ESSnet Big Data

Specific Grant Agreement No 1 (SGA-1)

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata

Framework Partnership Agreement Number 11104.2015.006-2015.720
Specific Grant Agreement Number 11104.2015.007-2016.085

Work Package 2
Web scraping / Enterprise Characteristics
Deliverable 2.1
Legal aspects related to Web scraping of Enterprise Web Sites
Version 2017-02-15

Prepared by:

Galia Stateva (BNSI, Bulgaria)
Olav ten Bosch (CBS, Netherlands)
Jacek Maślankowski (GUS, Poland)
Alessandra Righi, Monica Scannapieco, (ISTAT, Italy)
Matthew Greenaway, Nigel Swier (ONS, United Kingdom)
Ingegerd Jansson, Dan Wu (SCB, Sweden)

ESSnet co-ordinator:

Peter Struijs (CBS, Netherlands)
p.struijs@CBS.nl
telephone : +31 45 570 7441
mobile phone : +31 6 5248 7775
Table of contents

1 Executive Summary .............................................................................................................................. 2

2 Present Regulations for Official Statistics, Privacy and Copyright ....................................................... 3
   2.1 BNSI - Bulgaria National Statistical Institute................................................................................ 5
   2.2 CBS - Statistics Netherlands ......................................................................................................... 6
   2.3 GUS - Główny Urząd Statystyczny / Central Statistical Office of Poland........................................ 7
   2.4 Istat - Italian National Institute of Statistics .................................................................................. 10
   2.5 ONS - Office for National Statistics of United Kingdom ............................................................... 12
   2.6 SCB - Statistics Sweden................................................................................................................ 13

3. Challenges and recommendations from NSIs’ legal departments .................................................... 13
   3.1 BNSI – Bulgaria ............................................................................................................................ 14
   3.2 CBS – Statistics Netherlands ........................................................................................................ 15
   3.3 GUS - Główny Urząd Statystyczny / Central Statistical Office of Poland....................................... 15
   3.4 Istat - Italian National Institute of Statistics .................................................................................. 16
   3.5 Office for National Statistics (ONS) - United Kingdom ............................................................... 16
   3.6 SCB – Statistics Sweden................................................................................................................ 17

4 The access to Job Portals: a case study ................................................................................................. 17

ANNEX 1 - Netiquette for web scraping for official statistics: a proposal...................................................
1 Executive Summary

This report is aimed to review the EU and Member States legislation on the Regulation of Official statistics and in areas such as Personal data protection, Copyright protection and Database protection, in order to understand the real possibilities for the NSIs to perform activities of web scraping - on small or large scale - on the websites of enterprises.

The report shows that all the six Member States (Bulgaria, Italy, Netherlands, Poland, Sweden and United Kingdom) involved in Workpackage 2 on web scraping / Enterprise Characteristics have a statistical law guaranteeing to the NSIs similar prerogatives regard to the data access and the data processing.

The laws on personal data protection, the protection and copyright protection on the database have been ratified in almost all countries and some countries in addition to these legislative measures are also mentioned the development of ethical codes (Italy, UK, Netherlands).

A picture quite favorable to the data collection through web scraping of websites of enterprises seems, in general, to emerge, but with different shades. It ranges from most favourite situations (Italy, Netherlands, Bulgaria) to expressing greater uncertainty (UK, Poland, Sweden).

This uncertainty is also reinforced by the conclusions of 2014 the EU Commission report suggesting that the act of creating a database from web scraped data could breach database rights set out in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996, at least if essential parts of the database are scraped.

As to the challenges and recommendations from NSIs' legal departments on possible alternative approaches to achieve the agreement with enterprises on websites web scraping on a large scale, different degrees of intervention appear to be necessary at country level.

