time and location resolution.

Additionally, AIS has huge coverage in terms of ships transmitting AIS signals since a very large number of vessels is fitted with AIS. Therefore, huge amounts of data can be obtained almost in real-time.

Generally, the usage of AIS data may contribute to:

- Improve timeliness of the statistics
- Reduce burden to current data providers
- Improve accuracy of the statistics due to the dependence on actual raw data, which do not require manual processing, such as manual filling of forms by ships and submission of forms from port authorities.

2.1.5 Access to data required as input for the derivation of Eurostat's variables

Data need to be primarily provided via a web-based service. There are a number of commercial maritime databases through which data on vessel routes can be obtained. An appropriate case is MarineTraffic since it provides data of good coverage.

MarineTraffic¹ is a service that provides real-time information about ship movements and ports, mainly across the coastlines of many countries around the world.

Vessel positions are recorded based on AIS. The MarineTraffic terrestrial-based AIS network provides coverage of vessel positions in real-time at several thousands of ports and coastal shipping routes worldwide. Additionally, in order to cater for increasing demand for global AIS coverage, Marine Traffic combines terrestrial ship tracking with Satellite AIS data. Satellite AIS data come as an ideal supplement, allowing to monitor vessels tracks well beyond coastal regions, including the oceans, while offering limited coverage at crowded areas near the coastline. The combination of Satellite and Terrestrial AIS gives a unique presentation of the global maritime traffic and provides a daily update of almost the entire global merchant fleet.

MarineTraffic thus handles millions of vessel position records daily. Data received are uploaded in the database in real time and are immediately available on a Google map and on other pages.

MarineTraffic provides five APIs² through which different data can be obtained. Users can use this service to receive vessels' position data, along with port calls, ship particulars and photographs.

From the list AIS data that can be obtained from MarineTraffic, the following elements are required for the computation of Eurostat's variables. These include:

- Dynamic information: vessel position (longitude, latitude), navigation status, UTC seconds, wind, speed
- Static information: vessel ID (MMSI, IMO number), vessel type, gross tonnage of vessel, year of built, width, length
- Voyage-specific information: port of destination, draught of the vessel, deadweight of the vessel

From the available MarineTraffic's APIs, the most relevant one through which data about the abovementioned variables can be obtained is the API on vessel positions.

The **API on vessel positions** provides data on the latest position of several vessels at once, at regular intervals. It works for a predefined fleet or area but it can be configured to provide data for all ships that arrive at European ports. For this configuration it is necessary to provide a list with these ports.

This API may provide data at different frequency options and level of detail.

The so-called "simple response" provides the following data at most once every two minutes:

¹ <https://www.marinetraffic.com>

² <https://www.marinetraffic.com/en/p/api-services>

- MMSI number
- Latitude
- Longitude
- Speed (in knots)
- Course
- Status
- Timestamp

The so-called “extended response” provides data at most once every hour. It includes the following additional information compared to simple response:

- Ship name
- Ship type
- IMO number
- Call sign
- Flag
- Current port
- Last port
- Last port time
- Destination
- Estimated time of arrival at destination (ETA)
- Length
- Width
- Draught
- Gross Tonnage (GRT)
- Deadweight (DWT)
- Year of built

The data can be received from MarineTraffic either in XML, CSV or JSON format. A sample of a CSV datafile is indicatively presented below. As it can be noticed, it includes a string of each event record.

MMSI	LAT	LON	SPEED	COURSE	TIMESTAMP	SHIPNAME	SHIPTYPE	IMO	CALLSIGN	FLAG	CURRENT_PORT	LAST_PORT	LAST_PORT_TIME	DESTINATION	ETA	LENGTH	WIDTH	DRAUGHT	GRT	DWT	YEAR_BUILT
237594800	37.44848	25.32671	0	177	2012-04-18T21:10:00	ORCA	65	0	SY2714	GR	MYKONOS	MYKONOS	2012-04-18T17:12:00	DELOS MYKONOS	1900-01-01T00:00:00	43	10	25			
240521000	37.46272	25.32613	0	71	2012-04-18T21:09:00	THEOLOGOS P.	60	9223150	SZNB	GR	MYKONOS	RAFINA	2012-04-18T15:09:00	AND-THN-MYK	2012-04-18T22:30:00	118	22	48	4935	3227	2000
237106400	37.46368	25.32642	0	0	2012-04-18T21:10:00	AGIA ELENI	31	0	SV4137	GR	MYKONOS	MYKONOS	2012-04-18T17:12:00	MYKONOS	2012-04-30T11:00:00	30	7	0			

Although the data provided by MarineTraffic contain detailed information about the ship routes with high accuracy and coverage, they may not include the whole set of variables required for the computation of the existing indicators. Data that pertain to vessels' characteristics may not always be available for each single vessel. Additionally, essential information for the estimation of emissions from maritime sector such as engine power, engine type is not part of the data provided by MarineTraffic.

However, there is a large number of international databases on ship characteristics that contain such information:

When a vessel is commissioned it receives an IMO number. At the same time its main characteristics are entered in the **IHS Fairplay** (previously Lloyd's Register of Ships) database that handles IMO numbering. IHS offers several commercial products at various levels of coverage. The most detailed is the Seaweb. It claims to contain detailed information on 180,000 vessels of 100GT and above and it is constantly updated with new buildings and casualties. The database includes up to 600 data fields, including tonnages, class, inspections, cargo, capacities, gear and machinery details. Significantly is also keeps a record of historic vessel movements for 5 years.

LMIU, short for Lloyds' Marine Intelligence unit, has a long history of providing maritime information. Currently it claims to offer detailed characteristics for over 120000 vessels including tonnages, class, inspections, cargo, capacities, gear and machinery details. Besides other information (owners, shipbuilders, inspections etc) it keeps historical ship movement data that go back as far as 1997.

Other databases that are not as extended in coverage may also be used to cover missing variables. Shipbrokers maintain large databases with ship characteristics that can be used for this purpose. For example one of the largest, Clarkson's, offers detailed information on 40,000 ships over 100GT.

EQUASIS: The Equasis information service was established in May 2000, following the signature of a Memorandum of Understanding by the European Commission, France, Japan, Singapore, Spain, the UK and the US Coast Guard. Since 2007, the Commission has been represented by EMSA in both the MoU and the governing bodies of Equasis. In June 2008, the Equasis Supervisory Committee mandated EMSA to take responsibility for the hosting of the management unit. A statistics team in Equasis produces the annual Equasis statistical publication "The world merchant fleet" (with contribution from EMSA) and supports the agency's information needs by coordinating and managing the procurement of maritime data from the commercial data providers. These sources, which include information on vessel characteristics, vessel movements, historical information about ships, casualties, inspections, deficiencies, detentions, owners, demolitions, new buildings and equipment on board vessels, are made available to agency staff.

EMSA has also developed a vessel characteristics database and is currently populating with data from commercial providers, information from member state's registers (information from contracting procedure EMSA/OP/09/2012³).

Again, EMSA plays a pivotal role in providing data about vessels through the Equasis information service that is based on data from commercial providers but available to the public for free. Some technical information about engine characteristics may not be available through Equasis or the EMSA vessel characteristics database currently under development. In this case it might be required that some data may have to be purchased from the commercial providers.

³ <http://emsa.europa.eu/work/procurement/calls/111-on-going-calls-for-tenders/1551-op-09-2012.html>

2.1.6 Computation of Eurostat's maritime transport statistics based on AIS data

The approach for computing Eurostat's statistics based on the AIS data that can be obtained from MarineTraffic is presented below.

- a. **Number of vessels:** This variable can be almost directly computed from the data provided by MarineTraffic's API on vessel positions. Since data quantify individual vessel activities and provide real-time information about the location and status of vessels, a typical process for the computation of the variable consists of a simple aggregation of the detailed position data at the desired time and location resolution. Thus, the number of vessel arrivals at a port can be derived by aggregating the number of those vessels that were at port during a period of reference. Information about the vessels that arrived at port can be derived from the reported navigation status, last port and last port time. It should be finally noticed that data on vessel arrivals are only required for the computation of the variable since relevant Eurostat's data refer to inwards traffic.
- b. **Gross tonnage of vessels:** The gross tonnage of vessels that arrived at port is another variable that can be calculated on the basis of the available data from MarineTraffic. Information about the gross tonnage of vessels that arrived at a port is provided by MarineTraffic's API on vessel positions. This is the key variable required for the computation of Eurostat's relevant variable on gross tonnage of vessels. The procedure that should be followed for computation of this variable is similar to the previous one.

For the derivation of the variables broken down by size class and size of vessel categories, the following information is required:

- Type of vessel. MarineTraffic's data provide information on the type of vessel. Actual data cover a detailed classification of vessel. This requires a list matching the classification obtained from MarineTraffic's database to Eurostat's classification.
- Size class of the vessel (in gross tonnage). The size of the vessel is determined by its gross tonnage. This information is available for a large number of vessels. In this case, the size class categories can be computed according to Eurostat's classification.

One main issue, however, is that information about a vessel's gross tonnage is sometimes missing.

Missing data for the gross tonnage may be statistically estimated. Domain experts could possibly develop a model that would receive as input the vessel's characteristics, namely the vessel's type, length, width, draught, deadweight and year of built in order to predict its gross tonnage.

Alternatively, missing information can be obtained from international databases on vessel characteristics. Data provided by MarineTraffic should be then matched to vessel characteristics data obtained from international databases on vessel characteristics based on their IMO number. Each vessel commissioned receives a unique IMO number that stays the same if the owner or the ship's name change.

- c. **Emissions from maritime transport activity:** as already mentioned there are two different approaches for the estimation of emissions from maritime transport. Bottom-up models are based on data on vessels and their activity. A typical bottom-up process combines vessel characteristics (especially installed power of main and auxiliary engines) with activity data to estimate energy produced which in turn is used to compute fuel consumption and then emissions. Several bottom-up models, such as the STEEM model, the ENTEC model, the EMS/MARIN model, have been developed for the estimation of emissions from maritime transport.

In order to produce the required statistics using a bottom-up approach data are needed for estimating emissions per trip and per vessel that can then be appropriately aggregated to produce the required estimates. The most desirable but not yet available data at this level is actual fuel consumption for each trip. However, at present, these data need to be approximated based on available data:

- **Route delineation.** The first set of information required is related to vessel activity that consists of trips between ports. Data obtained from MarineTraffic is the primary source of route data.
 - **Vessel speed.** Vessel speed is important in emission models and slow steaming is a main operational abatement method. Instantaneous speed is included in AIS messages and is made available by MarineTraffic. Average speed can be computed from subsequent position recordings and relevant timestamps. Average speed over the whole trip can also be approximated from the information included about time of departure from a port and time of arrival at a port.
 - **Capacity utilization.** Vessels travelling on ballast emit smaller amounts of GHGs. Whether a vessel is loaded can be determined by the type of ship and also the ship's draught (so the volume of cargo can be inferred if compared with maximum draught from vessel characteristics).
 - **Vessel characteristics:**
 - **Identification.** Ship identification codes are included in vessel characteristics databases and also AIS messages. They are needed to combine activity and vessel characteristics data. These include IMO number and MMSI (Maritime Mobile Service Identity), which are also made available from MarineTraffic's data.
 - **General Vessel Characteristics.** Gross tonnage (GT), Deadweight tonnage (DWT), Length (L), Breadth (B), Draught (d), Hull type, Build year, Design Speed. These data are either available from MarineTraffic's data or generally available in databases of vessel characteristics.
 - **Other data.** Winds affect vessel emissions as they affect the power needed to attain the speed over ground. Wind currents are modelled based on meteorological conditions. There are some models with global coverage and real time or near real time results like ESA's Globwave⁴, which provides values for wind (velocity, direction) characteristics. Although this information is not included in MarineTraffic's API, it can be provided – upon request – since it is already available in its database.
- d. Gross weight of goods:** Eurostat's variable on gross weight of goods cannot be directly derived from the available data. However, draught can be used to determine the weight of the cargo on board by calculating the total displacement of water. Additionally, tables made by the shipyards provide information about the water displacement for each draught. The density of the water (salt or fresh) and the content of the ship's bunkers have to be also taken into account. In the literature, there are models and methods that have been developed allowing the estimation of the gross weight of the cargo from draught. These algorithms can be incorporated in the algorithm computing the number of vessels arriving or leaving a port (inwards and outwards traffic) in order to estimate the required variable.

Since the cargo weight can be algorithmically estimated based on draught, the data processing algorithm can calculate the difference in cargo weight on ships' arrivals and departures, which in

⁴ <http://www.globwave.org/>

turn provides an indication of the net weight of the cargo (i.e. the weight of goods in a consignment, excluding any immediate packaging) handled at the ports. However, this net cargo can often result from loading or unloading of a part of cargo, which may lead to discrepancies.

Taking these issues into consideration, it can be deduced that the produced statistics may not be of high accuracy. In order to assess their accuracy, the estimates produced should be validated by comparing them with Eurostat's actual data.

2.1.6.1 Coverage

MarineTraffic uses a combination of Satellite and Terrestrial AIS information. The Terrestrial AIS service provides near real-time updates of vessel positions at areas covered by MarineTraffic's coastal receivers network. While Terrestrial AIS provides real-time data, Satellite AIS service covers position updates less often but over the entire world.

On average, several Satellite AIS updates per day should be expected for most vessels sailing at the oceans, equipped with a Class-A or B AIS transponders. Although, real-time position updates are crucial for following vessels near coasts and ports, a couple of position updates per day for following vessels at the open sea are usually enough.

As a result, data provided by MarineTraffic are of good coverage. The only issue is that vessels of less than 300 GT may not be well represented. However, the contribution of vessels of this size class in commercial traffic, which is of interest, is not significant.

2.1.7 AIS data: conditions for opening them to producers of official statistics

MarineTraffic's API is available at a cost, which is negotiable. There are data available that are not included in the API but they may be added, if necessary. There are no specific constraints in the conditions for opening them to producers, e.g. confidentiality constraints or non-disclosure.

Additionally, data from third parties (i.e. databases on vessel characteristics) can be provided at a cost.

2.1.8 AIS data: conclusions

There is a high potential in using AIS data in the production of current statistics:

- Number of vessels, by size and type of vessel
- Gross tonnage of vessels, by size and type of vessel
- Emissions from maritime transport activity sector (currently not compiled by Eurostat but their compilation is under investigation)
- Gross weight of goods

A potential data source for obtaining AIS data is MarineTraffic. Although some data about vessels' characteristics may be missing or may not be readily available, these can either be estimated or obtained from an international database on vessel characteristics.

It is, however, possible to derive statistics on the number of ships almost in a straightforward and simple way from data that can be made available from MarineTraffic. This is possibly the only indicator that could replace official statistics in the very near future.

2.2 Real estate classified advertisements

2.2.1 Real estate classified advertisements: presentation of the source

The data source used, namely XE⁵, is one of the biggest classifieds site/newspaper concerning house sales and rental prices in Greece. For the needs of the current research, all data for purchasing and renting residential properties will be acquired.

The data source contains information about the area, price of the house property, location etc. in a structured form whereas other information such as the number of rooms in the house, the view, etc. is provided in an unstructured format. These data are usually provided in a free text form, which actually includes the content of the advertisement in the form that it is being published.

The type of big data from this source refers to house/flat sales and rentals that are put in the market through Internet advertising. Data is entered by individuals or businesses (real estate agents) and contain information about a single house property (e.g. list-price, area, location, etc).

The available data refer to individual property and cover all house sales and rentals in Greece. They cover those house sales and rentals published through XE's site/newspaper. Data are updated on daily basis.

When an owner or agent desires to upload an advertisement, fills-in a descriptive questionnaire about the property's characteristics. The information that needs to be provided is the following:

Characteristic	Measurement	Obligatory ⁶
Price	In Euros	No (but is encouraged/promoted)
Floor Area	In square meters	Yes
Location	Region/municipality, locality (two levels)	Yes
Property Category	Apartment, detached house, maisonette, etc	Yes
State of the property	New house, under construction, unfinished, etc	Yes
Level	Basement, ground floor, first floor, etc	Yes
Construction/Renovation year	-	Yes
Number of bedrooms	-	Yes
Number of bathrooms	-	No
Property type	Residential, resort, etc	No
House type	Neoclassical, preserved, loft, traditional, studio etc	No
Action	Sale, rent, exchange, etc	Yes
Orientation	Corner, bright, etc	No
View	Sea view, mountain view, forest view, etc	No
Heating	Central, autonomous	No
Other	Pool, parking, storeroom, solar heater, gas, etc	No
Property availability	Immediately, date when the property will be available	No

⁵ <http://www.xe.gr/property/>

⁶ All the variables marked as *obligatory* must be filled-in during the post of the assignment (through the website).

Characteristic	Measurement	Obligatory ⁶
Photos/ video	-	No
Contact details	-	Yes

Similarly to most real estate ad sites, the XE site requires some fields that are the most important for a buyer or renter for evaluating a house (total area, location, number of bedrooms etc). These characteristics are accompanied by many more others that are provided as selections, i.e. as classifications or clickable fields for present or absent characteristics.

In addition to structured fields that take values from a classification there is also a free text description that can be text mined for further usable information.

2.2.2 Real estate classified advertisements: related official statistics

Housing is very important for households and usually constitutes the most important expense in their budget. The housing market also plays a key role in the economy as it affects consumer behaviour and (either directly or indirectly) macroeconomic policies⁷. In the last decade protracted housing boosts and bursts in the developed world, helped trigger the financial crisis of 2007 and the ensuing great recession. Therefore timely housing price statistics of high quality are of primary importance for academics and policy makers.

Three domains are accessible in Eurostat that provide official statistics reflecting price levels and trends for buying or renting a housing property across European countries:

1. Harmonised indices of consumer prices (HICP)
2. Housing price index (also named Residential Property Prices Indices - RPPIs)
3. Purchasing power parities

2.2.2.1 Harmonised indices of consumer prices (HICP)

The first indicator of the price domain in Eurostat's website is the Harmonised Index of Consumer Prices (HICPs). The main HICPs include the Monetary Union Index of Consumer Prices (MUICP), the European Index of Consumer Prices (EICP) and the national HCIPs. The responsibility to collect these data on a monthly and annual basis lies on National Statistical Institutes.

These are economic indicators (deflators) that measure the change of the prices of consumer goods and services acquired by households over time. In other words, they are a set of consumer price indices (CPIs) calculated according to a harmonised approach and a single set of definitions. The HICPs cover all expenditures within the territory, whether by residents or visitors.

The data for the prices come from surveys, visits to local retailers and service providers and from central collections via mail, Internet or telephone. An important consumption category according to COICOP-HICP (classification of individual consumption by purpose) that relates to the concept of real estate statistics is the '**Actual rentals for housing**'.

Rental costs are usually determined via special household surveys that record rent expenses and quality characteristics of residential property.

⁷ André, C., Gupta, R., & Kanda, P. (2012). Do House Prices Impact Consumption and Interest Rate?: Evidence from OECD Countries Using an Agnostic Identification Procedure (No. 947). OECD Publishing.
Bulligan, G. (2010). Housing and the macroeconomy: The Italian case. In *Housing Markets in Europe* (pp. 19-38). Springer Berlin Heidelberg.

Besides actual rents, owner occupied housing has been recently included in the HICP. It includes both acquisition and ownership (repairs, maintenance, insurance etc) expenses. A methodological manual has been provided to MSs that delineates best practises and ensures comparability and coherence of computed indices. The owner occupied index (based on net acquisitions) covers dwellings that are acquired by households for own use and that are new to the household sector. Therefore the index includes new dwellings constructed by self-builders and excludes dwellings bought from the non-household sector (e.g. for rent or re-sale).

2.2.2.2 Housing price index (HPI)

The Housing Price Index (HPI) shows the price changes of residential properties purchased by households (flats, detached houses, terraced houses, etc.), both newly-built and existing ones, independently of their final use and of their previous owners. Therefore self-build dwellings are excluded. Only market prices of residential properties are considered but the price of land is included in prices and weights.

The HPI should be seen as an independent indicator aimed at measuring the evolution of residential market transactions, independently of the institutional sector that were bought from and the purpose of the purchase. Thus, both new dwellings purchased and existing dwellings are taken into consideration in the compilation of the indicator. Moreover, it should be seen as a price indicator that attempts to:

- Measure house inflation across countries
- Assess housing affordability over time
- Measure specific price trends
- Monitor economic imbalances and financial stability
- Be used as input for national accounts purposes
- Be used as input to economic forecasting and analysis
- Be used as input for decision making in respect of the house market

HPI is computed as Laspeyres type annual chained index allowing weights to be changed each year. Its compilation is based on the final market prices that are paid by households (i.e. VAT and other taxes are included).

