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:	50721.2013.003-2013.169		

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

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 censors 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 being 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 CO_2 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 (CO_2, NOx, SOx, 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.

MarineTraffic[1] is a service that provides real-time information about ship movements and ports, mainly across the coastlines of many countries around the word.

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 APIs[2] 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/09/2012[3]).

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 MarrineTraffic. Information about the gross tonnage of vessels that arrived at a port is provided by MarrineTraffic'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 MarrinesTraffic'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), Breath (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 Globwave[4], which provides values for wind (velocity, direction) characteristics. Although this information is not included in MarrineTraffic'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 are 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 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 constrains 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 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 MarrineTraffic. 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 XE[5], 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 policies[7]. 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. InHousing 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 study[8] 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 price[9]. 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 being 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 price[10]. 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 being 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 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 fake[11]. 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 variables[12]:

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 content[13].
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 information[14].

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 exhaustive[15]):

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 tweet[16];
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:

- DataSift[17] - provides access to past data as well as streaming data
- GNIP[18] - provides access to streaming data only
- Topsy[19] - 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 possible[20].

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(word| keyword), with the overall probability of the word P(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 is 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 word'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 report[21], 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 Index[22] and the Sweden Expenditure Index[23] 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 Markit[24], 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

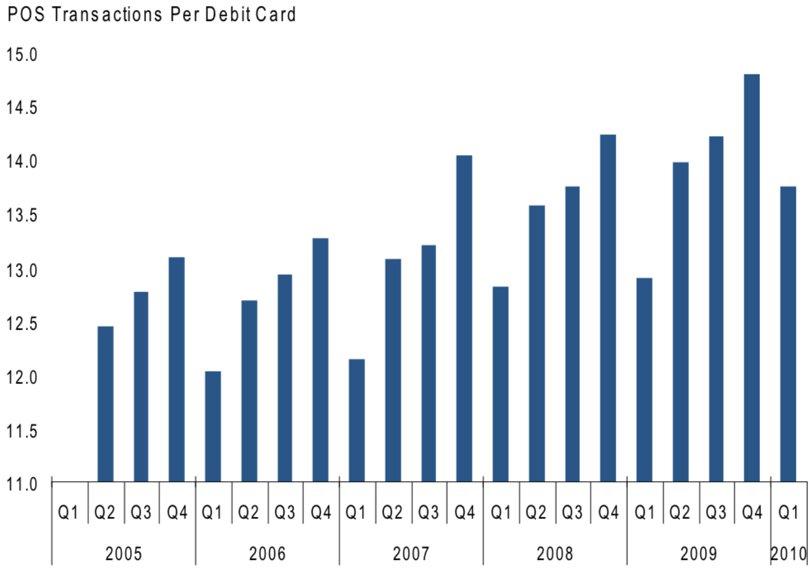
In Europe, there are more than 419 million Visa cards. Specifically for Visa debit card, the average number of transactions per card was 13.7 in the first quarter ending March 2010 (Figure 1).

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

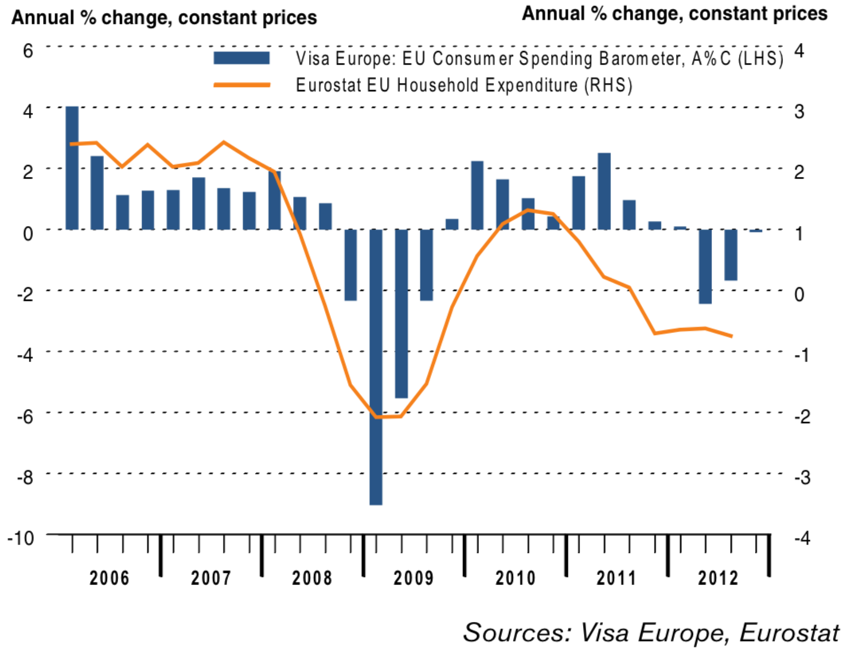
Data from the EU Consumer Spending Barometer indicate a strong relationship over time with the relevant official household spending data.