Overall, a strategy based on full compliance with the statistical laws, the measures for the privacy protection and other related regulations (i.e., on database protection or on database collection from governmental agencies) should be put in place, along with a strong communication strategy disseminated through the NSIs websites explaining in detail who is involved in the web scraping activities carried out, for what purposes (public interest, research,...) and how the data are collected. This last issue increases the need to focus both on the ethical codes which could ensure NSIs from reputational risks (a problem highlighted especially by ONS) and on the creation of a code of Netiquette for web scraping for official statistics. The proposal made by ONS and CBS in the Annex 1 of this Report therefore intends to meet this need. The proposed Netiquette has been discussed by the countries participating to WP2 and there is a general agreement on adopting it.

A good cooperation between NSIs and e-traders is proposed by Bulgaria and Poland that recommended to work for the introduction of a new law for non-commercial use of the data and further data processing with Big Data tools, as the current law or provisions in agreements / rules of websites refers to commercial use of data.

Furthermore, in order to understand the specific legal aspects of the different use cases, in collaboration with Workpackage 1 on Web scraping of Job Vacancies some recommendations on
possible solutions (on the short and medium term) have been proposed, addressing the issue of the access to job portals for the data collection of job vacancy statistics. The Project Team concludes with the recommendation that NSIs engaging projects to web scrape job portals should contact targeted job portals directly to explain the nature of the web scraping, how it will be done and how the data will be used or, alternatively, to negotiate agreement on getting data directly. If permission is granted (or at least not refused) then this should remove all practical legal, ethical and reputational concerns around accessing job portals data.

Finally, we mention that Eurostat launched a survey on legal issues related to big data conducted by a consortium led by SOGETI. The main purpose of the survey (designed for national legal correspondants) is to study both the current and upcoming relevant legislation (MS and EU) that could have an impact on issues related to access and use of Big Data for statistical purposes and the obstacles in national legislation regarding the use of Big Data for official statistics. The conclusions of this study will be compared with the positions of the law offices the National Statistical Institutes expressed in this report.

2 Present Regulations for Official Statistics, Privacy and Copyright

This chapter is aimed to review the EU and Member States (for countries participation in the ESSNet WP2 Project) legislation on data collection activities core subjects for the NSIs, i.e. Regulations for Official Statistics, Privacy protection and Copyright protection laws at national level for the countries involved in the WP2.

It should be remembered, in fact, that from the point of view of the enterprises, there are three main issues that can potentially raise when a party accesses via Web scraping the data they publish on their Web sites, namely the violation of:

- copyright protection;
- privacy rights;
- ethical principles in accessing data.

Copyright is a legal right created by the law of a country that grants the creator of an original work exclusive rights for its use and distribution. Copyrights are considered territorial rights, which means that they do not extend beyond the territory of a specific jurisdiction. While many aspects of national copyright laws have been standardized through international copyright agreements, copyright laws vary by country.

A preliminary analysis of the legislation on copyright protection of the EU has identified the following list of Directives:

- Directive 2009/24/EC of the European parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version)\(^3\);


Two international treaties refer to this topic also: the Bern Convention concerning the International Commission for the Rhine and the TRIPS agreement (trade-related aspects of intellectual property rights).

Furthermore, in 2014, a European Commission report titled “Analysis of methodologies for using the Internet for the collection of information society and other statistics” provides a legal opinion on the legal feasibility of web scraping more generally. The main issue focuses on the “sui generis” database right set out in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. This database right is considered as property right and it suggests that the act of creating a database from web scraped data could breach database rights, at least if essential parts of the database are scraped.

Nevertheless, a general reflection on the meaning of copyright for official statistics could be synthetized by underlining that there are no major problems with the copyright protection. The Law No. 633 of April 22, 1941, for the Protection of Copyright provides for an exception in case of use of the database for non-profit scientific research purposes (Art 64-sexies of law no. 633/1941 as amended). Though statistical research is not explicitly mentioned, it appears reasonable to argue that NSIs can access and use information from online databases to process this information into anonymized statistical results, even when the “sui generis” database right protects them. Moreover, in the statistical law is often expressed the concept that NSIs collect data basically for public interest purposes.