More specifically, European HPIs are calculated as weighted average of the national HPIs, using as weights the GDP at market prices (based on purchasing power standard) of the countries concerned. They are presented not only quarterly but also annually in Eurostat's database.

Data for the prices of the dwellings may come from various sources including real estate agents, construction companies, financial institutions, administrative sources and relevant surveys. In addition, national accounts, construction statistics and household budget surveys are the main data sources for the computation of the weights, which are taking account the total values of the houses' purchases.

As it is mentioned above, a survey can be conducted in order to collect real estate data. The surveys have the aim of asking directly the units to state information on transactions that were carried out in the relevant period. Additionally, surveys have the intention of following-up the price's evolution of "representative dwellings" throughout time.

The questionnaire of this relevant survey consists of a series of questions for the purpose of gathering information for the general characteristics of a dwelling. It is noted that those characteristics, which are listing below, depend on market characteristics of each country.

- Location (Municipality, Town, NUTS area, postal code)
- Type of dwelling (Detached house, flat etc.)
- Price of the dwelling
- Other expenditures (notary, registry fee, transfer taxes and other taxes)
- Total floor area (in square meters)
- Total usable floor area (in square meters)
- Facilities of the house (number of floors/rooms/bathrooms, garage, pool, basement, storage attic etc.)
- Existence of central heating
- Quality of neighbourhood (surroundings, shops, health services, accessibility, transport, schools)

2.2.2.3 Purchasing power parities

Purchasing power parities (PPPs) are indicators of price level differences across countries. In other words, it is a comparable measure for indicating how many currency units a given quantity of goods and services, costs in different countries. Therefore, they eliminate the effects of the differences in price levels between Member States thus allowing volume comparisons of GDP components and comparisons of price levels. In their simplest form PPPs are price relatives that show the ratio of the prices in national currencies of the same good or service in different countries.

Price Surveys are organized every year in order to compile prices for PPPs of actual and imputed rents. For those countries that have not a representative rental market, dwelling stocks estimates are used to estimate prices. Data for weights (country's expenditures) are compiled from national accounts, which are used then to aggregate the PPPs. In addition, most National Statistical Institutes (NSIs) use price collectors to obtain price data, and most other input data required are extracted from existing sources at the NSIs.

Actual and imputed rents

'Actual rentals for housing' and 'Imputed rentals for housing' are actually expenditure groups, which belong to consumer goods and services. However, they are covered by a separate survey.

Countries collect data on the rents paid by tenants and also on the imputed rents of owners and occupiers. The data refer to a number of precisely defined dwellings classified by type of dwelling (flat or house), number of rooms and availability of central heating. The data cover average area, average monthly rent per square meter and the relative importance of each dwelling class in the total expenditure of the relevant basic heading.

Countries that do not have a large and representative rent market and so are unable to supply the required data on actual and imputed rents, report data on the quantity and quality of their housing stock. The data comprises, separately for flats and houses, the number and total usable area of dwellings by number of rooms (Quantity data), and the number and share among the total of dwellings with availability of certain facilities such as electricity, running water etc. (Quality data). With data on housing stock volume measures are computed directly (Quantity approach).

Data are extracted from existing relevant statistical sources. The survey takes place annually under the responsibility of Eurostat and countries report data for the last three reference years t , $t-1$ and $t-2$.

PPPs for housing are obtained either directly with the price approach from the two basic headings on actual and imputed rents, or indirectly with the quantity approach from quantity and quality data collected on housing stock. Price approach, direct PPPs, are combined with the quantity approach, indirect PPPs, using as links the data from countries that supplied data for both approaches in order to produce the final set of PPPs on actual and imputed rents that cover all participating countries.

2.2.3 Real estate classified advertisements: feasibility of their use as input for official statistics

From the brief description of the source and the produced statistics the data obtained from Internet advertisement of real estate is quite relevant to

- Rents in the HICP
- Owner Occupied Housing in the HICP
- House price index
- PPPs for the housing heading (both rents and owner occupied housing) using the direct approach.

2.2.4 Main advantages

The main advantages of data from internet advertising over standard methodology are that:

1. They are continuously updated providing a constant stream of timely and fresh data
2. They can provide huge amounts of data at negligible marginal cost thus minimising sampling error
3. Internet based data across countries can be obtained and analysed in a unified fashion that can enhance geographical comparability.
4. It can restrict the often cumbersome and expensive price data collection to an automated process minimising cost and burden.

2.2.5 Issues that may preclude the use of Internet advertisement for the computation of housing indices

On the downside there are important issues that restrict the ability of these data to be used to compute existing indicators and need to be addressed.

List price vs. Sale price and transaction cost

The main problem with Internet advertisements data is that they contain the list price stated by the current owner. This is generally greater than or equal to the sale price. All indices mentioned in Section 2.2.2 are based on actual prices at which a transaction is made or is recorded with the administration, and in fact these are the relevant prices for policymaking.

The relationship between list price and actual price is notably missing from the literature, with few exceptions. The most recent study⁸ compares Internet advertisement data with official data from the Central Bank of Ireland for the boom-bust period of 2001-2012. The author estimates the correlation between hedonic price indices from the two data sources at 98% and concludes that *“using list prices*

⁸ Lyons, R. C. (2013). Price Signals and Bid-ask Spreads in an Illiquid Market: The Case of Residential Property in Ireland, 2006–2011. Available at SSRN 2205742.

when first posted is a very accurate gauge of changes in house prices, even in extreme market conditions”.

There are several ways to adjust data for the mismatch between asking and actual price.

The simplest and most costly one is to estimate a mean ratio of list to actual price. Then the actual price of an offering can be computed from the list price⁹. This requires a subsample of matching list and sale prices. The information can either be obtained from the site if the whole transaction is made through it or by a telephone survey of sellers.

In reality, houses are sold after some negotiation that leads to an agreed price. The outcome of the negotiation depends on some aspects that have been studied and can be used for the required adjustment. These include:

- **Seller’s patience.** Carrying costs (taxes, utilities, maintenance) of the house with no offsetting benefits such as rent income or occupancy affects the bargaining power of the seller and influences the sale price¹⁰. It is common in house descriptions to establish whether a property is vacant or not and thus the owner’s “patience”
- **Hot and cold markets.** In a hot market also referred to as a seller’s market properties for sale or rent are few and stay in the market for a limited time before being sold to the more numerous buyers. This gives sellers a better bargaining position and the final sale price is closer to the list price. In contrast in a buyer’s market, or cold market, a large number of properties is up for sale or rent and buyers are few. In this case properties stay in the market a long time before sold often in deep discounts over the list price. Whether listings correspond to a cold or hot market and thus the size of the difference between asking and sale price, can be indicated using variables that are found in the advertisements of sold / rented properties, provided that the site either removes these advertisements or marks them as “sold” / “rented”.
- The **selection spread** i.e. the difference between the listing price of the property under negotiation and the average price of the stock of advertised properties.
- The **time-to-sale.** A shift in the age of the advertisements indicates shifts in market conditions. Lyons (2013) reports that average time-to-sale in Ireland moved from two months in 2006 to six in 2009 and nine months in 2012.

Even if advertisements are not diligently removed from the site, the total number of offerings compared to a long-term average can provide an indication of the spread between list and sale prices.

2.2.5.1 Population Coverage

Dwellings offered that have not yet been built. This may create coverage errors because the actual transaction may happen months or years ahead and the owner (usually a construction company) is trying to attract customers in the process. These cases should be identified (it is usually easy even when in textual description to search for keywords such as “under construction”) and excluded from the dataset.

When computing the cost of rents for the HICP the relevant concept is related to the whole stock of rented properties. Usually in a rent survey the statistical unit is the dwelling itself. Internet

⁹ A standard assumption is that the price for sale or rent that is asked when a property is entering the market is never smaller than the negotiated final price.

¹⁰ Anglin, P., 1999. Testing some theories of bargaining, working paper, University of Windsor

advertisements refer to new contracts only and while the two are connected (sooner or later old contracts expire and are renewed or replaced so rental market shifts are slowly incorporated in the total stock) they are different concepts. Internet advertisements however can provide data for useful new statistics including leading indicators.

2.2.6 Real estate classified advertisements: conditions for opening them to producers of official statistics

One particular feature of Internet advertisements is that all descriptive data that is useful for statistical purposes is published so any Internet user (human or internet bot) can retrieve it without conditions.

More information, connected with a specific advertisement, and referring to the person or company that places the advertisement is of course private but it is not important for statistical purposes.

In some cases when the site not only hosts advertisements but also provide services that facilitate the transaction as well more data may be available including the actual price for the transaction. This is particularly useful since it removes the need for modelling the list price. However this is not at all common and advertisement sites still just connect the seller and buyer in the real estate market very much like what the advertisement pages of newspapers are doing.

2.2.7 Real estate classified advertisements: conclusions

There is a high potential in using Internet advertisement in the production of current statistics on the housing price index and PPPs related to rental and owner occupied housing.

On the other side, there is some potential to using Internet advertisement in production of the owner occupied housing sub index of the HICP, although there are differences in concepts.

It is unlikely that data from Internet advertisements can replace the rent surveys for the HICP but they can provide helpful new indices and facilitate the survey itself.

2.3 Social media message data

The scope of this feasibility study is the investigation of the potential of using social media content for the production of statistical information, complementary to typical official statistics. While Facebook and Twitter are taken as examples, the methodological outline described here also holds for other text based social media based on the «post» concept, i.e. a short text posted by a user. These may include tweets, Facebook posts and their comments' threads, YouTube comments etc.

The main characteristics of the data provided by these sources are twofold. First, one has to deal with unstructured text data in natural language, which implies the usage of text analytics methodologies for the extraction of concepts and classifications, with all implications and ambiguities of natural, informal language. Second, in contrast to usual applications of text analytics, the text is short and has to be interpreted in context (i.e. in the context of a thread, discussion etc.) in order to be correctly interpreted and classified.

The main concept behind statistical data extraction from social media can be summarised in (a) the classification of the post in a domain of interest according to the existence of domain-specific keywords (and - through a thesaurus - their synonyms and derivatives); (b) the ranking / scoring of the positive or negative «sentiment» expressed by the post, again according to keywords; and (c) the calculation of a sentiment index based on the aggregation of individual posts' scores, over a specified period of time.

Obviously, this method might be applicable only to official statistics related to subjective perceptions, and for this reason, the applicability to the domain of Quality of Life statistics (where subjective indicators are often used) is investigated. It must be noted that even in this case, the sentiment index produced is not directly equivalent to current statistical indicators. While the latter concern percentages of the population reporting a specific ranking of a concept (such as happiness, trust to institutions etc), a sentiment index provides an overall measure of the sentiment changes over time. Nevertheless, since these indexes can be calculated in almost real-time at low cost (in contrast to costly interview and questionnaire based surveys), they might provide interesting complementary statistics.

2.3.1 Social media message data: presentation of the source

2.3.1.1 Facebook

Facebook is an online social networking service which allows anyone who claims to be at least 13 years old to become a registered user of the website. Users must register before using the site, after which they may create a personal profile, add other users as friends, exchange messages, and receive automatic notifications when they update their profile. Additionally, users may join common-interest user groups, organized by workplace, school or college, or other characteristics, and categorize their friends into lists such as "People From Work" or "Close Friends". As of September 2012, it has over one billion active users, of which 8.7% are fake¹¹. Facebook (as of 2012) has about 180 petabytes of data per year and grows by over half a petabyte every 24 hours.

The study is narrowed down to Facebook data related to status updates (Status message). Within a status message, a sentiment can be exported (e.g. happiness, sadness, frustration, etc). The data source is able to provide all status update data in a structured form. Each status message contains information about a single person (e.g. content of the status update, date, etc).

In this case, a dataset refers to all status updates. This is the only dataset that will be used in this context. Each status update contains the following variables¹²:

1. ID: The status message ID
2. User: The user who posted the message;
3. Text: The status message content. The information related to sentiment can be extracted from this field. Facebook has updated status message content so that it contains a direct expression of sentiment. The user may select from a list the sentiment to be included in his post (e.g. feeling happy, sad, enthusiastic, etc). Nevertheless, the user may choose not to include sentiment expressions in his status or he may insert his own sentiment expression. However, in the case that the user does not use a direct expression to express his sentiment, several tools can be used to extract the sentiment out of the status message content¹³.
4. Place: Location associated with a status, if any;
5. Update time: The time the message was public;
6. Type: The type of the status message (e.g. mobile_status_update, created_note, added_photos, added_video, shared_story, created_group, created_event, wall_post, app_created_story, published_story, tagged_in_photo, approved_friend).

¹¹ Sharwood, Simon (November 9, 2012): "Facebook warehousing 180 PETABYTES of data a year". The Register. Retrieved August 8, 2013.

¹² <https://developers.facebook.com/docs/reference/api/status/>

¹³ <http://sentistrength.wlv.ac.uk/>

Moreover, a status message has the following connections by a single person:

1. Comments: All of the comments on this message.
2. Likes: The users that have liked this message.

For confidentiality reasons, only status updates that have their privacy set to 'public' are retrieved.

Data related to status messages are logged since the beginning of network monitoring (usage of Facebook Public Feed API or other means).

Data is updated every time a new status message is created (only the status messages with privacy set to 'public' are recorded).

Retrieving Facebook data

Facebook provides various Application Programming Interfaces (APIs) for retrieving and processing its data. In particular, The Public Feed API provides a stream of user status updates and page status updates as they are posted to Facebook. Only status updates that have their privacy set to 'public' are included in the stream. The stream isn't available via an HTTP API endpoint, instead updates are sent to an external server over a dedicated HTTPS connection. The stream only includes basic data about the given post. From that basic data the user may use the graph API to request additional metadata to supplement the updates received through the public feed API. Since users may delete or modify their privacy settings after posts are streamed, the API also sends reference to these actions.

Access to the Public Feed API is restricted to a limited set of media publishers and usage requires prior approval by Facebook. The current list of partners includes: BuzzFeed, CNN, NBC's Today Show, BSkyB, Slate and Mass Relevance.

Data mentioned above are provided but only for status updates that have their privacy set to 'public'.

Mass Relevance and Graph API

Mass Relevance is the first and only social experience platform to gain full access to Facebook's Public Feed API for display in broadcast and on digital properties. With the Mass Relevance Platform, users can draw data from any social conversation or interaction that is happening in the world. Moreover, the user can discover the right content by sourcing data from keywords, specific user accounts, geo-locations, client apps and more. Benefit of using the platform is the ability to pull data from multiple sources including Twitter, Facebook, Instagram, Google+, Youtube and more.

Using Mass Relevance platform all status messages data/metadata available through the Public Feed and other APIs are reachable and available for public use.

Facebook also provides Graph API. The Graph API is the primary way to get data in and out of Facebook's social graph. It's a low-level HTTP-based API that can be used to query data, post new stories, upload photos and a variety of other tasks that an app might need to do.

Data provided by Mass Relevance platform are delivered using RSS/Atom feeds, Javascript, XML and JSON APIs.

When using the public feed API to receive updates, all public user status updates and page status updates are received, in near real-time, as they are posted to Facebook. These updates will be streamed in the form of XML-based objects that will provide a basic set of information about the particular post. Moreover, the graph API may also be used to request additional details about the post to supplement these objects.

2.3.1.2 Twitter

Twitter is an online social networking and microblogging service that enables users to send and read "tweets", which are text messages limited to 140 characters. Registered users can read and post tweets, but unregistered users can only read them. Users access Twitter through the website

Data to be exported from Twitter are structured data. During the research, all user tweets will be exported and a sentiment analysis will be performed. These tweets consist of text (at most 140 characters) that express the opinion, beliefs or feelings of the user who creates the tweet. Entities for Tweets provide structured data from Tweets including resolved URLs, media, hashtags and mentions without having to parse the text to extract that information¹⁴.

Micro-data: Tweets are the atomic building blocks of Twitter, 140-character status updates, each tweet is created by a single user, with additional associated metadata.

A dataset refers to all tweets. This is the only dataset that will be used in the research. Each tweet contains the following variables (list is not exhaustive¹⁵):

1. ID: the unique identifier for the tweet;
2. Text: the actual text of the tweet (often called as status update);
3. User: the user who posted the tweet¹⁶;
4. Contributors: users who contributed to the authorship of the tweet;
5. Geographic location of the tweet as reported by the user of the client application;
6. Time: timestamp of the time the tweet is created;
7. Retweet count: number of times the tweet is retweeted;

Retrieving Twitter data

Users on Twitter generate over 400 million Tweets everyday. Some of these Tweets are available to researchers and practitioners through public APIs at no cost. The following types of information can be extracted from Twitter:

- Information about a user,
- A user's network consisting of his connections,
- Tweets published by a user, and
- Search results on Twitter.

APIs to access Twitter data can be classified into two types based on their design and access method:

- **REST APIs** are based on the REST architecture now popularly used for designing web APIs. These APIs use the pull strategy for data retrieval. To collect information a user must explicitly request it. Twitter provides the search/tweets API to facilitate searching the Tweets. The search API takes words as queries and multiple queries can be combined as a comma separated list. Tweets from the previous week can be searched using this API. Requests to the API return an array of Tweet objects. Parameters can be used to select between the top ranked Tweets, the latest Tweets, or a combination of the two types of search results

¹⁴ <https://dev.twitter.com/docs/entities>

¹⁵ <https://dev.twitter.com/docs/platform-objects/tweets>

¹⁶ The profile of a user in Twitter does not contain any personal information that can be used to extract statistical reports (e.g. gender, age, etc)

matching the query. An application can make a total of 450 requests and up to 180 requests from a single authenticated user within a rate limit window.

- **Streaming APIs** provides a continuous stream of public information from Twitter. These APIs use the push strategy for data retrieval. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user. Using the Streaming API, we can search for keywords, hashtags, userids, and geographic bounding boxes simultaneously. The filter API facilitates this search and provides a continuous stream of Tweets matching the search criteria. The input is read in the form of a continuous stream and each Tweet is written to a file periodically. This behaviour can be modified as per the requirement of the application, such as storing and indexing the Tweets in a database. There are three key parameters:
 - Follow: a comma-separated list of userids to follow. Twitter returns all of their public Tweets in the stream.
 - Track: a comma-separated list of keywords to track. Multiple keywords are provided as a comma separated list.
 - Locations: a comma-separated list of geographic bounding box containing the coordinates of the southwest point and the northeast point as (longitude, latitude) pairs.

Streaming APIs limit the number of parameters, which can be supplied in one request. Up to 400 keywords, 25 geographic bounding boxes and 5,000 userids can be provided in one request. In addition, the API returns all matching documents up to a volume equal to the streaming cap. This cap is currently set to 1% of the total current volume of Tweets published on Twitter.

They have different capabilities and limitations with respect to what and how much information can be retrieved. The Streaming API has three types of endpoints:

- Public streams: These are streams containing the public tweets on Twitter.
- User streams: These are single-user streams, with to all the Tweets of a user.
- Site streams: These are multi-user streams and intended for applications which access Tweets from multiple users.

The rate limitations of Twitter APIs can be too restrictive for certain types of applications. To satisfy such requirements, Twitter Firehose provides access to 100% of the public Tweets on Twitter at a price. Firehose data can be purchased through third party resellers of Twitter data. Currently, there are three resellers of data, each of which provide different levels of access. In addition to Twitter data some of them also provide data from other social media platforms, which might be useful while building social media based systems. These include the following:

- DataSift¹⁷ - provides access to past data as well as streaming data
- GNIP¹⁸ - provides access to streaming data only
- Topsy¹⁹ - provides access to past data only

¹⁷ <http://datasift.com>

¹⁸ <http://gnip.com>

¹⁹ <http://topsy.com>

2.3.2 Social media message data: Related official statistics

Subjective well-being is an aspect of quality of life that can be complementary to other measures of progress such as income and living conditions – to which it is only indirectly connected – as it provides information on how people are feeling in the light of those circumstances (Eurofound, 2012). Subjective well-being is a self-perception on one's quality of life weighting up by its different aspects.

The underlying concepts of happiness and life satisfaction, central to subjective well-being, are different, the former referring more to emotional aspects and the latter to a more cognitive evaluation of life as a whole (Eurofound, 2003).

Eurostat's database does not include indicators that concern quality of life or assessments of European individual's sentiments. Relevant indicators are either under development in EU-SILC 2013 module on Well-Being or to be developed in other surveys not defined yet.

EU-SILC: 2013 ad-hoc module on well-being

In May 2010 both the Living Conditions Working Group and the Indicators Sub-Group of the Social Protection Committee supported Eurostat's proposal to collect micro data related with well-being within the 2013 module of SILC in order to better respond to this request. With the implementation of the 2013 module, data for subjective indicators will start to be collected as European statistics on a regular basis. In the long term, EU-SILC should be developed further to serve as the core EU instrument connecting the different dimensions of quality of life on individual level and reflecting their dynamic interdependencies.