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.

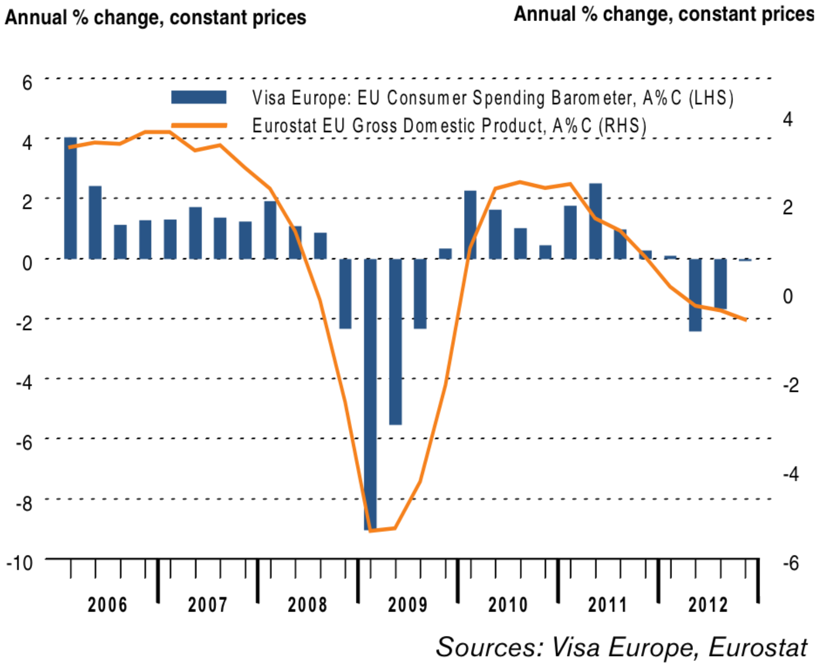
[25] http://www.visaeurope.com/en/idoc.ashx?docid=0f9b8b34-2bb9-4d08-a184-6b6d7b649c4e&version=-1

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).



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).



Source: Visa Quarterly report on EU Consumer Spending Barometer (March 2013)

Spending by product categories

[26] http://www.visaeurope.com/idoc.ashx?docid=7974d478-c525-4211-8e32-06660f7392f9&version=-1

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 recent then the HBS. However, NA cover expenditure from a macro level and thus expenditure cannot be correlated with characteristics pertaining to different households.

[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

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 expenditures[29]:

- **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» («δι@ύγεια») site. The law[30] ensures that a broad range of decisions of public entities are not enforceable if they are not first uploaded on the «δι@γεια» 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, «δι@ύγεια» is unique in the sense that it includes the totality of decisions, in a centralised infrastructure and in harmonised way.

«δι@ύγεια» 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 «δι@ύγεια» 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 («δι@ύγεια») 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 tags[32]
- 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 triples[33].

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.diavgeia.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» 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 NSI[36] are required to publish the relevant information in «δι@ύγεια». 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 data[37] 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» 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 authoritative[38] 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 «δι@ύγεια» 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 «δι@ύγεια» 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 «δι@ύγεια» 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 «δι@ύγεια» 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 «δι@ύγεια». 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 «δι@ύγεια»
    - Substantially improve timeliness.
- «δι@ύγεια» 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 data[40].

* 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 Netherlands[41] 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.

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

[41] See deliverable D2 of the present project.

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.