Furthermore, the “EC Regulation no. 45/2001 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data” sets out the principles that legitimize the collection and use of data collected and that make lawful their treatment. Consequently, a European body such as Eurostat, operating in compliance with established regulations, should not have major legal constraints doing the web scraping activity as third party with respect to Member states. Anyway, even if a possible web scraping activity made by Eurostat is not forbidden, it is necessary to pay attention and analyze all the circumstances in which it should be made, the type of processed data, the considered sources and the type of dissemination. Thus, practical reasons more than legal reasons make the direct web scraping activity not advisable, unless a specific in-depth analysis is carried out. In addition, this topic needs a further investigation on the different impacts at country level (e.g. some countries could have

---

7 https://www.wto.org/english/tratop_e/trips_e/trips_e.htm
adopted measures on cybersecurity or have some other country specific rules) before a conclusion could be reached.

With respect to privacy rights, data published on Enterprise sites are intended to be publicly available. However, for privacy guarantees there is a major issue to consider, namely: ensuring that the “purposes” for which data are accessed prevent from possible privacy breaches.

On the personal data protection, the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) provides more specific rules for the case in question.

Finally, ethical principles in accessing data, are also discussed. A Netiquette (Network Etiquette) should be established when accessing web sites (for instance, it could be advisable to identify oneself when reaching sites for scraping).

2.1 BNSI - Bulgaria National Statistical Institute

The legislation framework for the use of data for Official Statistics in Bulgaria mainly consists of:

- **Statistics Act**: The National Statistical Institute shall collect process and store individual and personal data and statistical information;
- **Personal Data Protection Act**: Personal data must be captured for specific, precisely defined and legal purposes and not be submitted to additional processing in a manner incompatible with such purposes; additional processing of personal data for historical, statistical or research purposes shall be allowable provided the administrator has ensured proper protection guaranteeing that such data are not being processed for any other purposes, except in the cases explicitly provided for in this Act;
- **Personal Data Protection Act**: Personal data may be processed only provided when processing is necessary in order to comply with an obligation imposed on the personal data administrator by a piece of legislation.

Individual data received and collected through statistical surveys shall constitute a statistical secret and may be used only for statistical purposes. Individual data received for the purposes of statistical surveys may not be used as evidence before the bodies of the executive and the judiciary. (as defined under the Statistics Act)

The National Statistical Institute and statistical authorities and their staff may not disclose or provide:

- individual statistical data;
- statistical data which can be matched in a way that enables the identification of a specific statistical unit;

---

• statistical information which aggregates data about less than three statistical units or about a population in which the relative share of the value of a surveyed parameter of a single unit exceeds 85 per cent of the total value of such parameter for all units in the population (art. 25 Statistics Act).

Privacy protection

The National Statistical Institute and statistical authorities shall have the obligation to ensure protection of individual data and prevention of any misuse thereof by undertaking appropriate organisational and technical measures and allowing such data to be handled only by persons who have signed an affidavit for protection of statistical secrecy.

Receipt, processing, usage and storage of statistical data representing statistical secret shall be carried out in a procedure set out in a regulation endorsed by the Chairperson of the National Statistical Institute (Statistics Act).

2.2 CBS - Statistics Netherlands

The legislation framework for the use of data collected from the internet for Official Statistics in the Netherlands is based on the following legislative and practical artefacts:

• Dutch statistical law (“Wet op het Centraal bureau voor de statistiek, 20 november 2003”), publicly available at [http://wetten.overheid.nl/BWBR0015926/2017-01-01](http://wetten.overheid.nl/BWBR0015926/2017-01-01)
• Dutch Database law / Legislation on protecting intellectual property rights (“Databankenwet, 8 july 1999”), publicly available at [http://wetten.overheid.nl/BWBR0010591/2008-03-26](http://wetten.overheid.nl/BWBR0010591/2008-03-26)