The well-being ad-hoc modules will be developed in order to complement the variables permanently collected in EU-SILC with supplementary variables highlighting unexplored aspects of quality of life.

The variables collected through the survey's questionnaire are presented below. The 8 categories we divided them in, serve the conceptual purpose of the description of this feasibility study.

1. Self-appraisal of life as a whole, Meaning of life
 - PW010: Overall life satisfaction (from 0-10)
 - PW020: Meaning of life (from 0-10)
2. Financial Situation of the household/ Household needs
 - PW030: Satisfaction with financial situation (from 0-10)
 - PW040: Satisfaction with accommodation (from 0-10)
3. Emotional well-being
 - PW050: Being very nervous (all the time, most of the time, some of the time, a little of the time, none of the time, do not know)
 - PW060: Feeling down in the dumps (all the time, most of the time, some of the time, a little of the time, none of the time, do not know)
 - PW070: Feeling calm and peaceful (all the time, most of the time, some of the time, a little of the time, none of the time, do not know)
 - PW080: Feeling downhearted or depressed (all the time, most of the time, some of the time, a little of the time, none of the time, do not know)
 - PW090: Being Happy (all the time, most of the time, some of the time, a little of the time, none of the time, do not know)

4. Professional activities and commuting time
 - PW100: Job Satisfaction (from 0-10)
 - PW110: Satisfaction with commuting time (from 0-10)
5. Time use
 - PW120: Satisfaction with time use (from 0-10)
6. Basic rights (Trust on the political, legal system and the police)
 - PW130: Trust in the political system (from 0-10)
 - PW140: Trust in the legal system (from 0-10)
 - PW150: Trust in the police (from 0-10)
7. Social interactions – social activities
 - PW160: Satisfaction with personal relationships (from 0-10)
 - PW170: Personal matters (anyone to discuss with) (Yes/No)
 - PW180: Help from others (Yes/No)
 - PW190: Trust in others (from 0-10)
8. Living environment
 - PW200: Satisfaction with recreational or green areas (from 0-10)
 - PW210: Satisfaction with living environment (from 0-10)
 - PW220: Physical security (Very safe, fairly safe, a bit unsafe, very unsafe, do not know)

Reference population: Information should be provided for all current household members, or if applicable, for all selected respondents, aged 16 and over.

Mode of data collection: personal interviews

Reference period: The reference period for all target variables is the current situation, except for the five variables on emotional well-being, which refer to the past four weeks.

Moreover, Eurostat produces publications using data from third parties. Overall satisfaction can be estimated from the database of Eurofound. Specifically, the results come from Eurofound's European Quality of Life Surveys (EQLS). The EQL surveys provide data on issues such as employment, education, housing, family life, health and life satisfaction.

In the questionnaire of the EQLS, respondents have to answer of how often they are affected negatively for example feeling lonely, downhearted and depressed or particular tense. The frequency was recorded by having answers with range from "at no time" to "all of the time". In addition, respondents have to answer whether they feel happy on scale 1 to 10, with the highest score is supposed to be a very happy person.

Messages from social media, for instance Facebook and Twitter, could be used as for estimating generally how individuals feel. The potential of obtaining estimates or figures on a daily or monthly basis can be approached, as opposed to the surveys mentioned previously, which data are available every 3 or 5 years. An exceptional example is Statistical service in the Netherlands, which analyses social media messages, to estimate a statistically significant relation between the sentiment towards the economic situation in Dutch social media and the Dutch consumer confidence.

2.3.3 Social media message data: feasibility of their use as input for official statistics

This section investigates the feasibility of deriving information on self-perceived/subjective topics i.e. happiness, job satisfaction, trust towards the legal system etc., and use it complementarily to official statistics.

Social media content may be exploited as a source for perception measurements, due to the voluntary expression of opinions and feelings by users. They provide data that are being characterized by great volume, extreme variety and rapidity.

Sentiment Analysis

In order to explain the sentiment analysis better we will describe the process by using as an example the extraction of text messages related to the third category of the questions of EU-SILC ad-hoc module of 2013, related with the emotional well-being of an individual (feeling happy, nervous, peaceful depressed etc.). Below we describe the steps of this process, which can be repeated for the rest of the categories of the ad hoc module on well-being.

Post classification is based on a domain-specific thesaurus that includes words (key words as well as their derivatives, synonyms etc) that are related with the specific domain (e.g. emotional well-being, trust to institutions, recreation etc.). The thesaurus also contains key words that provide negative or positive sentiments.

For each post a two-step procedure is followed:

1. Classification of the content domain (i.e. relevance to the specific statistical concepts for which indicators are to be calculated) according to the existence of domain-specific keywords;
2. Sentiment scoring or ranking, according to the existence of sentiment-specific keywords.

With the help of existing text classification algorithms (i.e. `classify_emotion` and `classify_polarity` in R “sentiment” package) the sentiment strength of the post can be analyzed and classified (different types of emotion: happiness, sadness, fear, joy etc. polarity: positive, neutral or negative). An example of a sentiment analysis algorithm is Naïve Bayes Classifier.

Final word scoring: Each word that will represent an emotion and will be classified according to its polarity then, it will be scored accordingly (+1, 0, -1, if the word is positive, neutral or negative respectively).

A simple example is shown in the table below:

Text	Emotion	Polarity	Score
I feel happy today	Happiness	Positive	+1
I just had my breakfast	Unknown	Neutral	0
It's raining and it's miserable!	Depressed	Negative	-1

The simplest answer to this question is to develop a scoring method. Each one of the keywords related to a feeling on a matter/topic that will appear on a user's profile or on a topic will be scored using simple emotion modelling. Simple emotion modelling combines a statistically based classifier with a dynamic model. The Naïve Bayes classifier employs single words and word pairs as features. It

allocates user utterances into positive, negative and neutral classes, labelled +1, -1 and 0 respectively.

Sentiment estimation results in a final score, which is computed as the sum of the scores of the individual words in every distinct topic/domain over a specific period of time. By using a specific point in time as a basis, an index may be constructed, providing comparisons over time.

2.3.4 Main Advantages

Subjective well-being and more specifically variables which include positive or negative moods and emotions like happiness or perceived mental health are very sensitive to changes over time. Short-term and long-term changes in subjective well-being should be separately assessed whenever possible²⁰.

Our hypothesis is that the feelings are visible on social media by an increased fraction of posts containing specific words-moods/opinions referring to one of the 8 categories of well-being (1.1.2) for example happiness, job satisfaction, trust in the legal system etc.

- Due to the fact that the above subjective information is collected and published by Eurostat in an ad-hoc way (i.e. the Well-Being module 2013), social media data can be used in a complementary way since they offer large samples of data whose trends can be explored over time.
- Social media provide a great volume of data, of the order of magnitude of millions of posts per day.
- There is an extreme variety of data due to the fact that new tweets are constantly being added.
- Data are being updated rapidly.

2.3.5 Issues that may make difficult the use of social media data for the computation of subjective well-being indicators

Creation of domain-specific thesaurus

The main problem with social media data is the complexity of detecting keywords and classifying / ranking posts.

The simplest way to solve this is to develop a thesaurus of associated words that express the positive and negative opinions and moods (stemming* algorithm) that are being created for each of the categories set on table above (section 2.3.3). Along with the keywords it is necessary to find the words that frequently appear in tweets containing the certain keyword of interest.

The procedure of finding salient words can be performed automatically with a t-test, which compares the probability of a word co-occurring with a keyword, $P(\text{word} \mid \text{keyword})$, with the overall probability of the word $P(\text{word})$. Words that co-occur with specific keywords that express certain feelings will also be included in the thesaurus. The above process will result in the creation of a «word cloud» around the main keyword of interest.

Population Coverage Issues

Another restriction is to identify the characteristics of the population we are interested in. One way to address this problem is by focusing in the users' profiles.

²⁰ E. Diener, Guidelines for National Indicators of Subjective Well-Being and Ill-Being, 2005

- Location Identification

In Twitter identification of users' country of origin may be erratic since it must be based on the content of the location field in their profile. Nevertheless, additional information provided by the API, the language of the tweet or other context-based classification may improve the accuracy.

- Age and Sex Identification

The API does not offer demographic characteristics for each Twitter user; such as sex and age, although it is possible by using estimators to automatically classify twitter users into age and sex categories. Based only on tweets the use of certain words, the name of the user as well as the variation on the language use can predict the gender and the age of the individual.

- Other Issues

One critical issue might be that social media users may tend to misrepresent their emotions and opinions to their friends in order to feel accepted by their friends. However this kind of bias is happening in all the data collection methods that collect subjective information. Satisfaction data from wherever they are collected are biased by varying participant attitudes towards the interview itself.

Previous sentiment analysis on Facebook data has shown that especially the feeling of happiness maximizes its peaks around holidays and other special days. For example the phrase "I am very happy today" and the conventional phrase "Happy New Year" do not weight the same. It is feasible to deal with this issue by eliminating from our analysis words related with specific occasions.

Note on the usage of Facebook

Facebook has some characteristics that are different from other social media, and can provide richer metadata. These are:

- The users' registration form collects demographic information such as place of residence, sex and age of the user
- Facebook distinguishes among status updates, comments, links to external multimedia contents and full articles
- Status updates are usually connected to other social activities that users can take, for example, users can either "like" or comment on a status update

As far as the happiness feeling is concerned, Facebook itself developed a Gross National Happiness Index (GNHI) that measures how happy Facebook users are from day-to-day by looking at the number of positive and negative words they're using when updating their status. When people in their status updates use more positive words—or fewer negative words—then that day as a whole is counted as happier than usual.

Privacy Issues

Another issue that derives from Facebook data analysis is that we can include in our analysis only the profiles set as public. Facebook has a very strict privacy policy regarding about user data. Eurostat can make an arrangement with Facebook for obtaining access to their data, after all personally identifiable information has been removed for confidentiality issues, that they could be used complementary as input for statistical indicators about well-being and life satisfaction.

2.3.6 Social media message data: Conclusions

There are a lot of benefits from using social media in the production of subjective indicators, which are used in the current statistics.

It is worth noting that Twitter and Facebook are two potential fascinating sources of sentiment information, however it is important to highlight that those sentiments cannot replace the existing official statistics and its indicators.

The measures of sentiments and their scoring can be used complementary to official statistics and provide us with useful trends over time as well as with comparisons among the different European countries.

2.4 Credit card transaction data (Visa Europe)

2.4.1 Credit card transaction data: presentation of the source

Visa Europe Ltd. is a membership association of more than 4,000 European banks and other payment service providers that operate Visa branded products and services within Europe. It is comprised of 36 countries across Europe, the EU states, and non-EU countries (Andorra, Iceland, Liechtenstein, Norway, Switzerland, Turkey, Israel, Greenland and Gibraltar).

Visa Europe has issued more than 419 million Visa debit, credit and commercial cards in Europe. Visa/PLUS is also one of the world's largest global ATM networks, offering cash access in local currency in over 200 countries.

In addition to its well-known transaction processing services, Visa is able to respond quickly to the specific market needs of European Banks and their customers – cardholders and retailers. Payment security knowledge is also offered to business and government.

Visa's network also runs many information services such as business intelligence and report generation, as well as risk management services such as fraud monitoring and encryption. In fact, every transaction is checked against 100 fraud-detection parameters in real-time.

Data from Visa concern daily transactions of each credit card holder. These are accompanied by each holder's personal data. The data recorded for each transaction are:

- Card credit number
- Customer Identification number
- Date and time of transaction
- Transaction type
- Expense type
- Total amount of transaction
- Transaction currency
- Country where the transaction took place
- Value-added tax rate of transaction
- Exchange rate (seven decimal points)
- Description of service provider
- Visa type (debit, credit, prepaid)

Additionally, the following information is requested when applying for a visa card:

- Full name and father's name
- Identity card or passport number

- Issuing authority
- Data and place of birth
- Customer's Income
- VAT Registration Number
- Current home address and Telephone number
- Profession and current business address

2.4.2 Visa Europe: EU Consumer Spending Barometer

Visa Europe already compiles an Index, named "EU Consumer Spending Barometer" using real-time card transaction data. Its aim is to provide a robust indicator of total consumer expenditure at a European level. Through this index a uniquely comprehensive and timely insight of the consumer spending across all payment methods is provided. Currently, it is used by a range of stakeholders to gain insights into consumer spending.

The Barometer is compiled using Visa's data on transactions at EU level for a reference quarter. A report²¹, analysing the trends of household consumption in the EU is published 2 months after the end of the reference quarter. Similarly, two indices, namely the UK Expenditure Index²² and the Sweden Expenditure Index²³ are compiled to reflect consumer spending in the UK and Sweden, respectively.

About the EU Consumer Spending Barometer

Visa's EU Consumer Spending Barometer is based on actual spend data on all Visa debit, credit and prepaid cards. These are adjusted to allow for Visa card insurance, consumer payment preferences and inflation. The index is compiled by Markit²⁴, a private company providing financial information services, on behalf of Visa Europe. A model has been developed for adjusting raw Visa transaction data for a number of factors and for ensuring that the data provide an accurate indication of consumer spending trends.

More specifically, data on transactions are firstly deflated by changes in the number of Visa cards in order to account for the expansion of Visa's card operations (deflating the data by changes in Visa card numbers helps to provide a better indication of the underlying nominal spending patterns), particularly on the debit side. At a second stage, data are adjusted to offset changing consumer preferences for card usage. This is based on an assessment of the trends in cash withdrawals and point-of-sale (POS) transactions on Visa cards. The data are then deflated by changes in the consumer price index.

POS Transactions

In Europe, there are more than 419 million Visa cards. Specifically for Visa debit card, the average number of transactions per card was 13.7 in the first quarter ending March 2010 (Figure 1).

The data highlight the growing role that debit cards play in consumer spending behaviours. A significant percentage of consumer spending in Europe, 11.2%, concerns point-of-sale transactions with a Visa card, of which more than 70% is with Visa debit cards. Currently, more than €1.5 millions every minute in Europe are spend on Visa-branded credit cards. This signifies that visa transaction

²¹ http://www.visaeurope.com/en/newsroom/all_reports/european.aspx

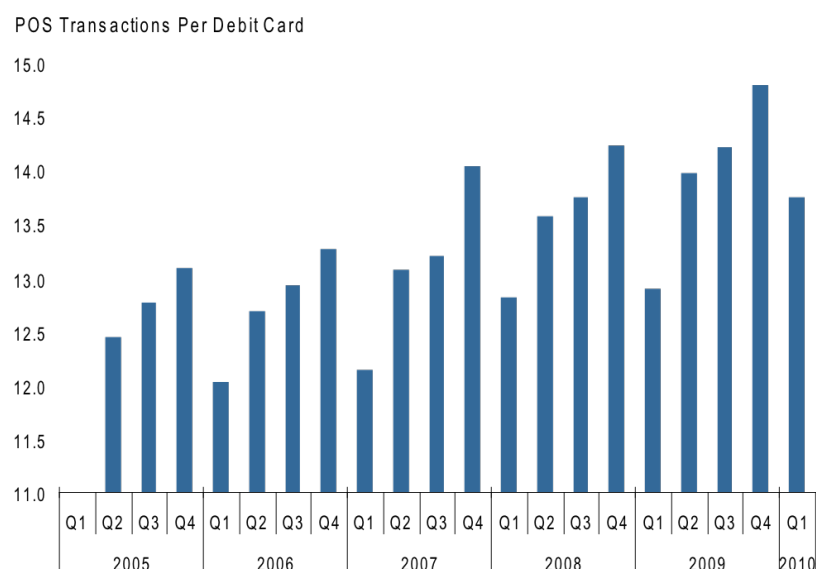
²² http://www.visaeurope.com/en/newsroom/all_reports/uk.aspx

²³ http://www.visaeurope.com/en/newsroom/all_reports/sweden.aspx

²⁴ <http://www.markiteconomics.com>

data can provide a strong indicator of total spending. Taking also into consideration the speed with which Visa data can be processed and analyzed, the indicator can provide a timely insight into the spending patterns of EU consumers.

Figure 1. Average POS Transactions per Visa debit card.



Source: Visa Quarterly report on European Spending Trends (May 2010)²⁵

Visa's EU Barometer and official data

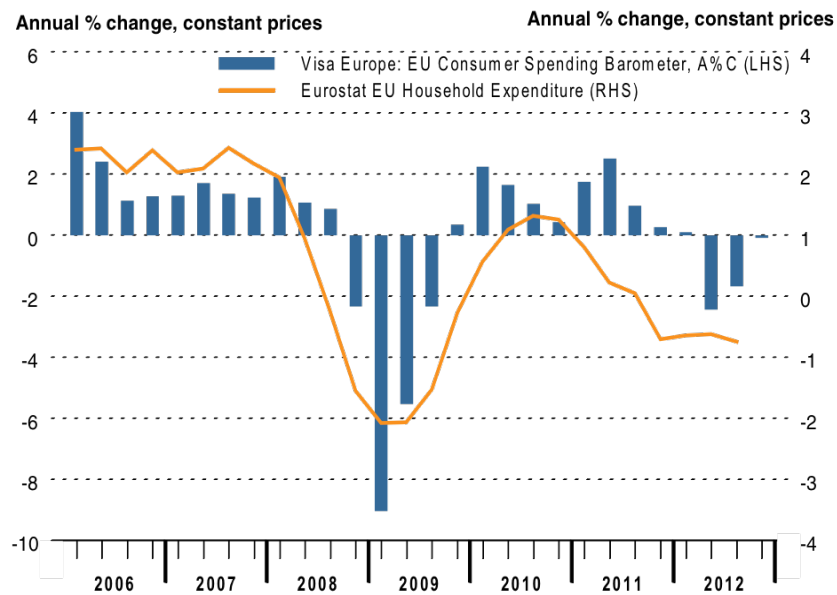
Data from the EU Consumer Spending Barometer indicate a strong relationship over time with the relevant official household spending data.

Figure 2 provides an indication of the relationship between the two data series, although data from Visa's Barometer tend to move in a wider range than the equivalent official data. The latter, may be attributed to different factors, such as the tendency to use cards for higher valued purchases or different attitudes to card usage across age groups.

Additionally, Visa's Barometer is positively correlated with Eurostat's Gross Domestic Product (GDP). As it is shown in Figure 3 the two data series have a similar trend over time (2006-2012). This is foreseeable considering that the consumer expenditure constitutes a significant part of the total economy.

²⁵ <http://www.visaeurope.com/en/idoc.ashx?docid=0f9b8b34-2bb9-4d08-a184-6b6d7b649c4e&version=-1>

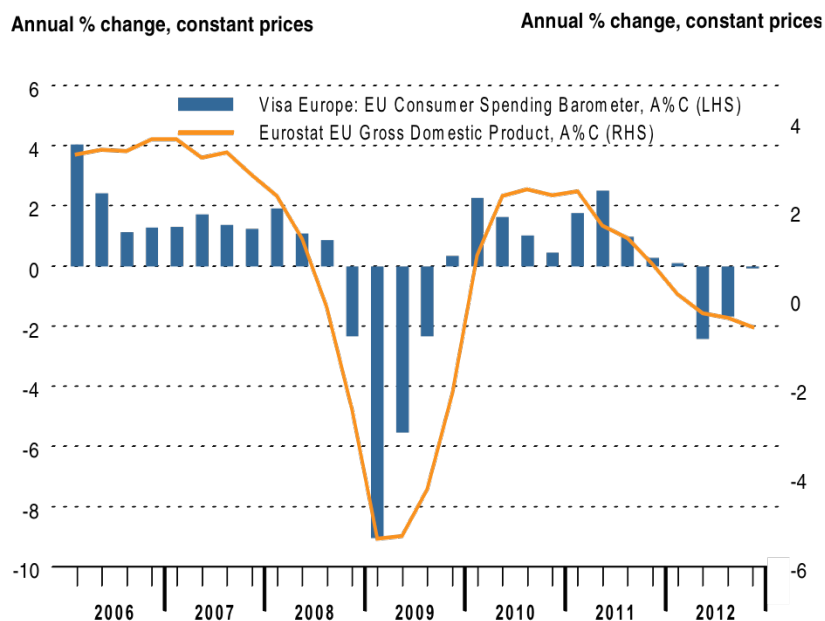
Figure 2. Year-on-year relative change of Visa Europe's EU Consumer Spending Barometer (left-hand-side) and EU Household Expenditure (right-hand side).



Sources: Visa Europe, Eurostat

Source: Visa Quarterly report on EU Consumer Spending Barometer (March 2013)²⁶

Figure 3. Year-on-year relative change of Visa Europe's EU Consumer Spending Barometer (left-hand-side) and EU Gross Domestic product (right-hand side).



Sources: Visa Europe, Eurostat

Source: Visa Quarterly report on EU Consumer Spending Barometer (March 2013)

²⁶ <http://www.visaeurope.com/idoc.ashx?docid=7974d478-c525-4211-8e32-06660f7392f9&version=-1>

Spending by product categories

Although, Visa has not published a report on EU Consumer Spending Barometer by product categories, the relevant indices for UK and Sweden are compiled by product categories. These categories consist of the following standard Classification of Individual Consumption, according to Purpose (COICOP) groups,²⁷ which are in accordance with Eurostat's classification:

Product Category	COICOP Group
Food, Beverage & Tobacco	1, 2
Clothing & Footwear	3
Housing & Household Goods	4, 5
Health & Education	6, 10
Transport & Communication	7, 8
Recreation & Culture	9
Hotels & Restaurants	11
Miscellaneous Goods & Services (including Visa card spend n.e.c.)	12

Therefore, the computation of the Visa's EU Barometer according to COICOP classification is feasible and provides an indication of the practical and computational feasibility of producing/supplementing the official statistics according to the official COICOP classification.

2.4.3 Credit card transaction data: related official statistics

Consumption expenditure is what people, acting either individually or collectively, spend on goods and services to satisfy their needs.

Data on consumption expenditure combine three sources in Eurostat's database: (a) the Household Budget Survey (HBS), (b) National Accounts (NA) and (c) the Harmonised Index of Consumer Prices (HICP). These are organized according to the Classification of individual consumption by purpose (COICOP).

HBS and NA provide information both on amounts and on the structure of the consumption expenditure, whilst the HICP provides only a structure of the expenditure. In fact, the HBS shows amounts of expenditure per household and per adult equivalent in PPS, whilst the NA show data in current prices and volumes, as well as price indices. The three sources are related, but they do show some differences, due to the way that data are collected, differing definitions and the publishing timeliness.

The HBS deals with households and all the information is gathered directly from them. The information about consumption expenditure is accompanied with information about the income, place of residence, and some characteristics of the reference person.

On the other side, the NA rely on several sources to estimate consumption expenditure, both from the demand and supply sides. This information is published much more frequently and is more

²⁷ <http://www.visaeurope.com/idoc.ashx?docid=94e6248b-10eb-44a9-b1ff-7a960feb379f&version=-1>
<http://www.visaeurope.com/idoc.ashx?docid=8b39fbb2-e15d-4a52-90fc-8ddac4737cf4&version=-1>

recent than the HBS. However, NA cover expenditure from a macro level and thus expenditure cannot be correlated with characteristics pertaining to different households.

From the brief description of the source and consumer expenditure statistics produced by Eurostat, it can be drawn the conclusion that the statistics compiled on the basis of the HBS data can be compiled based on Visa credit card transaction data.

2.4.3.1 Household Budget Survey (HBS)

Background information

The HBS is among the most comprehensive household surveys, conducted in all Member States of the Union. The HBS mainly focuses on consumption expenditure of households on goods and services. Its primary aim (especially at national level) is to calculate weights for the Consumer Price Index (used as measures of inflation).

As its name implies, the HBS is a survey, which is run on a sample of households (big institutions, such as hospitals, hotels, institutes and prisons are excluded) in the participating countries and collected, aggregated and published by Eurostat on an informal basis.

Data collection involves a combination of one or more interviews and diaries or logs maintained by households and/or individuals, generally on a daily basis. The basic unit of data collection and analysis in the surveys is the household. However, the reference person is often the head (or reference person) of the household (i.e. the person designated in each original national survey)²⁸. The socio-economic group, occupation and employment status, income, sex and age of the reference person are often used to classify and present results.

There are two relevant conceptual bases in the European System of Accounts (ESA) for household consumption expenditures²⁹:

- **household final consumption:** the acquisitions households obtain through their spending on consumption goods and services in their own country or abroad;
- **household actual final consumption:** household final consumption and, in addition, acquisitions from the government and non-profit institutions serving households, which are essentially provisions in kind to the households.

Taking into consideration the practical difficulties for the measurement of the 'household actual final consumption' in many Member States, Eurostat recommends that the 'household final consumption expenditure' as the basic conceptual basis of the Household Budget surveys.

Household final consumption expenditure has a monetary and a non-monetary part. The monetary part covers all cash payments, whereas the non-monetary part includes (a) services of owner-occupied dwellings (measured as an imputed rent) and (b) income in kind, such as goods and services received as income in kind by employees or goods or services produced as outputs of incorporated enterprises owned by households that are retained for consumption by members of the household.

Statistics disseminated in Eurostat's disseminated database

²⁸ A common practice used in some countries (Ireland, Luxembourg, Portugal and Finland) is to consider as head, the person designated as such by the household concerned. Some countries use more objective and specific criteria such as the person contributing most to the income of the household (Belgium, Denmark, Germany, the Netherlands, Austria and Spain); the person owning or renting the household accommodation (United Kingdom); or the oldest active male (Greece).

²⁹ Eurostat (2003) *Household Budget Surveys in the EU. Methodology and recommendations for harmonisation*. Luxembourg: Office for Official Publications of the European Communities.

Eurostat's consumption expenditure of private households statistics provide data about: (a) the mean consumption expenditure for households and per adult equivalent and (b) the structure of consumption expenditure, (c) households' characteristics.

Thus, consumption expenditure as an indicator of the standards of living of the households is studied both in level and in structure. In level, the average expenditure is analyzed and is expressed in Purchasing Power Standard (PPS). The structure of consumption expenditure aims to determine the share of the total consumption expenditure devoted by a household to a particular type of consumption.

The statistics are disseminated broken down by degree of urbanization, detailed COICOP, by employment status of reference person, number of active persons, income quintile, age of the reference person, type of household, main source of household's income.

Additional data about households' characteristics (covering data about the distribution of households, number of households in the sample, average household size and number of adult equivalents) are also disseminated broken down by employment status and age of the reference person.

As already mentioned, the HBS collects information on Consumption Expenditure according to the Classification of Individual Consumption by Purpose (COICOP). The main divisions of COICOP include:

- Food and non-alcoholic beverages (CP01)
- Alcoholic beverages, tobacco and narcotics (CP02)
- Clothing and footwear (CP03)
- Housing, water, electricity, gas and other fuels (CP04)
- Furnishings, household equipment and routine maintenance of the house (CP05)
- Health (CP06)
- Transport (CP07)
- Communications (CP08)
- Recreation and culture (CP09)
- Education (CP10)
- Restaurants and hotels (CP11)
- Miscellaneous goods and services (CP12)

Information about expenditure on insurance and gambling is not collected. Besides, information on consumption expenditure on COICOP headings linked to activities considered as non-socially correct (e.g. consumption of alcoholic beverages, narcotics or prostitution) is usually under-reported by the surveyed households. Therefore, these figures are not reliable.

Quality of disseminated statistics

The data are collected approximately every five years. It takes between one to four years after the end of the reference period to be published. Since there is no legal basis, there are many methodological issues, which restrict the comparability of the data across countries. Efforts are made after each collection round to increase the harmonisation of these statistics.

2.4.4 Credit card transaction data: feasibility of their use as input for official statistics

This section investigates the feasibility of producing or supplementing Eurostat's consumption expenditure statistics based on credit card transaction data from Visa.

Visa debit, credit and prepaid cards are mature payment instruments used by hundreds of millions of consumers in the EU for a number of transactions (either for high or low valued purchases). Therefore, Visa Europe can be exploited as a source for deriving information about the amount and structure of the consumption expenditure of households.

The EU Consumer Spending Barometer compiled by Visa provides a proof of concept of how a relevant but more elaborated Index, which would fit Eurostat's needs, could be feasible to be produced from Visa's data.

Principles of computation

Eurostat in cooperation with Visa can use the EU Barometer as a prototype for the production of an Index about EU Consumer Spending accompanied by demographic and household characteristics of the reference population and according to the official COICOP classification.

- A. In fact, Visa can provide information not only about the total spending and number of transactions of cardholders, but also about their profiles and characteristics (e.g. age, sex, income, marital status, etc.). This information is recorded when applying for a Visa card. Additionally, information about the number and type of Visa cards owned by each household is also available (or is estimated based on a model).
- B. Information about the type of the total expenditure can be deduced, with a high probability, from the type of merchant (e.g. clothing stores, secretarial schools and business, physicians and pharmacies, restaurants, etc.) that the transaction is made. This permits to classify the expenditure at the relevant product category. Visa already uses the official standard COICOP classification for the categorisation of products into categories.
- C. Based on the penetration of card usage in each country, the different attributes to card usage (such as cardholders' age, income, etc.), as well as COICOP category, a weight indicating the intensity of card usage can be allocated to each category.
- D. To account finally for inflation, data should be deflated by changes in the Consumer Price Index for each given COICOP category.

Based on these principles, a model, which would use as input all the above-mentioned information, can be developed by domain experts to estimate the amount and structure of household expenditure. The estimates produced should be validated by comparing them with Eurostat's actual data and cross-checked with NA data.

2.4.5 Credit card transaction data: conditions for opening them to producers of official statistics

Visa's data are imposed to privacy and confidentiality restrictions. The compilation of consumption expenditure statistics from Visa's data can only be achieved in cooperation with Visa, providing that Visa undertakes the computation of the required data. Eurostat can make an arrangement with Visa for obtaining access to its aggregated data; after all personally identifiable information at the individual cardholder level or individual merchant outlet level has been removed. Taking into consideration that Visa already produces indices based on these data, it is very probable that Visa provides these data at a regular and frequent basis.

2.4.6 Credit card transaction data: conclusions

There are a lot of benefits from using Visa's data in the production of consumption expenditure statistics. Currently, the HBS survey from which input data come is carried out at an informal basis every five years.

It is worthwhile using Visa as a source, in a complementary way, for the production of flash estimates about the structure and amount of consumption expenditure. However, it is important to highlight that an Index similar to Visa's Barometer, cannot replace the existing official statistics and its indicators.

Although, such a Barometer can be used complementary to official statistics, it can only provide a robust indication of real consumer spending trends over time and among the different EU countries.

2.5 Government financial transparency portal data

2.5.1 Financial transparency portal data: presentation of the source

In 2010 an important legislation aiming at improving the transparency of public administration was enacted in Greece. According to law 3861/2010 government agencies are obliged to upload their decisions on the Internet, through the «Clarity» («διδύχεια») site. The law³⁰ ensures that a broad range of decisions of public entities are not enforceable if they are not first uploaded on the «διδύχεια» website.

There are similar datasets in many countries depending on relevant transparency legislation. Notably in the UK the office of publications provides 450,000 post-1980 records from over 2000 public bodies as well as distributed records from the websites of public entities. These are aggregated by independent initiatives.³¹ However, «διδύχεια» is unique in the sense that it includes the totality of decisions, in a centralised infrastructure and in harmonised way.

«διδύχεια» covers all public institutions, regulatory authorities and local government; in all as of 2013 there were 3900 public entities registering 2.141 million decisions in the system. The «διδύχεια» program introduces the obligation to publish all the decisions on the Internet, with the exception of decisions that contain sensitive personal data and/or information on national security. The use of Internet guarantees openness and access to information, progressively contributing to a culture change in the whole of the Public Administration.

Uploading is done by the public entities and each uploaded document is digitally signed and assigned a transaction unique number automatically by the system.

The data of this source are produced by Public entities thus it belongs to the traditional business systems type of Big data. The data source is able to provide all data in a structured form in XML format. Each dataset contains information about a single decision (e.g. protocol number, date, etc). As a result, data is updated constantly whenever a new decision is issued.

Each decision issued by a public entity and published in («διδύχεια») contains at least the following metadata:

- Protocol Number
- Issue date
- Subject of the decision

³⁰ As amended by law 4210/2013

³¹ <http://wheredoesmymoneygo.org/>

- Email address of the Decision Registrar
- Organisation ID
- Organisation Unit ID
- Decision type
- Various tags³²
- Signer ID
- Relative Government Gazette Issue (FEK), etc.
- For each decision related to expenses the following metadata are also available:
 - Type of VAT Registration Number of the Entity (Payer)
 - VAT Registration Number of the Entity (Payer)
 - Legal name of the Entity (Payer)
 - Type of VAT Registration Number of the Contractor (Payee)
 - VAT Registration Number of the Contractor (Payee)
 - Name of the Contractor (Payee)
 - Short description of the decision's content
 - Amount of the expense/transaction (including VAT)
 - Common Procurement Vocabulary (CPV code)
 - Expense Code Number (based on the national budget classification of income and expenses)
 - Category of the Expense (this determines the stage of a payment)

The content of the site is huge. An analysis of the information that was obtained by the publicspending.net initiative included approximately 2 million payment decisions valued 44.5 billion Euros that have been paid from 3,900 payers to 204,000 payees and form 63 million triples³³.

2.5.2 Financial transparency portal data: related official statistics

Government finance statistics (GFS) data show the economic activities of government in a harmonized and comparable way. They differ noticeably from the budget presentations or public accounting presentations that are nationally specific and not harmonized between countries. GFS data include both the financial (borrowing and lending) and non-financial (income and expenditure) activities of government.

Government Finance Statistics are found in the theme, Economy and finance, of Eurostat's Data Navigation Tree, which are presented in millions of Euro, millions of national currency units and percentages of GDP. The main indicators and their breakdowns of this theme are the following:

1. Government expenditure by COFOG function and type notified by national authorities (annual data)

³² Clarity supports various tags that can be assigned to a decision. Each decision may contain several tags. A list of all supported tags can be found in the following link: <http://opendata.diaugeia.gov.gr/api/tags.xml>

³³ Vafopoulos, M., Meimaris, M., Anagnostopoulos, I., Papantoniou, A., Xidias, I., Alexiou, G., ... & Loumos, V. (2013). Public spending as LOD: the case of Greece. Semantic Web Journal. Available at <http://www.semantic-web-journal.net/system/files/swj464.pdf>

2. Main revenue and expenditure items of the general government sector,³⁴ notified by national authorities (annual data)
3. General government total expenditure and total revenue, as well as their breakdowns by ESA95 categories and the resulting quarterly government deficit/surplus (quarterly data).

Moreover, the data for the computation of GFS usually derive from annual national accounts, national authorities, administrative and other records of general government.

According to the European System of Accounts 1995 (ESA 95), the categories that comprise the total general government expenditure are the following:

- Intermediate consumption: the purchase of goods and services by government;
- Gross capital formation: gross fixed capital formation, changes in inventories, acquisitions less disposals of valuables
- Compensation of employees: the gross wages of government employees plus non-wage costs such as social contributions
- Other taxes on production
- Subsidies payable
- Property income: interest, payable and other property income, payable
- Current taxes on income, wealth, etc.
- Social benefits other than social transfers in kind
- Social transfers in kind related to expenditure on products supplied to households via market producers
- Other current transfers
- Adjustment for the change in net equity of households in pension fund reserves
- Capital transfers payable
- Acquisitions less disposals of non-financial non-produced assets

Currently, the Greek NSI (ELSTAT) is collecting, every quarter as well as annually, data from ministries that refer to all entities under each ministry's jurisdiction and include:

- Characteristics of each entity (VAT number, legal framework, number of employees etc)
- Its debt if it is allowed to borrow
- Its income from sales including subsidies and excluding taxes in accrual basis.
- Government grants received
- Expenses incurred in accrual basis and broken down in expense categories (salaries, intermediate consumption, taxes etc)

The data collection is based on the statistical law and in specific agreements between the NSI and each ministry that sets the content, responsibilities and standards for the data and its transmission.

³⁴ Sub-sectors of general government: central government, state government, local government and social security funds

2.5.3 Financial transparency portal data: feasibility of their use as input for official statistics

The Financially transparency portal «[δι@ύγεια](http://diavgeia.gov.gr)» can in technical terms be used for national accounting purposes as it includes financial information of great detail and also uses a publicly available API. The API uses RESTful-like calls and returns the data in XML format, according to a published XSD.³⁵ However, there are some issues both conceptual and methodological that need to be addressed during data processing so that official statistics about public finances can be produced.

2.5.3.1 Coverage

The transparency legal framework applies to all public entities and to entities owned by the state as well as entities that are receiving regular funding for at least 50% of their budget. This is in agreement with the delineation of the public sector in ESA95 and in practice all organisations that are included in the public entities list of the Greek NSI³⁶ are required to publish the relevant information in «[δι@ύγεια](http://diavgeia.gov.gr)». There are, however, some exclusions from this obligation that affect data coverage, if only marginally

- Some public entities are excluded from the obligation, including the presidency of the Republic and the Parliament.
- Some decisions are excluded from the requirement for publishing. Exclusions are explicitly stated for purposes of protecting sensitive personal data³⁷ as well as classified information including state and company secrets. It is not clear if these restrictions affect the publication of the decision per se or parts of it thereof that are considered classified.

The financial information contained in «[δι@ύγεια](http://diavgeia.gov.gr)» is included in published decisions, so only expenditures that require a decision are published each time they occur; recurrent expenses such as salaries of permanent personnel are included in the published documents once e.g. when a salary level is decided (upon hiring, promoting etc) and implemented without additional records at later times. Therefore, while there is a huge amount of information, coverage is not complete and does not include all parts of government expenditure evenly. Public procurement is covered at a very high level but expenditure on salaries, remuneration or pensions are not.

2.5.3.2 Accuracy issues

While the source is authoritative³⁸ the text of the decision and the way data is included creates issues that should be addressed.

Double counting. Based on public accounting rules there are 5 stages for each payment, in which the public entity decides to undertake the obligation, clears out whether the undertaking is lawful, issues a payment order and finally executes payment. All these decisions are recorded for each payment usually with the same amount. The problem is that currently the type of decision is a field that is not required to be filled and thus a payment can be counted more than once. It is important that each stage is identified as such and the connection between stages is established so that double counting can be avoided.

³⁵ <http://opendata.diavgeia.gov.gr/?lang=en>

³⁶ Available for years 2010-2013 at:

http://www.statistics.gr/portal/page/portal/ESYE/BUCKET/A0701/PressReleases/A0701_SEL08_DT_AH_00_2013_00_2013_01AB_F_GR.pdf

³⁷ According to law 3471/2006 "Sensitive data" shall mean the data referring to racial or ethnic origin, political opinions, religious or philosophical beliefs, membership to an association or trade-union, health, social welfare and sexual life as well as criminal charges or convictions and membership to a society relating to the above.

³⁸ In case of conflicting versions of a decision's text the published version is considered the authentic, by law.

Number format. The entry that refers to the amount is coded as text. This and the fact that data entry is manual means that the amount is entered in a great variety of formats (e.g. Commas and dots are used interchangeably for thousands separator and decimal point, the euro symbol and sometimes the word is entered). The problem is more acute with the number of decimals that can be zero, one or two. This issue is a problem but is a tractable one and in any case is expected to be overcome with a new version of the «διδύμεια» system currently under development.

Validation of metadata (codes). Some fields correspond to codes from a classification (e.g. CPV). Currently, there is no validation of each entry and typing errors may occur. An automated system will not have the ability to assign the correct category. A validation system is expected to be implemented in the revised version of the system.

Expense and income codification is currently an input field but it is not required in order to complete the entry. Although in practise it is included in most decisions it is not guaranteed that it is available in all and furthermore it is not validated (see above). This is an important problem for statistical use because the codification of expenses is required for essential breakdowns by type of expense (e.g. whether it is consumption investment etc).

2.5.3.3 Relevance

Accrual vs. Cash basis. National accounts are computed in an accrual basis rather than cash basis. So when a payment is made it should not be assigned to the period of the date of payment but to the period of the date when the product or service was delivered. It is important to be able to establish the later. The content of «διδύμεια» provides enough information for this distinction. A payment is the last of a series of decisions and the analysis of the sequence can provide successful delineation of important events and establish when a particular income or expense was incurred. For instance a government entity that needs a product and has budget available needs to:

1. Decide to request the product
2. Implement a procurement procedure
3. Receive the product (incurred the expense)
4. Initiate the payment procedure
5. Implement the payment (Cash basis)

Software that is able to connect the decisions and establish the sequence of events is needed to correctly assign expenses or income to the accounts of the entity in a correct manner compatible with National Accounts methodology.

2.5.3.4 Timeliness

Currently, the Greek NSI (ELSTAT) is collecting quarterly and annual data from ministries. Their deadlines for data submission are:

- 60 days after the end of the reference quarter,
- 60 days after the end of the reference year for preliminary annual data, and
- nine months after the end of the reference year for final annual data.