Dutch Statistical Law

According to the Dutch statistical law, enterprises and individuals are obliged to provide Statistics Netherlands (CBS) data on request. In case of refusal Statistics Netherlands may enforce this using a fine. However, it is also a strategy in the Netherlands to reduce the response burden created by government agencies as much as possible. These two facts combined have led to the strategy that whenever it is possible to retrieve the same data from other sources, Statistics Netherlands will not ask the respondent. The web is also considered a valuable source of data. Retrieving data from the web instead of collecting them from a respondent thus helps reducing response burden and helps implementing the Dutch statistical law.

A second aspect to be noticed in this context is that the Dutch statistical law states that Statistics Netherlands uses its data for the production of official statistics and for research purposes only. Statistics Netherlands is not allowed to publish data that discloses information about individual persons or enterprises. For web scraping this means that even if the scraped content contains
privacy-sensitive data, it will never be published as such. We conclude that the use of data collected from the internet for the production of official statistics is allowed, as long as the data is processed with the same care and under the same conditions that hold for data collected using traditional means, which is the case.

**Dutch Database law / Legislation on protecting intellectual property rights**

The Dutch database law states that commercial re-use of intellectual property in which others have invested a considerable amount of effort, is forbidden. Since Statistics Netherlands will never publish or re-sell scraped content as such and only use it for producing aggregated statistics, web scraping by Statistics Netherlands does not infringe the dutch database law. The same holds for legislation on intellectual property rights.

**Legislation on data protection**

Statistics Netherlands processes data collected from the public web with the same care as any other possibly privacy sensitive data. This means that, once collected, the data is processed and analysed in a safe environment and that statistical disclosure control is applied to the statistics before publishing. We conclude that the under these conditions the use of data from the internet for official statistics is allowed with respect to privacy protection law.

**Netiquette**

Statistics Netherlands strategy is to comply to best practices and informal guidelines for scraping from the internet community, also known as “netiquette”. In particular this means that:
- Statistics Netherlands respects the robots exclusion protocol (robots.txt)
- Web scrapers operate using a clearly defined user-agent string (CBSbot), so that website owners are able to distinguish a statistical visitor from regular visitors
- Statistics Netherlands provides information on web scraping activities on its website
- Web scrapers operate “nice” to websites: they leave some time between requests and operate preferably at quiet periods (nights).

In addition, Statistics Netherlands has the policy to inform website owners if a considerable amount of data is collected on a regular basis.

### 2.3 GUS - Główny Urząd Statystyczny / Central Statistical Office of Poland

The legal framework of official statistics is based on the following laws:
Laws and regulations relevant for web scraping for general purposes

The law issued on 29 June 1995 on official statistics in article 5 p. 1 gives a general permission to collect and store the data from all data sources specified in this legal act as well as in other regulations or in different legal acts. Therefore it is a legal act that is used by official statistics to confirm formally the basics of data processing. Next articles concerns detailed description of data acquisition and processing, such as article 13 – gathering data from administrative data sources, article 18 – programme of statistical surveys of official statistics, article 21 p. 2 – conduct at individual requests statistical surveys not included in the programme of statistical surveys of official statistics, art. 18a and 28a – the way of sharing the data including the detailed format of the data to be transmitted, chapter 4a processing of personal data for statistical purposes in art. 5 p. 1. Those rules apply to methods of data gathering already used in official statistics, i.e., providing this data by respondents or administrative data owners, in which official statistics is a receiver of ordered statistical data. The rules of this law do not apply to other forms of data gathering, e.g., by official statistics by its own activity, without informing the second part.

Big Data analysis for experimental purposes is possible because of the obligation for official statistics to conduct research work as well as research and development work, including methodology or mathematical methods and applying IT methods in statistics. This work must be done bearing in mind the protection of personal data.

The law issued on 18 July 2002 on providing services using electronic forms generally concerns electronic trade, but some regulations may have an impact on web scraping. Based on art. 3a of this Law, providing services using electronic forms is dependent on law of EU and EFTA member country – parties of agreement of European Economic Area – in which service provider has a headquarter or a place of residence. However this rule does not apply to copyrights protection and other forms, according to the Law issued on 4 February 1994 on copyrights and similar laws. Art. 8 of this Law concerns necessity of defining and sharing the rules of providing services in electronic forms by the service provider. Art. 12 and next ones regard disclaiming the responsibility of service provider in terms of providing services by electronic forms. Chapter 4 concerns personal data protection by providing services electronically. Indicated in article 18 personal data of customers, which service provider can process, is partially covering personal data that official statistics can process based on the Law on official statistics.

A particular aspect is on personal data processing. General remark: because of publishing the Regulation (EU) 2016/679 of The European Parliament And of The Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), it is necessary to change the Polish law according to this regulation. Conclusions below are according to current Polish law.
Law issued on 29 August 1997 on personal data protection includes rules of personal data processing. Based on article 23 of this law, data processing is possible when it is necessary to execute all rules from the law for the public good and when it is necessary to execute privileges or fulfil duties written in the law.

Moreover art. 26 p. 2 of the law issued on 29 August 1997 on personal data protection, allows the possibility of personal data processing for other purposes than the primary purpose for which the data collection had been conducted (secondary processing of the personal data). It can be done only with respect to law aspects and freedom of the person, whose data is processing and this processing must be done for statistical purposes.

Official statistics has additional detailed regulations on personal data processing (see chapter 4a of the Law on official statistics).

**Privacy and database protection**

Law issued on 27 July 2001 on database protection is determining the rules on which databases are protected. According to the Polish Law, databases are protected based on the Law issued on 27 July 2001 on database protection, and based on the Law issued on 4 February 1994 on copyrights and similar laws the database is protected as a whole or a part of the database.

If the database is not legally protected based on the Law referred above, important is a Judgment of The Court of Justice of the European Union, issued on 15 January 2015 (C-30/14) in which the Court rules:

“Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases must be interpreted as meaning that it is not applicable to a database which is not protected either by copyright or by the sui generis right under that directive, so that Articles 6(1), 8 and 15 of that directive do not preclude the author of such a database from laying down contractual limitations on its use by third parties, without prejudice to the applicable national law”.

Law issued on 29 June 1995 on official statistics has several rules for statistical data protection, according to the standards of the European Statistics Code of Practice. In Poland the guarantee of data confidentiality is a statistical confidentiality – art. 10 of the Law. Identifiable individual data collected in statistical surveys are absolutely protected. It means that the data cannot be published without anonymization and aggregation. These data can be used only for preparing statistical publications and analysis as well as to make a frame for the statistical surveys by the President of the Central Statistical Office. Sharing or using this data for purposes different than in the Law is prohibited (statistical confidentiality).

All persons having access to individual data are obliged to obey the law. These persons are submitting written promise (an agreement to respect the privacy of the data) based on the art. 12 of the Law.
2.4 Istat - Italian National Institute of Statistics

The legislation framework for the use of data for Official Statistics in Italy is mainly based on the three following legislative artefacts:

- National Statistical System\(^{10}\) (NSS) law (d.lgs nr. 322, Sept 6 1989);
- Personal data protection law\(^{11}\) (d.lgs. nr. 196, June 30 2003, in the following named as Privacy Code);

With reference to such legislation, the following major points are relevant to the task of accessing, storing and processing data from enterprise web sites:

- Istat and other components of the NSS can use personal data for their institutional functions in compliance with principles of ‘pertinence’ and ‘no excess’.
- Personal data can be collected only for ‘defined aims, which have been made explicit and legitimate, and used for other tasks that are compliant with such aims’ (Privacy Code art. 11).
- ‘Handling personal data for historical, statistical or scientifical aims is considered as complaint with aims according to which data have been collected or handled previously’ (Privacy Code art. 99).
- The NSS decree introduces the concept of the National Statistical Program. This program is a three-year program that has a complex iter ending with the approval by the National Authority for Privacy (in Italian Garante per la Privacy).
- Our regulation does implies duties to notify enterprises for which specific usage the data will be treated.