The collection of data from «διδύμεια» has the potential to generate government finance data much faster, with the ability to have most of the data at the end of the reference period.

2.5.4 Financial transparency portal data: conditions for opening them to producers of official statistics

Data from «διδύμεια» is not only government owned and thus generally available for official statistics but in fact it is available to anyone. All data is available under a Creative Commons - Attribution license.³⁹

This means that it can be captured by software in an automatic fashion thus minimising the substantial burden to the general government entities involved. Currently, personnel (at least two persons) in each ministry are assigned the role of statistical correspondent and many more are involved in the production of primary data in each entity. This burden can be reduced substantially if part of all of the reporting can be done automatically.

2.5.5 Financial transparency portal data: conclusions

A huge amount of data on public expenditure is available through the financial transparency portal «διδύμεια». Main conclusions from analysis of its content and availability include:

- Data can be retrieved and processed for statistical purposes as it is publicly available and contains fields that can be linked to statistical classifications.
- There are several issues affecting data quality, primarily having to do with data entry errors and shortcomings in the current software that was prepared as a pilot. Most of them are expected to be solved with a new version currently under development that is expected to be released on September 2014.
- There are important impediments in terms of coverage; only expenses that require decisions are included. Therefore the source can't become a single source for all government finance data but it can be used as a supplementary source and in that way to:
 - Reduce the burden to public administration entities by requiring them to report to the NSI only data that has not being published in «διδύμεια»
 - Substantially improve timeliness.
- «διδύμεια» can serve as a primary source for statistics in certain areas where coverage is complete or near complete (e.g. public procurement, R&D spending).

3 General conclusions

The volumes and variety of data being generated nowadays mean that the five use cases presented in this report are a very small, purposefully selected sample of potential data sources. Nevertheless, even this sample represents a wide palette of data providers and potential applications in official statistics, summarised in Table 1.

Table 1. Overview of the characteristics of the big data sources examined in this report.

Source	Potential statistical domains	Data owner	Type of source	UNECE classification ⁽¹⁾	Degree of openness	Structured?
AIS	Transport	Small private	Crowd-sourced	3122 - cars*	No	Yes

³⁹ <http://creativecommons.org/licenses/by/3.0/>

Source	Potential statistical domains	Data owner	Type of source	UNECE classification ⁽¹⁾	Degree of openness	Structured?
	Environment	enterprise	data		confidentiality constraints Bulk data available at a fee	
Real estate classified ads	Housing price statistics	Small private enterprise	Classified advertisements	--*	Bulk data availability not clear	Partly
Social media	Public sentiment Well-being	Large private enterprise	Social network posts	1100 - social networks	A subset of the data is open Bulk data mainly available at a fee	No
VISA Europe	Consumer expenditure	Large private enterprise	Business transaction data	2240 - credit cards	No release of the data to third parties	Yes
Diavgia	Government expenditure	Government	Government data	--*	Open data	Yes

(1) Categorization of the source according to the draft classification of types of big data, prepared by UNECE's task team on big data⁴⁰.

* Exactly fitting class not available in the classification.

The cases demonstrate that there are big data relevant, at the outset, to various existing official statistics as well as data that can produce new statistics (e.g. AIS data and emission statistics or social networks and well-being statistics). Ever more personal and professional activities are carried out online or have online counterparts and leave 'digital footprints' behind. The chances of finding data relevant to a given statistical domain therefore increase and should not be overlooked by NSIs.

Big data offer **several potential benefits** to the production of official statistics. Their sheer volume makes them similar to very large 'samples'. They carry information about a very large number of statistical units and therefore provide potential for statistics about very detailed sub-groups of the studied populations. The real estate classified ads for example provide data about a far larger number of dwellings than what any sample survey could offer. Moreover, the data offer the possibility of producing statistics at a very fine geographical resolution, as shown again by the real estate ads or by the AIS data.

The second main characteristic of big data, the very high speed of updating or accumulation of data, means that statistics of very high timeliness and frequency can be produced. This is a very useful property for the study of volatile phenomena (e.g. consumer confidence) or, more generally, for the

⁴⁰ <http://www1.unece.org/stat/platform/display/msis/Classification+of+Types+of+Big+Data>.

production of flash estimates of key indicators. Even when not of ‘flash estimate’ speed, statistics based on big data can supplement official statistics of very low frequency (e.g. VISA transaction-based statistics versus household budget survey-based ones).

Big data that are generated without the intervention of human reporting (e.g. AIS messages or VISA transaction data) reduce the burden imposed on individuals and enterprises for statistical data reporting, one of the major considerations of every NSI. Moreover, they lead to more accurate reporting of information. Recall errors or intentional retention of confidential information (e.g. the purchase of goods or services that an individual may find undesirable to report) are avoided to a large extent. Finally the data collection costs of NSIs may be reduced since they avoid the need for sample surveys (see also arguments to the contrary below).

Finally, if a big data source has geographical coverage greater than a single country (e.g. the AIS messages have global coverage) this means that geographical comparability will be higher than that of survey-based or administrative data for the same countries.

On the other hand there are **potential disadvantages** too. The big data sources may be applying different concepts than those required by the corresponding official statistics. For example, the real estate classified ads contain data on asking price but not on the final price at which each property is sold or let.

The coverage of the intended target population may not be the desired one. For example AIS messages cover vessels larger than 300GT and only a voluntary subset of the smaller ones; expenditure data in Diavgia omit some sensitive expenditure items. Therefore, either the target population of the statistics must be modified or the big data need adjustment or combination with additional sources.

The need to combine several big data sources also emerges because a single source may not contain all required variables. For example AIS message data must be combined with technical data available from separate sources in order to estimate emissions. This means that NSIs face the need to link data sources. Data linking is not a new issue but it may be something that has not been confronted by all NSIs.

Additional data processing needs, which may not appear in a well-designed survey, emerge in the case of big data. On one hand they are needs for data validation and cleaning, as the example of Diavgia shows: this dataset may contain double or multiple-counting of the same expenditure item, may report amounts of money in text format, making all types of spelling mistakes possible, and may contain no expenditure type codes or wrong codes. The large amount of data increases further the processing needs for validation.

Moreover, processing is needed in order to convert data to useful quantitative data. Tweets for example must be analysed with the help of a thesaurus and perhaps semantic analysis of their content so as to be transformed into scores of positive or negative sentiment. In fact Statistics Netherlands⁴¹ receives processed statistics generated from social media messages by a private company. Statistical modelling may also be needed to convert data into measurements of variables (e.g. ship draught into weight of cargo) that can be aggregated for the production of statistics.

Some big data sets represent self-selected samples. For example, not all individuals have Facebook accounts, arguably choose what they want to post on Facebook and moreover probably make public only a subset of it. Therefore, regular statistical inference may not be correct without modifications, which is a research topic at present.

⁴¹ See deliverable D2 of the present project.

Finally, there may be impediments to the NSIs' access to the data. Some of them may be confidential (e.g. credit card transactions) and heavily 'guarded' by the source owners. Access to them may be very difficult or impossible. Others may only be available via private intermediaries (e.g. Facebook status updates) who will charge for access.

Cost of access to the data combined with cost for processing them may in fact offset the gains from not having to run a sample survey.

The examined cases show that **each statistical domain and each possible big data source is a unique case**. Each one represents different possible benefits and different difficulties for NSIs pondering its use. It would be imprudent for NSIs to ignore big data but they should not embrace them uncritically either. Each potential source must be examined carefully versus statistical needs and the other sources with which it could be combined. It seems that at least a subset of the currently produced official statistics can be supplemented by statistics based on big data, while new indicators can also be produced.

12.7. D5 – Accreditation procedure for statistical data from non-official sources

European Commission – Eurostat/G6

Contract No. 50721.2013.002-2013.169

‘Analysis of methodologies for using the Internet for the collection of information society and other statistics’

D5: Accreditation procedure for statistical data from non-official sources

February 2014

Document Service Data

Type of Document	Deliverable		
Reference:	D5: accreditation procedure for statistical data from non-official sources		
Version:	3	Status:	Draft
Created by:	Michalis Petrakos, George Sciadas, Photis Stavropoulos	Date:	6/2/2014
Distribution:	European Commission – Eurostat/G6, Agilis S.A.		
Contract Full Title:	Analysis of methodologies for using the Internet for the collection of information society and other statistics		
Service contract number:	50721.2013.002-2013.169		

Document Change Record

Version	Date	Change
1	27/11/2013	Initial release
2	31/12/2013	Revision following comments received at the progress meeting of 3/12/2013
3	6/2/2014	Revision following comments received on 16/1/2014

Contact Information

Agilis S.A.
 Statistics and Informatics
 Acadimias 98 - 100 – Athens - 106 77 GR
 Tel.: +30 2111003310-19
 Fax: +30 2111003315
 Email: contact@agilis-sa.gr
 Web: www.agilis-sa.gr

TABLE OF CONTENTS

1. Introduction	3
2. The quality environment	4
2.1. Quality Approaches for Administrative Data	30
3. Accreditation.....	9
3.1. Conceptual underpinnings of the proposed approach	9
3.2. Foundational principles.....	10
3.3. Refinements and Interdependencies	122
3.4. Procedure for Accreditation	16
4. Casting the net wider	25
4.1. Certification.....	266
5. Summary and conclusions	27
References.....	29
Annex 1 Overarching quality frameworks.....	30
Annex 2 Examples of secondary sources	33

1. Introduction

The ongoing barrage of creative thinking concerning the potential of all kinds of new data from a variety of sources entering the official statistical system, and specifically the world of the National Statistical Institutes (NSIs), has also brought to the fore the need for some sort of accreditation. Naturally, this is closely related to the issue of quality, traditionally a hallmark of NSIs.

The same is echoed by the work undertaken in this project, which investigates the potential of new sources and methods for the compilation of ICT statistics and beyond. Examining issues related to quality and eventual accreditation procedures for secondary data sources is therefore a worthy issue.

What follows is a think-piece that addresses directly the issue of quality, proposes an accreditation procedure that producers of official statistics can use to assess the quality of data from non-official sources, and discusses broader interrelated matters that will certainly be faced in the near future. The utilisation of new **data sources** should be differentiated from exploiting digital footprints, such as scraping web sites of enterprises or gaining access to individuals' smartphones as explored in this project. The latter are in reality new **collection methods**. The accreditation procedure is not meant for this purpose but the intent is to:

- i) Situate the issue of accreditation within the existing and overarching environment that guides the statistical system, and step on well-established frameworks and procedures for quality that historically constitute one of the system's key strengths.
- ii) Use this as a springboard to expand the examination of quality specific to secondary data sources by taking stock of the current state of affairs and linking to the body of knowledge that is already available.
- iii) Contribute some new observations, analytical commentary and, hopefully, insights that can help advance decision-making in light of the reality facing us today.
- iv) Combine all the above, with appropriate adaptations, in a way that:
 - enables the articulation of useful specific steps and procedures for accreditation
 - puts on the table a set of issues to stimulate further dialogue and exchange.

At the end, a range of forward-looking issues is discussed. For the most part, these relate to the “big picture” that will drive developments in the area of new data and data sources for some time to come.

2. The quality environment

This section presents a synoptic overview of the notion of quality that governs the work of the official statistical system, as well as its application in the production and dissemination of statistical outputs at the NSI level. These quality principles and their resulting practices are ultimately responsible for the reputation enjoyed by organisations within the official statistical system, an attribute paramount to the credibility of the whole effort. The intent is not to be exhaustive but rather to ensure that this “background” is carried on in our minds and serves as a reference against which to judge our comprehension of what is involved in much more detailed issues as we move on.

Over many decades NSIs have been supplying the bulk of data needed to understand the state and evolution of our economies and societies. These statistics got thoroughly integrated into the fabric of countless decisions, by government in policies, businesses in decision-making, and researches in illuminating issues of interest. Moreover, they have been used constantly by the general public in their many capacities, from students at schools, to readers of current affairs or books, to voters, and to ordinary conversations among informed citizens.

Whether from well-established regular and ongoing programs with significant history, like censuses, the National Accounts, the CPI and the Labour Force or emerging issues that require quantification, such as the information society, one key characteristic of all has been their acceptability for common and wide use. It would be painful to sort through arguments, say, in contract negotiations if the credibility of the CPI was at stake or if competing CPIs were at the table. (A present-day example of the consequences that occur if something goes wrong comes from Argentina).

Such gains did not come to pass because NSIs were granted monopoly rights over these statistics. Rather they represent the outcomes of the creation, implementation and adherence to important frameworks and thoughtful principles that safeguard the overall quality of outputs for all to see. These have led to quality standards employed by statistical programs in their operations, and which have been developed and evolved over time. Moreover, they are based on orthodox, state-of-the-art statistical theory developed by professional statisticians, methodologists and other practitioners. Virtually every NSI adheres to a set of quality attributes. Quality standards for data and metadata, together with transparent methods and processes, and accompanied with limitations and caveats, are standard fare and readily available and communicated by NSIs to any and all users. The procedures in place guide the

conceptualization, design, collection, processing, analysis and dissemination of data, and they are crucially linked to their widespread acceptance and use.

Relevant material with regards to quality frameworks in Europe and internationally can be found in Annex 1.

2.1. Quality Approaches for Administrative Data

For a long time, though, NSIs also utilise other data sets that do not originate in surveys. These administrative data are of a different nature and typically they have been integrated into statistical programs – either as replacements of survey content or independently. Such data sources are now being increasingly sought after and are expected to proliferate. Moreover, as discussed earlier, they may well be augmented by numerous other sources, including the Internet.

Truth is that, with some specific exceptions of administrative data particularly in countries with advanced registers, the knowledge and understanding of procedures for quality assurance are not at the level of advancement or sophistication of survey data. Tackling this issue becomes therefore a timely endeavour as we seem to be at an historic junction when new data will be coming principally from non-survey, overwhelmingly digital, sources. Again, coming closer to our theme, we must make good use of what exists and there has been literature to match the expanding use of administrative and other sources into the official statistical system.

A good example comes from the paper “Quality Assessment of Administrative Data for Statistical Purposes”, Eurostat (2003). While the paper observes quite correctly that *“It must be kept in mind that frequently it is very difficult for a SI to assess fully the quality of administrative data. For example, the SI will not be able to assess the measurement errors in an administrative dataset if the producing organization has not studied these errors itself and does not permit the SI access to micro data either”* (p.4), it then proceeds to make a point which we consider crucial in the eventual development of any quality-based accreditation procedure for any secondary data. That is, *“...the uses and operations create the requirements which in turn define what is considered as good quality of administrative data”*. This effective definition of data quality as not absolute but highly conditional on the intended use/s of the data has been restated since then in various ways, some even more specific, but it continues to be the cornerstone insight for our thinking.

For example, more recently it has been stated that *“Broadly defined, data quality means ‘fitness for use’. Different users of the same data can have different assessments of its quality. Administrative data were gathered for a particular purpose – running a program – and can have qualities that are well-suited for that purpose. When the data are adapted to a new purpose, issues of data quality become especially salient”* (Iwig et al 2013, p. 2). This is not to say that “quality is in the eye of the beholder” without adding that “most beholders see alike”! It is not a case of perceived vs. objective quality but rather the more classic “the right tool for the job at

hand”. (In any event, in our case not only quality must really be present but must also be perceived to be present. Both are indispensable for the statistical system or any other accredited organisation).

An additional key insight of the 2003 paper was that administrative data have multiple uses, something that must somehow be factored in any approach to quality. *“Data from the same administrative source may be used in different ways, in more than one statistical products of a SI. For example, they may provide raw data for a product and may be used as a sampling frame for another one. On the other hand, one statistical product may use administrative data from more than one sources”* (p. 5).

Secondary data may be used for survey design, survey planning, data collection, enhancement of a survey’s coverage, data verification, auxiliary data collection (use in weighting and estimation), data edits and imputation, the creation of statistical registers etc.

In the same vein, Statistics Canada states that *“Statistical uses of administrative records include: (i) use for survey frames, directly as the frame or to supplement/update an existing frame, (ii) replacement of data collection (e.g. use of taxation data for small businesses in lieu of seeking survey data for them), (iii) use in editing and imputation, (iv) direct tabulation, (v) indirect use in estimation (e.g. as auxiliary information in calibration estimation, benchmarking or calendarisation), and (vi) survey evaluation, including data confrontation (e.g. comparison of survey estimates with estimates from a related administrative program)”* (<http://www.statcan.gc.ca/pub/12-539-x/2009001/administrative-administratives-eng.htm>).

Eurostat’s “Handbook on Data Quality Assessment Methods and Tools” (2007) also contains analogous references

(<http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf>).

With all that ammunition, we feel well equipped to take on to the next level that fitness-for-use is the crucial point here and that it is not an abstract notion but can be used to establish quantifiable criteria in an accreditation procedure.

This is quite consistent with the distinction between data source and data outputs, each of which has peculiarities to consider. The message is that while the quality of a data source can be assessed through a set of indicators, the assessment of the data in the production of outputs must be subject to criteria above and beyond those of the source. This is so because this is where the fitness for use criterion is applicable. In our report, we shall make such a distinction explicit, and outputs will be subjected to much more detailed accreditation steps.

Additional insights have been generated by research specifically aimed at assessing the quality of administrative sources. Daas et al. (2009) in “Checklist for the Quality evaluation of

Administrative Data sources” proposed a quality framework and a checklist. While it is well understood that quality is a multi-dimensional continuum and not a dichotomous affair, they distinguish three hyperdimensions, each of which influences the usability of the data in different ways. These are: Source, Metadata and Data and they suggest that all three should be used to determine the quality, and therefore the usability, of the data source. In this conceptualisation, each hyperdimension consists of many dimensions. Dimensions offered are:

Source: supplier, relevance, privacy and security, delivery and procedures

Metadata: clarity, comparability, unique keys, data treatment

Data: technical checks, over coverage, under coverage, linkability, unit non response, item non response, measurement, processing, precision, sensitivity

Then, one or more quality indicators are proposed for each dimension, and methods are suggested that can be used to measure each indicator. The end result is expected to come from a combination of all the above.

A different approach also comes from Statistics Netherlands, this time under the name Object Oriented Quality Management (OQM). In “A New Model for Quality Management” Nederpelt (2010) refers to an **object** as anything about whose quality we care, e.g. data of an administrative data source, metadata of an administrative data source etc., while a **characteristic** of interest can include virtually any attribute of quality (such as relevance, reliability, accessibility etc.) The object and the characteristic constitute a **quality area**. Any number of quality areas can then be assessed as the organisation cares to exercise control over them (and experts in different domains can undertake such tasks).

An additional exercise attempted to reconcile the two approaches outlined above (“Application of the object oriented quality management model to secondary data sources”, Daas and Nederpelt, 2010). The conclusion was that *“The comparison made between the two methods, reveals that a combined approach seems the most fruitful way to assure the coverage of all quality areas for a particular object”* (p. 17). This is so because the top-down approach of the OQM model misses indicators that are part of quality areas that had not been identified while the bottom-up approach of the QADS framework misses quality aspects belonging to new areas. Under that approach, tests of secondary data sources have included not only administrative data but also data from surveys by others, registers, the Internet, and offline routing information.

In a subsequent paper the two authors recommend *“49 factors that influence the quality of secondary data sources”* (Nederpelt and Daas, 2012). These areas were clustered in five categories: respondent, system, data supplier, statistical agency and regulations, agreements and cooperation.

An additional view of how to assess the usability of an administrative data source from a statistical point of view was proposed by Laitila, Walgren & Walgren (2011) in “Quality

Assessment of Administrative Data”. The authors make a case for the systematic analysis of an administrative source, and differentiate between *producer and consumer views* of the administrative data. *“The consumer view concerns the quality of the final product, or the ‘Output quality’. The producer view concerns two problems: i) ‘Input data quality’ – the preparations of the input needed for use in the production process and, ii) ‘Production process quality’ – the gains in the production efficiency of using the input”* (pp. 9-10). In that setting, the quality assessment of the secondary source must be done for each of the three components: output, input data, and production process. Outputs are then assessed by means of indicators for each sub-component of each quality component (relevance, accuracy etc.). Similar procedures are followed for the other two quality concepts. One of the results is that *“The strongest requirements on an administrative register are found when it would be used as the single source for producing statistics”* (p. 12). Considering the actual operations of NSIs, these insights too are valuable and they will be exploited in our approach.

A recent contribution comes from the USA by Iwig et al. (2013) in “Data Quality Assessment Tool for Administrative Data”. This paper re-iterates the importance of the fitness for use criterion, as effectively synonymous to data quality. Moreover, it asserts that quality assessment can benefit both the NSI and the (secondary) program area and proceeds to develop a data quality assessment tool. Identifying the information/knowledge asymmetry between source and user it states: *“This Tool is developed to support a conversation between a user of an agency’s administrative data—either a user who may be initially unfamiliar with the structure, content and meaning of the records or a user who repetitively acquires the data but may not be aware of recent changes to the system—and a knowledgeable supplier or provider of administrative data. The Tool provides questions that are pertinent to helping a user assess the fitness for their intended use”* (p. 2). We shall make maximum use of these insights too.

Very much like what we have encountered earlier, it is recognised in the paper that quality has many dimensions, leading to the tool having six already-familiar dimensions: relevance, accessibility, coherence, interpretability, accuracy, and institutional environment. The tool contains 43 questions, but not all need to be answered at the same time. Instead, it is organised in three phases: discovery, initial acquisition, and repeated acquisition. Within each phase, the questions are organized by the dimensions of data quality that are relevant to that phase and thus only a subset of the questions must be answered at any one time since different activities and decisions rely on different kinds of information. The organizing principle is the signing of a legal agreement (MOU). The discovery phase contains 12 questions focusing on the dimensions of relevance, accessibility and interpretability. In the initial acquisition phase, accessibility and interpretability become central dimensions and account for 29 questions. The third and final phase (repeated acquisition) has 11 questions but only 2 are new since 9 are repeated from the previous phase. The paper also includes a detailed data dictionary template.

According to the authors, “*Using the Tool does not result in a single overall numerical measure (or metric or index) for data quality. Instead, the Tool provides questions for which some answers are quantitative and others qualitative. The usefulness of the Tool lies in providing a well-developed set of questions that prompt the user to consider certain key attributes of data quality; the Tool does not result in a judgment or recommendation apart from what the user develops. It is the user’s own interpretation of the answers—and the user’s prioritization of which ones are especially germane for the data application at hand—that constitutes the user’s own assessment of data quality*” (p.3).

The ABS follows along with general purpose questions intended to ascertain the quality of admin data. They are modeled after the 7 dimensions of the Bureau’s Data Quality Framework which also includes the ***institutional environment*** in addition to relevance, timeliness, accuracy, coherence, interpretability and accessibility. They do capture aspects of the beginning of data, series start, revisions, method of collection, under/over counts, representation of population, non-reporting items, comparability issues etc. In this case, the questions are more open-ended than most, leaving it to the respondent (administrative source) to provide lengthy answers. (ABS, Data Quality Statement Questions, Data Quality Online). https://www.nss.gov.au/dataquality/PDFs/DQO_Admin.pdf

3. Accreditation

Depending on the institutional arrangements of a country, NSIs have been using data from administrative sources for some time. In the process, many issues have been dealt with, kinks have been ironed out, and much experience has been accumulated on how to integrate such data with surveys for the production of outputs. At any rate, the fact that use of administrative and other secondary sources is expected to intensify calls for the establishment of basic accreditation procedures that will guide the acquisition, treatment and uses of such data.

3.1. Conceptual underpinnings of the proposed approach

The design of an accreditation procedure must accommodate a multitude of dimensions that vie for attention. The divisions of each such dimension can delineate several focus areas. Right from the outset, it becomes evident that the most prominent dimensions alone can delineate an extremely large number of areas of interest. Consider, for instance, the following:

- The existence of diverse secondary sources - at different degrees of advancement
- The distinction among source, metadata, and data – at least
- The ways in which secondary data can be used - auxiliary, standalone outputs etc.
- The need to consider content, administrative, and technical matters separately
- The many quality dimensions of outputs to examine (relevance, accuracy etc.)

- The need to examine both inputs and outputs
- The different “models” of quality that can be used

Even assuming a few categories for each of the above, their “intersections” can delineate an unwieldy number of individual areas, impossible to negotiate – as evidenced by their permutations, which will be in the thousands. To cut through such a spaghetti-like conundrum and identify manageable pieces of work, a good deal of pragmatism becomes a definitive asset.

With that in mind, our thinking internalises all that is fundamentally useful from the discussion so far, while remains simultaneously rooted at the actual workings of an NSI. To further underpin and solidify the approach, foundational principles are explicitly spelled out through the discussion that follows. These, then, support the proposed accreditation procedure.

3.2. Foundational principles

This issue of secondary data sources is linked to an ongoing evolution and has its time and place. Therefore it cannot be examined in isolation. There is no need to re-invent the wheel; much of what we need is already in place. New data will come into well-established statistical norms and practices and will be integrated into the whole system. Such integration in no way invalidates, or somehow renders outdated, the existing quality frameworks and the practices of adhering to existing quality standards.

Principle 1: Accreditation procedures must be fully compliant with well-established principles of quality frameworks that guide the world of official statistics, and consistent with quality assurance practices embedded deeply in the work of NSIs.

On the other hand, it is not a far stretch to say that a negative predisposition to the new could lead to a level of standards impossible to attain. At a time when NSIs are amenable to the idea of quality levels that fit the need, under the logic of the fitness-for-use criterion, we cannot overplay the quality card and raise the bar at a height where nothing can possibly jump over it. That would be akin to hiding behind some high quality morale and become insular, something detrimental to any NSI. The heterogeneity of potential sources requires, at a minimum, research that would lead to more in-depth knowledge and experience. The whole issue of new sources must be approached with an open mind and a welcoming attitude.

Principle 2: Any accreditation procedure must be flexible in a way that does not unduly prejudice or rule out new opportunities without serious examination.

At the same time it is recognised that venturing into the examination of all kinds of new sources will undoubtedly consume a fair amount of effort, energy and resources. Seen under the prism of investment, therefore, it should be leveraged prudently to maximise returns. A corollary of this

is that the accreditation procedure should contain incremental decision-making and allow early “gating” and front-loading of work rather than require a lengthy and large-scale investment, at the end of which no fruitful outcome may materialise.

Principle 3: An accreditation procedure should include sequential decision-making based on a pragmatic step-wise approach, so that we spot early on new data sources that won’t work, while we always invest in new sources that will work.

In light of the asymmetry of knowledge between owners of new data sources and NSIs, assessments may be subject to both type 1 and type 2 errors. Good sources may inadvertently be disqualified and bad ones qualified, only to find out much later at high cost. While the literature explicitly identifies the need to assess the data, some of the proposed methods stake much on answers to open-ended questions by owners. Consistent with the previous principle, we find it constructive to differentiate between aggregate data and microdata. NSIs know well that there is no adequate substitute to microdata as a building block for statistical products.

Principle 4: The accreditation procedure must contain an empirical assessment with real data, and it must be carried out by NSIs directly. It cannot be delegated to filling out questionnaires by the source owners.

As early as 2003, Eurostat noted: “Reference to specific coverage problems (over-coverage, under-coverage, misclassification, duplication) may not be possible with no specific statistical product in mind” (p. 12). It is by now well understood that new data can serve many uses. Among them, they will be used as inputs in the production of statistical outputs – whether existing or new. So, they should be assessed with regards to the impacts they have on the quality of those outputs. (In the case of a new output, rather than assessing the impact it becomes a matter of establishing the best quality possible). Yet, new data sources can also be assessed in their quality as inputs, something that is not expected to mirror the familiar quality dimensions of the outputs. As part of an accreditation procedure, we can always map which quality dimensions of the new data as inputs correspond to the key quality dimensions of the outputs. (This is explained in detail in section 3.3).

Principle 5: A systematic accreditation procedure must assess the quality of the statistical outputs, the quality of the statistical inputs (including the source and metadata), as well as the quality of the statistical processes involved.

While the fitness for use is a powerful quality criterion of a statistical output, it is more oriented towards the inherent subject-matter itself (e.g. can these data be used in a meaningful way for this type of analysis) and thus constitutes a narrower notion than the one encompassed by all quality dimensions. In addition, fitness for use does not lead to dichotomous outcomes but to trade-offs concerning acceptable levels of quality vis-à-vis intended use. Based on this criterion, the same statistical output may be produced with different quality levels. Moreover, quality

cannot be compared across different outputs. This is so because the desired quality dimensions do not lead to absolute quality measures as such. For example, the timeliness of the CPI (say, 3 weeks after the reference period) cannot be considered superior to that of the GDP (say, 6 weeks after the end of the quarter).

We must therefore come to terms that statistical outputs can be of different relative significance for NSIs. While every output matters to influential groups of users, some are more critical than others. This becomes evident during contingency planning for business continuity due to distractive events (e.g. interviewer strikes). Depending on which output will rely on the new data source, outside the direct control of the NSI, brings into the decision-making the issue of risk management (which we do not believe can be subsumed under quality indicators). Accreditation must explicitly account for that, and the process should provide all necessary information, including measurements of the vulnerability of critical outputs.

Principle 6: The final decision for the accreditation of a new data source must incorporate a combination of corporate criteria, broader than strict data quality. The accreditation procedure must compile adequate supporting documentation, including measurements.

3.3. Refinements and Interdependencies

There is agreement in the literature than in case of administrative and other secondary data, not only the data and the metadata must be assessed but the source too. Eurostat (2003) identified: *“We believe that two different types of internal quality reports on administrative data are needed. One type will refer to particular administrative data sources (source specific) and the other will refer to particular statistical products (product specific)”* (p.5).

The source is frequently referred to as institutional environment. We choose to refine the “source” by decomposing it explicitly in two parts: first, as it relates to content and subject-matter involved in the data, and; second, as an institution with regards to the power to negotiate, conclude and sign legal or binding contractual agreements (e.g. MOU). (Roughly speaking, the distinction can be thought of as that between the working-level subject-matter unit and its work objective, and the senior level of the organisation’s management).

As well, we find it useful to decompose the hyper-dimension of data to explicitly account for aggregate data and microdata, as they practically have different implications – both in their assessment and their use. Thus, we utilize five hyper-dimensions:

- source – content, metadata, aggregate data, microdata, source – institution.

Furthermore, elaborating on Principle 5, to avoid some unnecessary confusion in the literature and to facilitate understanding, we define input qualities that do not overlap with the terms used in the quality dimensions of outputs. These are: **potential usefulness, usability and cooperation**. Each of them affects different parts of the needed assessment, and at different times. Assessing potential usefulness can take place early, usability requires much more effort, while cooperation extends beyond strict quality. How they all map together is explained below.

Some ambiguities arise in quality assessments from the desired quality attributes of outputs vs. those of inputs to statistical outputs. The quality of a statistical output released by an NSI (say the CPI) adheres to the known quality dimension : relevance, accuracy etc. Clearly, this is a relative rather than absolute measure of quality. It means that the product was tested against each of these qualities and passed a specific threshold, and that each of these qualities of the product is as good as it can be. As every other released product in the mix has undergone the same procedure (in the pass/fail sense), the qualities of different outputs are not directly comparable. For example, a product (e.g. quarterly GDP) may have “worse” timeliness than the CPI, in the sense that it is released 6 weeks after the end of the reference quarter but the CPI is released only 3 weeks after the reference month. These are clearly not comparable. Since, then, quality attributes are specific to a product whether or not they improve or deteriorate over time will be judged against the product itself and not others. This becomes material in our accreditation procedure since new input data will impact on output quality.

To illustrate, insights can come from production processes outside the statistical world. Assume that as part of being committed to TQM, a car manufacturer defines, measures and advertises the following quality dimensions of a car: Functionality (handles well, reliable), Performance (acceleration, speed etc.), Fuel-efficiency (kms per litre), Good looks (aesthetically pleasing exterior and interior). To produce the car (output) and achieve these quality attributes, the manufacturer uses all kinds of inputs. Presumably, he has a keen interest in the quality of such inputs. However, these inputs are the outputs of other businesses (parts manufacturers) and their quality attributes have been determined by their makers. Although the two sets may overlap, they should not be expected to be the same. For example, the engine manufacturer may also have fuel efficiency as a quality attribute but not good looks. On the other hand, the manufacturer of leather seats may include good looks as a quality but not fuel efficiency.

If the car manufacturer is looking at changing engines and seats, each of them will affect differently each of the quality criteria for his car. Although at some point everything affects everything (e.g. not inconceivable that lighter seats may improve fuel-efficiency), the effect of a new engine is expected to affect primarily the fuel efficiency indicator and the new seats the good looks indicator.

Such matters can become more complicated, the closer the inputs and the outputs are in nature. That is, if statistical outputs are used to produce other statistical outputs the qualities of inputs and outputs will be much more alike. This will represent well the case of Eurostat vis-à-vis NSIs. When we look at new secondary sources, though, they may or may not relate to statistical outputs (such matters are discussed more in Section 4). Sensitivity analysis will be helpful to identify such issues, particularly when the contemplated new data sources are of the Big Data variety. Unlike other secondary sources, such as administrative data, Big Data tend to be unstructured or their structure is largely unknown and difficult to decipher. A specific problem that may well occur with Big Data is imprecision, that is, the presence of rather qualitative or categorical points instead of numerical values in data sets. Then, the use of methods, techniques and tools that would unveil possibly hidden structures in a meaningful and usable way becomes an arduous but necessary task. This is true of data mining and/or data visualisation techniques and the like, which assume additional important if adequate metadata do not exist.

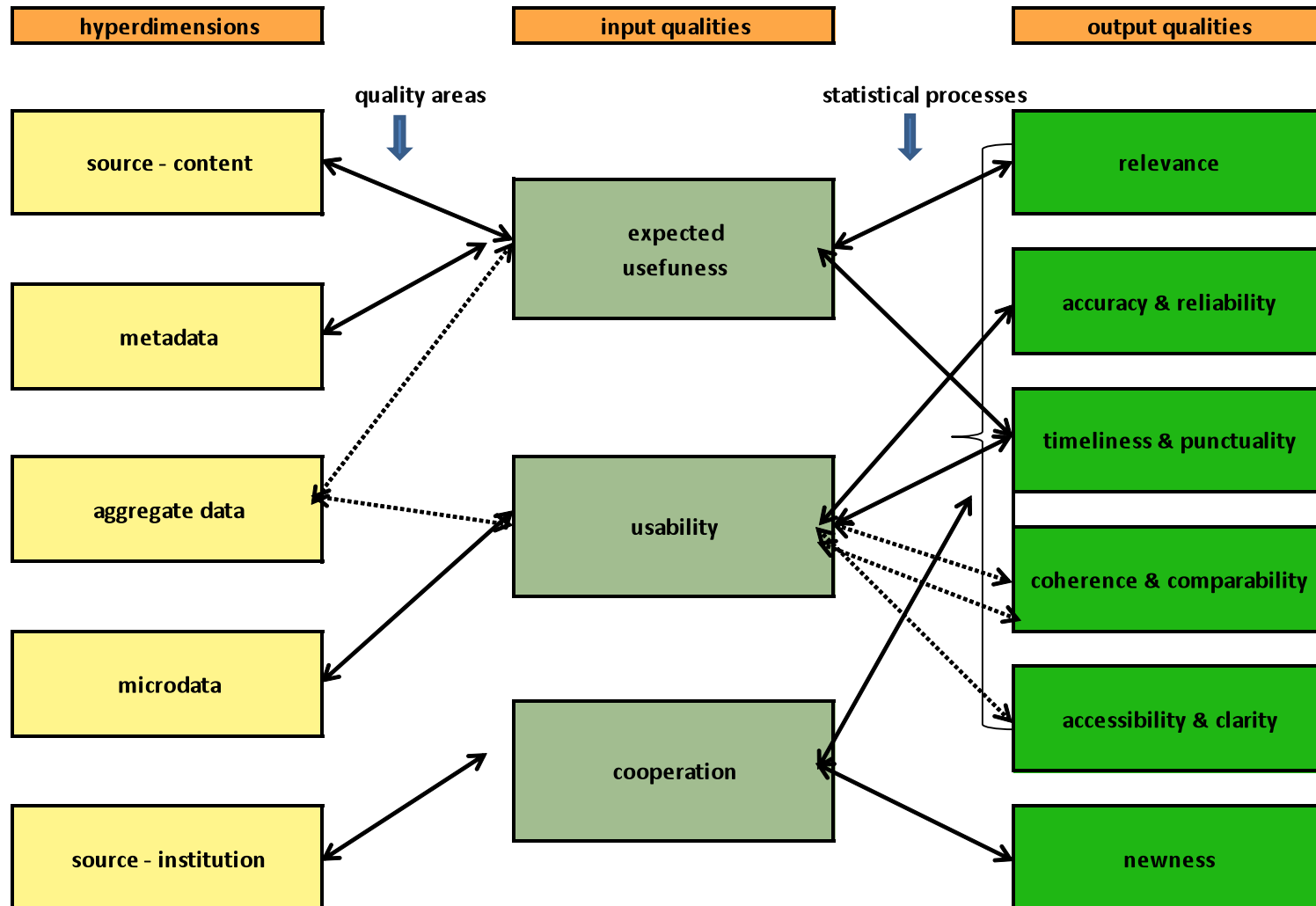
In our case, the above is particularly important if data from the new source are used as inputs in existing outputs. As we have already seen, they can be used to produce brand new outputs too. For instance, *The Role of Big Data in the Modernisation of Statistical Production* (Fiona Willis-Nunez, 2013) identifies the following potential uses in official statistics: i) experimental uses, ii) complementing existing statistics, iii) supplementing existing statistics, iv) replacing existing sources and methods.

In the latter case we should not be looking at impacts on existing quality but establish the best quality possible for the new output. In any event, even in such a case, we don't believe that new data can be used directly without going through the statistical production process – if not combined with other data, they will surely involve methodology etc. We use the term **standalone output** to denote this.

The following schema (Figure 1) describes the primary and secondary linkages between hyper-dimensions and quality characteristics of the secondary source data (bold and dotted connectors, respectively). Then, these are mapped in a similar manner to the quality dimensions of the end outputs. As explained, in the case of existing products which may use the new inputs what matters is the marginal change in each of those quality dimensions.

Consistent with the OQM, we can effectively define quality areas with hyper-dimensions as the objects and the qualities of the input as the characteristics. Moreover, from a global perspective we can view all known output qualities as one, and add **newness** as an extra quality for a new standalone output.

Figure 1. Hyper-dimensions and quality characteristics of secondary source data and of the statistical outputs to which they contribute.



3.4. Procedure for Accreditation

Consistent with the preceding analysis and the principles developed, the proposed accreditation procedure evolves in a step-wise fashion. It consists of five stages with gradual assessments involving indicators measured through scales and hard data, which in turn lead to recommendations associated with six decision points. This section is accompanied by tables, which also contain illustrative examples. At the end of the section additional comments are offered, together with Figure 2, a flowchart of the procedure.

Stage 1: Initial examination of source, data and metadata

In order for an NSI to even contemplate acquiring and using an external data source some knowledge of it, or at least exposure to it, is surely a necessary condition. That is, some individuals have become aware of that source at some level, have a decent idea of what statistics it might produce, or perhaps have come across published outputs or third-party references in a way that picked their curiosity.

At this stage, an early assessment of the data, the metadata and the source is needed. Anything that can be gauged from the outside or through limited and rather unofficial interaction with the working level at the source organisation should be collected, shared internally, and examined. Such material can come from the media, Web sites, releases, publications or articles and should cover the *raison d'être* of the organisation behind the source and as many aspects of content, data and metadata as possible.

Here the emphasis is placed squarely on the potential usefulness of the data. There should be no concern with the feasibility of actually acquiring the data, and much more doing so routinely, timely or under what terms and conditions. Similarly, the quality of eventual outputs should not enter the picture, not even the quality of the data themselves yet.

The overarching question (same as the quality of the input) is: ***potential usefulness***. Detailed questions can examine the population coverage, units of measurement, variables, timeliness, frequency, as well as provide some information on the organisation. They should also include possible uses of the data that will help the decision at this stage. The emphasis on potential usefulness has the practical implication that at this stage we do not need hard data to decide.

At the end of assessing various indicators with a scale ('high-medium-low' as in Table 1, or different) a Yes/No answer is needed to the question: "Is this data source potentially useful and for what"? This will lead to a recommendation to proceed to the next step or not, which constitutes an early decision point.

Table 1. Fictional example of implementation of stage 1 of the proposed accreditation procedure.

Accreditation procedure for secondary data Stage 1: Initial Examination - Potential Usefulness				
questions	description	possible outputs	score	comments
Theme	A-social networking/B-retail superstore		L, M, H	web site/many outlets
Source-content		social networking		
organisation type	private/private	new output	M	new area of data, networks
area of inquiry	social media network/sales transactions	input to Internet use	M	
scope	anyone who registers/all retail stores	input to time use	L	
size of effort	about 1/2 million users/1.5 million per month	data confrontation	M	
method of collection	online registrations/cash register software	other	M	
since when	5 years/more than 20 years	overall assessment	M	variety of social stats possible chatter data
releases	some aggregate data/none, proprietary			
prospects	unknown/very stable			
other	-			
Metadata		retail superstore		
population	census/sample	replacement for survey	H	only for this retailer only for this retailer various commodities daily sales, if all retailers
units	individuals/transactions	input to monthly retail	H	
variables	numerous/prices, total bill	input to CPI	L	
reference period	any point in time/day	data confrontation	M	
timeliness	immediately/next day	other	L	
frequency	stock at some time/probably even by hour	overall assessment	M	
classifications	none/own product codes			
dissemination method	online/none			
relation to other data	online time/retail trade, prices			
paradata	unknown			
Aggregate data	users, time online/total sales, daily			
by type	age & gender/by product, method of payment			
by city	no/yes			
other	perhaps traffic flows/-			
Microdata				
identifiers	name & address/own product codes			needs verification
linkage potential	remote/none			needs verification
Additional comments: Retailer accounts for 15% of monthly sales in its NACE				
DECISION POINT 1	Recommendation:	A-Proceed	B-Proceed	

Stage 2: Acquisition of data and assessment

This stage entails negotiations with the source with a view to acquire a set of files or file extractions adequate for rigorous testing. The hands-on testing itself will be the main object of this stage.

The primary objective is to clarify whether the source is willing and able to deliver files or extractions at the record level, as well as keep open a communication channel during the testing process. Without the cooperation of the source data cannot be obtained and no real progress can be made. A number of issues must be discussed in a professional manner with the data source, albeit not with the burden of formalizing a legal agreement yet (e.g. MOU) - which is more demanding. Certain details pertaining to what, how, and when will be delivered will be prominent among them. These include specifications of files or file extractions, time and method of transmission, as many metadata as possible, and any particular conditions that must be known. In the process, we can update the results of Stage 1 with more accurate information that becomes available. This is not a repetition of Stage 1. It adds the revised results of that stage to those of stage 2.

As a guide, the target should be to obtain enough data and metadata from the source in a way that these would be comparable to the amount of data and information typically available to NSIs immediately after collection in a survey process. There is no reason to put the bar higher. At that point, we have ample information in the form of questionnaires, glossaries, interviewer guides, as well as a collection file with coded data. The effort should be directed to acquire the same, and a close concordance with the source material can be developed. For instance, the questionnaire corresponds to their input form, the glossary to instructions provided to individuals who must fill and submit the form or register, and the like.

It is understood that at this point the files have missing and incomplete data, item non-response, outliers and many other issues that will be part of a later clean-up phase. The point remains that despite such issues in the file immediately after collection, none of them is a showstopper. We recommend the same for the acquisition of microdata from secondary sources – match as closely as possible this situation with which we are quite familiar. In exchange, perhaps, the NSI can commit to share some of the intelligence that will be gleaned during testing, something that typically is of interest to data sources.

Again, through the systematic capture of information and scoring across all items we can ascertain whether we can arrive close to a post-collection survey situation and make an appropriate recommendation for this decision point.

Table 2. Fictional example of implementation of stage 2 of the proposed accreditation procedure.

Accreditation procedure for secondary data					
Stage 2: Acquisition of Data - Potential Usefulness					
questions	Y/N		additional information	score	comments
	A	B			
Source-content					
<i>willingness/ability to share</i>	N	Y	most data holdings	H	source is cooperative
provide data for testing purposes	N	Y	need to specify extractions	M	need some time but no other issue
resources/dedicated contact	N	Y	will nominate an analyst	H	asked that we also have one direct contact
time to communicate	N	Y	pre- and post-delivery	H	as required
Metadata					
input forms/reg	Y		layout of system receipts	H	
definitions of variables	Y		item descriptions	M	
filer instructions	N/A			-	
mandatory fields	N/A			-	
coding	Y		own	L	extremely detailed coding, e.g 6 types of apples
info material	N			-	
size of files	N/A			-	we need to specify extractions, by day, week, month
% of target	N/A			-	every transaction is captured
transcription/capture	Y		electronic	H	
issues encountered	N/A			-	
Aggregate data	Y			M	
existing internal reports	Y		limited	M	mostly for company management, not for sharing
custom	Y		we can design some, not complex	M	they can use our specs, useful to compare later
Microdata				H	
file extract specs	Y		format and time period/s	H	
time period	Y		they have detailed, even by hour	H	they can meet our time
platform, file format	Y		they can export to SAS EG	H	
file size	-		no issue, within reason	H	
record layout	Y			M	they prefer we stay with main variables
method of transmission	Y		secure e-channel	H	
Additional comments: No money, no other conditions, just that we need to be reasonable. If test works, they will ask to be removed from surveys					
DECISION POINT 2			Recommendation: Proceed		

Stage 3: Forensic investigation

This represents a critical step and requires a fair amount of work by the NSI. It can be sub-divided in four distinct phases: i) producing a clean microdata file (halfway through which we meet a decision point); ii) using the file to produce and analyse aggregate statistics iii) producing pilot new outputs or using the file in the production of existing outputs, and; iv) assessing the capacity of the existing statistical tools to handle the new data.

a) During the first phase of this stage, all the known steps taken for the processing of collection files apply. Everything must be scrutinized and verified. Duplicate records will be identified and removed, specifications for various kinds of edits will be developed (flow, validity and consistency edits) in a way that will correct erroneous, inconsistent or contradictory entries, outliers will be detected and dealt with, and documentation will be kept. A number of quantitative indicators can be constructed during this stage that will speak volumes for the quality of the files (included in the table). It is conceivable that before the end of this stage, in particularly under circumstances where the file/s are deemed to be in a really bad shape (e.g. effective response rate too low for the production of any meaningful aggregates), a judgment may be made to recommend that we should not proceed further. This is similar to having an exceptionally low response rate that proceeding to estimation is unacceptable.

If we proceed, we perform weighting (if applicable), imputation and make any other adjustments necessary to arrive at a final microdata file, which will be used for estimation. All along, we continue to document through quantitative indicators.

b) In this phase, we use the clean microdata file to produce actual aggregate statistics, which are then analysed and compared with any existing data, such as prior publications by the source, or confront their levels and movements against related series. If the quality of the resulting aggregates is deemed satisfactory, as captured by additional qualitative indicators, we proceed to the last step of this stage.

c) This entails the use of the microdata in the production of actual statistical products, which can cover one or more of the initially intended uses, and can be a standalone output or parts of one or more existing outputs. This must be accompanied by detailed analyses of the impacts of using the new data on the quality of existing outputs. Generally, they should be accompanied by a good identification of pros and cons, which will serve as additional indicators in the assessment. For example, an output may gain in timeliness but lose in accuracy.

d) Moreover, during this stage it would be opportune, if not inevitable, to assess whether the available statistical tools in an NSI can adequately deal with the potentially new data. That is, issues of storage and processing must be examined explicitly, as the amounts of data may be vast and conceivably may require special software and analytic tools. These will have implications not only of a technical nature but also on skills required to manipulate and use such data.

At the end of this stage, we shall have adequate information to assess the strengths and the weaknesses of the new data. Whether meeting our initial expectations or having found new uses, we will have worked with the data for some time and will have documented all that matters to make a recommendation that will get us past this decision point.

Table 3. Fictional example of implementation of stage 3 of the proposed accreditation procedure.

Accreditation procedure for secondary data Stage 3: Forensic Investigation - Usability			
actions	instructions	indicators	notes
a) Final microdata file <i>edit specs</i> duplicate records flow validity/wrong values consistency/wrong values outlier detection/treatment mandatory fields	remove check data patterns, if applicable correct through edits check internal consistency, correct define criteria, eliminate identify	% % % % % %	combine % of target (non-response) with findings from editing, outliers, missing mandatory fields, and set a standard. If below, stop. Here assume the standard is exceeded.
DECISION POINT 3	Recommendation: Proceed		
imputation, rules absent records item non-response mandatory fields weighting estimation derived variables b) Produce aggregates by domain time series analyze levels analyse movements c) Produce outputs standalone input to X input to Y d) Assess technical tools systems and software skills	missing or wrong specify, define and document in metadata compare to source and elsewhere	% % % % - - - M M L M M H M L Y Y Y	very difficult to reconcile the two data sets existing tools adequate (Y) or new needed (N) if new software, what skills are needed
Additional comments:			
DECISION POINT 4	Recommendation: Proceed		

Stage 4: NSI decision

Having come thus far, it is time for a corporate decision. This stage is dedicated to the assessments necessary for such a decision to be made based on as much information and knowledge as possible. Much of the work needed has already been accomplished, and it becomes a matter of putting it all together in a comprehensive and coherent fashion.

As a first step, we need an account of the outputs and the indicators quantified during the previous stage. However, they must be re-packaged to fit the occasion. What is needed is an itemisation of the exact uses of the new data and their impacts. What specific new output/s can be produced that will expand the NSI's offerings, which output/s can benefit, to what extent, how, and what would be the implications and trade-offs? For example, *“new POS data can replace half the retail trade survey. If we proceed, we eliminate the response burden on half the respondents (X thousand and XX million hours of burden) and save Y euros per year. This*

change will not affect the release's relevance, it will improve its timeliness by one week, but it will decrease its accuracy by an estimated 2%". Estimates of the impact on timeliness, for instance, can be obtained by comparing the time lapse from the reference period between the existing and the new data source, while estimates of the impact on accuracy can be had by assuming that the old estimates are correct and computing the difference arising from the utilization of the new data. This is then a management call and, to the extent possible, such summary must be done in a sharp and "clinical" manner.

A second step entails a top-level cost-benefit analysis, which focuses on the financial picture. Best we know, what are the extra costs and savings from the introduction of the new data? For one, we may not have to pay the source but we may have to reimburse some expenses they will incur to accommodate our needs or may have to dedicate resources for a reasonable quid pro quo. We may generate efficiencies in our survey-taking because of the new data but, on the other hand, we will likely have to absorb extra costs to integrate the new data into existing products. Which outweighs which? This is the time to bring all that together in a concise way. The suggested indicators are consistent with those of Blue-ETS¹.

The third step places the emphasis on the risks that need to be undertaken and managed by the NSI. Aside from output issues and financial matters, what else could be the impact on the NSI from such a decision? How vulnerable will be the outputs involved, and by consequence the reputation of the NSI, to factors outside its control? What will be the mitigation strategies? This is where some outputs may be of paramount significance to the NSI. What if, despite many benefits from the new data, the release of the CPI or the GDP is in jeopardy? From a risk management perspective, whatever decision is taken must be an informed one.

A final step before making a decision at this point involves the need to go beyond the purely statistical and practical matters discussed above. Effectively an analysis leading to an assessment of the feasibility of incorporating a new source into the gamut of an NSI's statistical operations from a legislative and socio-political point of view would be desirable. Such issues are dealt with in deliverable D2 of this project.

Quality, indispensable as it may be, is not the only issue on which corporate decisions are made. NSI management has to weigh in multiple, and at times conflicting, interests and make decisions based on the totality of issues. In the process of assessing the new data source, data quality issues must be combined with financial, legal, and risk management issues. Moreover, examples of negotiations that started with legal and jurisdictional issues lasted long and did not go far. This is one of the reasons why it is more prudent to start with the data. The benefit can be twofold: information needed for the next level will be known, and there will be a clear identification of trade-offs to guide and facilitate negotiations.

¹ <http://www.blue-ets.istat.it/>.

Table 4. Fictional example of implementation of stage 4 of the proposed accreditation procedure.

Accreditation procedure for secondary data Stage 4: NSI Decision - Usability			
questions	description	indicator	impact
outputs			
standalone (new)	interesting quartely release	H	meets quality standards
input to X	can substitute for some survey data	M	survey content will be reduced
scale		% of variables	
response burden		% of hrs	burden lower
timeliness		% of days	timeliness of X up by 1 moth
accuracy		% of survey estimates	accuracy of X estimated at 2% lower
input to Y	partial to a few domains	L	complicates work flows
accuracy		% of previous	reduced by 10%
coherence	some incompatibility with Y	L	problematic due to heterogeneous sources
auxiliary	useful to confront data, help seasonal	H	no quality affected
cost-benefit		+/-	
new source	extra effort is required	% of total costs	costs for X will increase by XXX
survey	survey sample can be reduced	% of survey cost	costs for survey will decrease by YYY
efficiency gains		% of total X cost	YYY-XXX
...			
risks/vulnerability		H	
punctuality	fixed delivery date cannot be missed	H	X is a critical output, any delay endangers ability to release
changes	any unilateral change disruptive	-	
technical specs	softwarecycles and compatibility	H	
costs	no costs now	L	
program	occasional changes introduced	M	
other	unknown	H	
feasibility analysis		M	
legislative matters		M	
socio-political factors		L	
Additional comments: Data can produce a new standalone output on a quarterly basis. Also, quite useful for auxiliary purposes. Not conducive as inputs to output Y. Useful for output X, eliminating X% of survey content reducing response burden. Gains in timeliness partially offset by some loss on accuracy. Source is currently free but integrating it into the production of X means that we cannot afford non-delivery or surprise changes.			
DECISION POINT 5	Recommendation: Proceed		

Stage 5: Formal agreement with source

This final stage involves high-level negotiations with the source as an institution to secure cooperation and arrive at a formal and comprehensive agreement. The NSI is now well equipped with the information it needs for such deliberations. The initial information asymmetry vis-à-vis the source has been largely eliminated.

At the outset a good understanding is needed that willingness to cooperate is not an abstract notion but matched by deeds. The early implications of this translate to obligations by the source to commit needed resources, and the NSI to respect lines that the source may not want crossed. In defining the ability to cooperate much will depend on the type of the source – public or private, statistically inclined or not, stage of advancement etc.

Then issues of reciprocity involved in a fair deal must be explicitly clarified. Terms and conditions of the agreement will be discussed in detail, supported by accompanying documentation from the working teams. At the end, the issue of governance needs to be articulated, including change management and a dispute resolution mechanism.

This stage in the accreditation procedure can also be subject to quantifiable indicators as they emerge both from knowledge of what is involved and attitudes. They will serve well in subsequent rounds, complemented of course with the experience accumulated at that time. (The table uses a scale from 1-5 but alternatives are possible).

Table 5. Fictional example of implementation of stage 5 of the proposed accreditation procedure.

Accreditation procedure for secondary data Stage 5: Formal Agreement - Cooperation		
questions	remarks	indicators (1-5)
Source - institution		
willingness		5
provide microdata & metadata on ongoing basis		5
engage in continuous communications		5
ability to share		4
resources/dedicated contact	both organisations will name contacts	4
commit to adhere to quality standards	best effort	4
commit to undertake actions as needed	best effort	4
disclosure controls	source satisfied with our confidentiality standards	4
reciprocity		3.3
costs involved	no direct costs, only some related to requirements of data transfers	3
quid pro quo	source could use some help in the process from our personnel	3
other	they appreciate credits, and sharing analytical findings relevant to them	4
terms of agreement	see attached documentation	3.5
detailed delivery specs		4
specific timelines		4
technical, IT matters		3
corrective actions		3
governance		3.5
interactions	quarterly meetings at senior level	4
sharing of responsibilities		4
change management	yes for major revisions, but for smaller changes limited	3
reciprocal timely information sharing	at the working level	3
joint quality assurance	they are open to recommendations	4
dispute resolution	signatories of the MOU	3
other	in the future, we can have input to redesign of program forms	4
Overall assessment		3.9
Additional comments:		
DECISION POINT 6	Recommendation: Proceed	

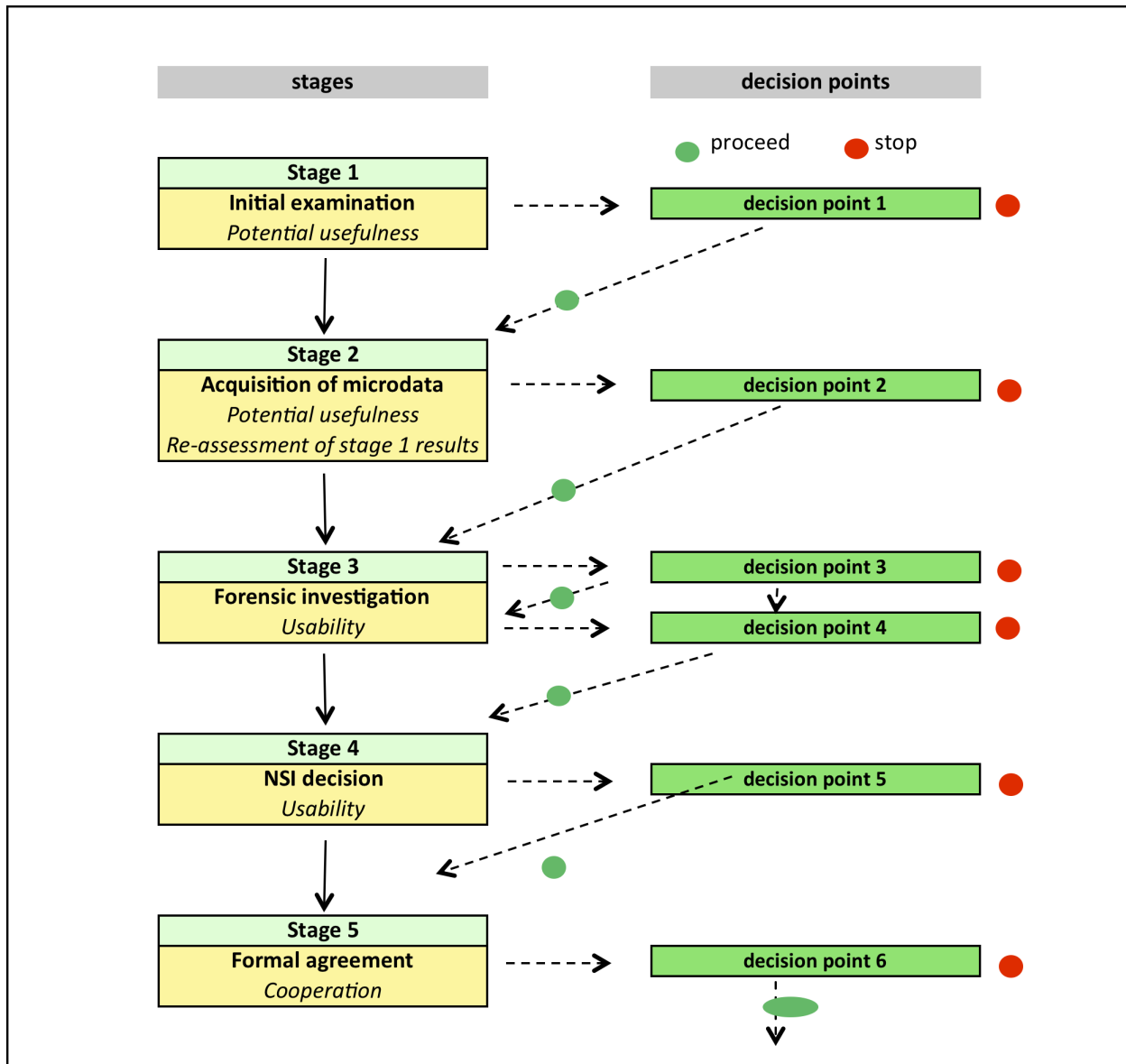
Summary

The stages of the accreditation procedure are depicted synoptically in the flowchart below. They are accompanied by a mapping of the decision points discussed earlier. In applying the accreditation procedure, a few additional issues must be taken under consideration.

Depending on the domain of interest it may be that more than one data sources should be examined at the same time and a comparative assessment be made. In the process, NSIs could well consult with others and take advantage of work done, whether in assessing data sources or having actually acquired them. This is likely to occur as in the integrated European system many needs and practices are common. As a minimum, applying the accreditation procedure to a potential data source should be communicated to others – although it is not undesirable to have more than one assessments depending on the data source and its national significance. Such cross-fertilisation becomes particularly useful when there are sources of multi-national coverage. Under such circumstances, cooperation among NSIs would be beneficial both for their resource

implications as well as for the eventual comparability of data across countries. In such a case, an NSI might consider making use of the resources, methods, tools and overall experiences of other NSIs.

Figure 2. Proposed procedure for the accreditation of non-official statistical data sources.



Sources vs. methods

The accreditation procedure outlined above is flexible for application in various contexts. It can be applied to assess secondary data from public and private sources, such as a Web site and a ministry register; it can be applied to more than one data sources within the same organisation, such as drivers' licences and vehicle registrations from two registers in a ministry of transport; it

can be equally applied for similar sources across many organisations, such as information from all credit card companies, if all of them are desired for the data to be useful.

Frequently, in talks about big data, the utilisation of data from secondary sources (as organisations) is mixed with discussions of data sources with reference to tapping digital footprints. The two must be differentiated, though. New data sources, such as scraping web sites of enterprises or gaining access to individuals' smartphones as explored in this project, are in reality new **collection methods**. Web sites are proxy respondents for businesses, providing the information we "request" and which the respondents have already put there themselves. The same holds true for individuals. "Source" in these cases is not one or a few organisations but thousands of respondents. While, then, all existing or augmented quality assurance apparatus for collection methods apply, the accreditation procedure is not meant for this purpose.

4. Casting the net wider

The world of data is changing rapidly. As we contemplate new collection methods and develop accreditation procedures for the acquisition of secondary data, we can benefit from developing an understanding of the major forces underway that are already shaping the overall statistical landscape. There, all kinds of data proliferate and co-exist. This overview is driven by a practical orientation stance and is intended to advance this line of work through stimulating more thinking and exchanges.

In the new order, statistical data, as we knew them, are no longer the (almost) exclusive prerogative of the "official" system. Until recently, NSIs were the key providers of most statistical information needed for the functioning of an economy and society. Their statistics covered a wide area, yet not everything, and generally they were credible and enjoyed a good reputation. Mathematically and inevitably – and probably fast – NSI data are becoming a diminishing fraction of all available data.

Like everything of a transformative nature, this is associated with both advantages and drawbacks. A key drawback would be the possible inability to navigate through a vastly expanded array of data and differentiate legitimate from illegitimate data for the same object of investigation. What happens in this case, when we are clearly outside the realm of official statistics? Are we entering a vacuum with free-for-all? In some ways, this is reminiscent of what transpired a bit earlier with the Internet as a whole. While, all the knowledge has come to within everyone's reach, questions linger as to what is accurate and solid and what is not. The early days of Wikipedia serve as an example. Short of assuming a perfect user, who can ascertain at a glance which offering is good, what else can be done?

With no claim of being exhaustive, a few thoughts are offered here to that effect. To separate matters from the earlier analysis, we make use of the term certification here. It must be clearly

understood that this is not related to the accreditation presented in this document, and that it is clearly a longer-term prospect.

4.1. Certification

Our thinking starts with whether or not it is desirable that some quality standards are established to allow users to sift through the world of statistics with a certain degree of confidence, and in a way that separates the good from the not-so-good. If yes, who will do that, and how?

Most data producers today, deliberate or accidental, do not have to abide by known quality standards. Worse, standards as such do not exist except for those specific to NSIs. At this point, there is no widespread agreement, established approach or mechanism to take this matter on – in a way comparable to ISO certification. While the official statistical system has neither a monopoly on data nor can it become the police of the data world, it does have a moral authority and a protagonist role to play by virtue of its history on quality.

There is more to this. Ascertaining the quality of data and their sources, and eventually arriving at some certification, presupposes that someone is asking for it. To our knowledge, the doors of the official system are not flooded by applications to do so. On the contrary, the ongoing discussions - and the work in this project – concentrate on the NSIs going after new data sources. Through that lens, the balance of powers in negotiating is not one of strength. External sources may be willing to accommodate such needs only up to a certain point. Even if that was not an issue and all sources eagerly cooperated, what is the limit of today's official system in absorbing all that is useful before being inundated and paralysed? Can it really continue to ever-expand? The main implication from this analysis is that alternative courses of action may be worth exploring.

Potential certification would certainly be one of those, and could be used to expand what is “official”. Several possible scenarios can be contemplated, depending on the type of source. Some may well see statistics as part of their business, whether as a primary or secondary activity. These should be encouraged and supported. Others would be negative to the whole idea and become “accidental” data providers with no interest to enter that space. Yet others may pose additional challenges, as not only they see statistics as part of their business but approach it strictly from a commercial, profit-making point of view. Different solutions will be needed tailored to the particular circumstances encountered.

The impact on our overall approach starts to be visible with the example of an organisation with substantial data holdings, advanced-enough in its ways, and a positive predisposition. In such a case, the opening in our Stage 5 would be quite different. Rather than trying to establish the organisation's willingness to cooperate and share their data, it would start with whether the organisation wants to be certified as a data producer in that particular area. The issues and

questions asked would assume a very different approach. For instance, they would be aimed at ascertaining if the organisation would consider adopting the existing quality frameworks, issuing quality statements, adopting and abiding by provisions of confidentiality including penalties for their breach, and generally adhering to most principles that guide the work of the statistical system.

Alternatively, it may be that the organisation neither wants to be certified as a statistical producer not to share data with an NSI but to work instead towards the idea of federated data. A modified set of standards might be applicable in this case.

The above discussion should factor in the fact that there are already examples of credible data producers. Whether implicitly considered authoritative or not, central banks are the sources for data on interest rates, exchange rates, money supply measures and more, stock markets for stock prices, volume of transactions etc. There are also weather statistics, sports statistics (FIFA has a statistical team) and many more. These tend to have “exact” data, not subject to sampling or revision, and they release data systematically and historically. Others will be very different. Moreover, their data that may not be relevant for ever and will vary tremendously. Passports and drivers’ licenses are expected to continue to be issued; utility billings or POS data are also expected to continue to exist, even if a specific utility or retailer does not. The longevity of a particular online social network may or may not materialize but this affects only the time horizon of the data and not their utility.

5. Summary and conclusions

The statistical system continues to evolve and is constantly looking for new sources of data and modern methods of collection. While administrative data are used for some time, to varying degrees depending on the institutional set-up of countries, there are more systematic efforts underway for NSIs to acquire and use data from secondary sources. Such sources could come in many different types, from public to private, profit and non-profit, statistically-inclined or not. Moreover, the data from such sources can lead to the production of new outputs in areas that expand the reach of NSIs, can be used as inputs in the production of existing outputs with different importance to the NSI, or can be used as auxiliary sources for a variety of other uses. Naturally, all these bring to the fore the need for some accreditation to guide such efforts.

Much of the work involved in this task is quality-related. Recent literature exists, and the approach proposed in this report relied on that. Some refinements were also introduced and foundational principles were formulated. The hyper-dimensions used cover the source – both as content and institution – metadata, aggregate data and microdata. Quality is assessed both for the source and its data as inputs into the production of statistical outputs and for the outputs in their own right. The quality characteristics of potential usefulness, usability and cooperation were postulated for the source and its data as inputs, and they were combined with the hyper-

dimension objects to define quality areas. The quality attributes used for statistical outputs are the same well-known dimensions embedded in the existing culture: relevance, accuracy, timeliness/punctuality, coherence/comparability, accessibility/clarity. Newness was introduced as an additional attribute for new outputs.

The resulting accreditation procedure is step-wise, with five stages involving six decision points. It is front-loaded with early gating, and allocates work as necessary. In such a process, the role and responsibility of the NSI is brought to the forefront and not delegated. Moreover, the approach recognises explicitly that in addition to the issues of quality there are other elements at play that must be balanced by a corporate entity, such as an NSI. Financial considerations, risk tolerance, and associated trade-offs are all matters that must be examined together for a responsible end decision.

At the end, the analysis expands the horizons of this report by linking to aspects of the broader picture that drive today's evolution, and with an eye on tomorrow. The message is that the accreditation procedure must be situated in its time and place, and that it may be viewed as one of several possible answers to what the future may hold.

References

- Australian Bureau of Statistics, (2013), “Data Quality Statement Questions, General Purpose: Administrative Data”, https://www.nss.gov.au/dataquality/PDFs/DQO_Admin.pdf
- Blue ETS Project, (2010), “Measurement Methods for the Indicators”
- Iwig William, Berning Michael, Marck Paul, Prell Mark, (2013), “Data Quality Assessment Tool for Administrative Data”
- Daas Piet, Ossen Saskia, Vis-Visschers Rachel, Arends-Toth Judit, (2009), “Checklist for the Quality Evaluation of Administrative Data Sources”, Statistics Netherlands
- Daas Piet, Nederpelt Peter, (2010), “Application of the Object Oriented Quality Management Model to Secondary Data Sources”, Statistics Netherlands
- Eurostat, (2003), “Quality Assessment of Administrative Data for Statistical Purposes”
- Eurostat, (2007), “Handbook on Data Quality Assessment Methods and Tools”
<http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf>
- Eurostat, (2011), “European Statistics Code of Practice”,
http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice
- EU, (2012), European Statistical System “Quality Assurance Framework of the European Statistical System”
http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF
- Laitila Thomas, Walgren Anders, Walgren Britt, (2011) “Quality Assessment of Administrative Data”, Statistics Sweden
- Nederpelt Peter, (2010), “A New Model for Quality Management”, Statistics Netherlands
- Nederpelt Peter, Daas Piet, (2012), “49 Factors that Influence the Quality of Secondary Data Sources”, Statistics Netherlands
- Statistics Canada, (2010), “Quality Guidelines”, <http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm>
- UK, Office of National Statistics, (2011), “Guidelines for Measuring Statistical Quality, (2011),
<http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html>
- UNECE, (1992), “Fundamental Principles of Official Statistics”,
<http://www.unece.org/stats/archive/docs.fp.e.html>
- Willis-Nunez Fiona, (2013), “The Role of Big Data in the Modernisation of Statistical Production”,
<http://www1.unece.org/stat/platform/display/msis/Final+project+proposal%3A+The+Role+of+Big+Data+in+the+Modernisation+of+Statistical+Production>

Annex 1: Over-arching Quality Frameworks

Even before statistical outputs per se, this conditioning starts with the reputation of the organisations involved. At a high level, official statistics internationally are guided by the UN's Fundamental Principles (adopted by UNECE in 1992 and by the UN Statistical Commission in 1994). Notions of impartiality and freedom from interference, transparency and trust based on scientific and professional standards and ethics, and paramount respect for confidentiality are central. Even no discrimination by source, specifically between survey and administrative data, is spelled out. In many ways, albeit at a high level, these principles indeed set the stage of the qualities expected of a modern statistical system.

Fundamental Principles of Official Statistics

1. Relevance, impartiality and equal access

Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

2. Professional standards and ethics

To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

3. Accountability and transparency

To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

4. Prevention of misuse

The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

5. Sources of official statistics

Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

6. Confidentiality

Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

7. Legislation

The laws, regulations and measures under which the statistical systems operate are to be made public.

8. National coordination

Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

9. Use of international standards

The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

10. International cooperation

Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

In Europe, in order to realize the vision and the mission of the Statistical System, the quality framework is epitomized by the more detailed European Statistics Code of Practice (http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice). The Code elaborates 15 principles that cover the institutional environment, the statistical processes, and the statistical outputs. Each of those principles, in turn, contains a reference set of indicators of good practice that should guide the implementation of the Code among all the organisations that are part of the European Statistical system. Many of the principles and indicators are relevant and applicable to the theme of this report and they will be used, particularly those related to quality as it relates to accreditation.

Principles of European Statistics Code of Practice	
Institutional Environment	
	1. Professional independence
	2. Mandate for data collection
	3. Adequacy of resources
	4. Commitment to quality
	5. Statistical confidentiality
	6. Impartiality and objectivity
Statistical Processes	
	7. Sound methodology
	8. Appropriate statistical procedures
	9. Non-excessive burden on respondents
	10. Cost effectiveness
Statistical Output	
	11. Relevance
	12. Accuracy and reliability
	13. Timeliness and punctuality
	14. Coherence and comparability
	15. Accessibility and clarity
Source: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice	

Moreover, to guide and assist with the implementation of the Code, the supporting Quality Assurance Framework of the European Statistical System (ESS QAF) has been developed (http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/QAF_2012/EN/QAF_2012-EN.PDF). This is an instrument that contains an even more detailed prescription of activities, methods and tools that can facilitate the practical steps needed to adhere to each indicator and principle. The ESS QAF covers the principles of the Code that relate to statistical processes and statistical outputs, as well as principle 4 (Commitment to Quality) of the institutional environment – with

four indicators devoted to it specifically. Other national and international organisations follow suit with quality assurance frameworks, also largely based on ISO standards. Such materials too are at the core of the theme of this report and they will be factored in the overall approach.

In an expanded sense of quality, NSI responsibility to users has come to the point to transcend national boundaries and work through Eurostat and global agencies to harmonise standards and definitions across countries as well, to facilitate international comparability.

At the level of outputs, wide acceptance that quality is multi-dimensional has led to the following quality dimensions generally embraced by Eurostat and, with minor variations, all NSIs: relevance, accuracy, timeliness and punctuality, comparability and coherence, accessibility and clarity. See, for instance, the quality guidelines by the ONS (<http://www.ons.gov.uk/ons/guide-method/method-quality/quality/guidelines-for-measuring-statistical-quality/index.html>), Statistics Canada (<http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm>), and the ABS (<http://www.abs.gov.au/ausstats/abs@.nsf/mf/1520.0>)

At a more detailed level, as far as surveys are concerned, many and detailed measures are produced to accompany the data in a way that quantify the sampling error. Whether in the form of standard errors (SEs) or coefficients of variation (CVs), which in turn may translate in country-specific quality scales, quality measures are ever-present. Estimates not considered of good quality are either not published or their lower reliability is explicitly flagged. However, it is fair to say that for the most part this apparatus focuses on 20th century data production, a big part of which was relying on surveys.

Annex 2: Examples of secondary sources

Drivers' licenses: Assuming one register for the whole country (or aggregation of all individual registers by state, province etc.) and without knowing the details of database fields, legal requirements, missing data etc. let's take a look at questions that the data can answer:

Since it's a census of all individuals with a driver's license, it can provide an exact answer to the question: "How many individuals in the country have a driver's license?" This is not the same as "How many individuals in the country know how to drive", which we may collect from a survey, and which may be higher due to unlicensed drivers (suspensions, under-age drivers etc.).

Now, if the register contains mileage driven too, or if it could be added, so much the better. We can expand the range of statistical information from the same source. The same applies to types of vehicle etc. Since it's a census, the quality should be excellent. A very important part of quality of such data, not part of the traditional arsenal at NSIs, is the personal involvement of the individuals and their vested interest in the correctness of the data. They are involved effectively as part of data clean-ups for their own interest.

Passports: This source contains the total number of passport holders, which is a big part of the population. Any available information, for reasons related to national and personal vested interests is deemed good. Its analytical usefulness, though, may be limited to specific inquiries related to passport holders.

Credit cards: Frequently mentioned as one of the potential sources for big data. This is not a census of the passport, drivers' licences, birth certificates or ID types. Without having seen an external actual database as it exists in a bank, clearly the grand total must be the number of individuals holding a credit card from the company – linked perhaps with other family members who may be supplementary cardholders on the same primary account. An important key will be the credit card number. As well, fields will contain a lot of personal information, including full name, address etc. which can be considered accurate for billing verification, but also occupation, estimated incomes etc. that may not be.

The important thing is that the number of cardholders is a total with limited usefulness, unless we have the same for all companies in the country. Then, we can answer questions like "How many individuals have a credit card?", and do so by type of card, credit limits, monthly purchases, and other details too.

This example raises the issue of additional quality across many secondary sources (perhaps reduced to lowest common denominator?)

12.8. D7 – Powerpoint presentation to the Information Society Working Group



Internet as a data source project

(Project implemented by **Agilis S.A.** and the **Greek Free / Open Source Software Society**)

Data everywhere

- Proliferation of digital data around us, generated by
 - the activities of individuals in the Internet (e.g. social networks, blogs, YouTube, etc.)
 - the interactions and transactions of individuals and private enterprises with public authorities and other private enterprises
 - the automated interactions between devices (e.g. between sensors and servers): “the Internet of things”
- Increasing number of interconnected devices
 - computers, tablets, smartphones, etc
- 70% of data are created by individuals
- 80% of data are stored and managed by enterprises.

Internet / Big data for Official Statistics

- A potential data source that cannot be ignored
- Not all available in the Internet; proprietary data
- Ways to exploit them: still a research topic
- Research and technological developments (software, hardware) are expected to facilitate the production of accurate and reliable official statistics

Internet as a data source project

- Definition of IS indicators based on Internet data
- Feasibility of IS indicators based on:
 - automatically generated Internet usage data
 - information available on the websites of enterprises
- "Cookbook" for implementing the methods and processes for IS indicators
- Potential of big data repositories as data sources for official statistics (any domain, not only IS)
- Procedure for accreditation of big data repositories by producers of official statistics

IS data collection: enterprise websites

- Random sample of enterprises → get site address
- OR –
- Sample of websites: from list or by crawling
- Site owners accept that a 'crawler' harvests data
- The crawler collects and transmits data about site facilities to the producer of official statistics
 - keywords
 - identified website technologies
- A questionnaire is also administered

Indicators about sites: pros and cons

- Reduced burden
- Speed; rich detail
- Coverage, when business register not available
 - Isolate business websites
 - Isolate national websites
 - One enterprise \leftrightarrow many sites
- What about randomness?
- Need for crawlers for dynamic content
- Need for very site-specific crawlers
- Fear of breach of privacy
- Legal constraints

Indicators about sites, through IaD

- Language options
- Last update date
- Secure access (https)
- Certified communication (SSL)
- Registration facility
- Registration technology (e.g. openID, Facebook)
- Site map
- Usage of online web analytics tools

Indicators about sites, through IaD

- Facility for reception of orders
- Number of orders received via the site
- Links to multimedia content (audio, videos etc)
- Links to social networks or blogs
- Content linked to multimedia sharing sites (YouTube, Flickr, etc)
- Links to wikis and wiki-based sharing tools

Indicators about sites, through IaD

- W3C accessibility guidelines compatibility
- Search tool availability
- Ability to sign up for alerts (e.g. RSS feeds)
- Accessibility by visually or listening impaired persons
- Availability of version for mobile devices
- Provision of calendar of events
- Subscription functionality (e.g. newsletter, listserv)
- Online surveys/polls; Comments; forum; chat or instant messaging

Pilot survey of websites (ongoing)

- Country: Greece
- Sample design of ICT survey was not replicated
- Convenience sampling from available list of sites
- Sample size: 316 sites
- Collection of data about 16 indicators with the use of
 - keywords and
 - Google's Custom Search Engine
- All indicators correspond to availability of site facilities and take values 0 and 1

Pilot survey of websites: results

- Keywords do not offer enough specificity
- For example: detection of keyword “email” does not always mean that the site provides the enterprise’s contact email.
- Mistaken identification of indicators hovers around 10% of all identifications
- Moreover, more than 2/3 of sites that possess a given facility are not detected

IS data collection: individuals

- Random sample of individuals
- They accept to install monitoring software on their devices (e.g. PC, smartphone, tablet)
- The software records and transmits activity data to the producer of official statistics
- Users can switch it off at will
- A questionnaire is also administered for variables that cannot be collected by the software

IS indicators – individuals: pros and cons

- Reduced burden
- Speed; rich detail
- No recall issues
- Coverage? Track the same user on many devices
- Measurement? Distinguish specific types of activity
- Difficulties with iOS devices → undercoverage
- Fear of breach of privacy
- Users may turn the software off occasionally
- Legal constraints

Indicators about individuals, through IaD

- Access to ICT – technical characteristics
- Distribution of “Internet session” duration
- Distribution of total daily session duration
- Value of goods / services bought or ordered over the Internet
- Interaction with e-gov. sites
- Number of emails with attachments
- Volumes of movie or music files downloaded
- E-skills: activities carried out

Pilot survey of individuals (ongoing)

- Country: Greece
- Sample design of ICT survey cannot be replicated
- Sample selection: panel from market research company
- Initial contact by email; screening questions
- Installation of monitoring software on PCs and Android devices; kind of parental control software
- Online questionnaire for additional characteristics
- Response rate is expected to be small
- Possible software issues

Big data repositories

- 'Big data' not necessarily the same as 'Open data'
- Federated Open data: (big) data from business and public sector shared in an agreed and defined way with other partners (e.g. producers of official statistics).
- Can such data sources be used for official statistics?
- Under which conditions?

Five examined repositories (ongoing)

- Ship location and movement data (AIS, LRIT) → transport and emissions statistics
- Real estate classified ads newspapers → data on asking prices for renting / selling property
- Facebook → consumer sentiment indices
- Greek government transparency service. Data about all expenditure decisions of all levels of government → financial statistics
- Credit card transaction data → financial statistics

Accreditation procedure

- Big data repositories are one more potential data source for producers of official statistics
- Users of official statistics trust producers that they provide statistics of high quality
- Producers of official statistics must therefore evaluate repositories before adopting them as data sources
- The project proposes a draft accreditation procedure of five steps

Outline of draft accreditation procedure

- Step 1: Assessment of potential usefulness, without concern about the feasibility of acquiring the data
- Step 2: Assessment of what data will be provided
- Step 3: Data validation; evaluation of quality of official statistics based on the data
- Step 4: Corporate decision of the producer of official statistics about whether to use the data source: cost-benefit and risk analysis
- Step 5: Negotiations with the source to arrive at a formal, comprehensive agreement for regular data provision