Given point 3, it is always possible to use for statistical purposes personal data that third parties have collected according to whatever purpose.

Given the point 4, Istat proposed a project about scraping of enterprise Web sites in the last National Statistical Program. In the project, Istat had to provide details in particular on the following items:

- Purposes and ways according to which Web scraped data would have been handled.
- Data sources, i.e. Enterprises web sites.
- Collaboration with other parties.
- Phases of the statistical process in which data would have been handled.
- For personal data, time after which data would have been deleted.

Given the point 5, Istat must design a strategy according to which notify enterprises of the scraping activity and of the related usages of data that will be done.

\(^{10}\) http://www.istat.it/en/files/2010/12/dlgs322_89.pdf

\(^{11}\) http://www.camera.it/parlam/leggi/deleghe/03196dl.htm

\(^{12}\) http://www.camera.it/parlam/leggi/deleghe/99169dl.htm

\(^{13}\) http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1556573
The Legislative Decree of May 6, 1999, nr. 169 concerning "Implementation of the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases defines ‘database’ as a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. Furthermore, it states that the copyright protection of databases shall not extend to their contents and shall be without prejudice to any rights subsisting in those contents themselves.

The author of a database shall have the exclusive right to carry out or to authorize:
(a) temporary or permanent reproduction by any means and in any form, in whole or in part;
(b) translation, adaptation, arrangement and any other alteration;
(c) any form of distribution to the public of the database or of copies thereof. The first sale in the Community of a copy of the database by the rightholder or with his consent shall exhaust the right to control resale of that copy within the Community;
(d) any communication, display or performance to the public;
(e) any reproduction, distribution, communication, display or performance to the public of the results of the acts referred to in (b).

The Legislative Decree nr. 169/99 provides for limitations on the rights set out if there is use for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved. As part of these activities to access and consultation, any permanent playback operations of all or substantial part of the contents of other media are still subject to the authorization of the rightholder.

Copyright and Privacy

The copyright is taken into account by a controlled use and distribution of data gathered through the Web scraping activity. Indeed, Istat provide this kind of detailed description within the National Statistical Program regulated by the NSS Decree. The NSS Decree regulates the obligation to provide statistics not only for all government bodies and public organizations, but also for private entities, provided that they are included within the National Statistical Program. As mentioned Istat has also the obligation to notify enterprises for which specific usage the data will be processed. This is a way to take into account the “creator permission” on one side and the “fair usage” on the other side needed to respect copyright.

According the Privacy Code personal data are “any information concerning a natural person, identified or identifiable, even indirectly, by reference to any other information, including a personal identification number”. The Privacy Code identifies the following steps for processing personal data; “Personal data undergoing processing shall be: a) processed lawfully and fairly; b) collected and recorded for specific, explicit and legitimate purposes and used in other processing operations in terms compatible with those purposes; c) accurate and, where necessary, updated; d) adequate, relevant and not excessive in relation to the purposes' for which they were collected or subsequently processed; e) kept in a form which permits identification of data for a period of time not exceeding that necessary for the purposes for which they were collected or subsequently processed.”

In the case of scraping of enterprises web sites personal data can be present but are not the main focus of the activity. However, Istat will take into account such possible presence in order to be compliant to the Privacy Code.
**Ethics**

The Ethics Code ensures that the processing of personal data will be in respect for human rights and fundamental freedom, as well as dignity of the person concerned, with particular reference to confidentiality, the identity and the right to protection of personal data. As remarked, the scraping of enterprises web sites could occasionally involve the processing of personal data. In such cases Istat will process such data in order to be compliant to the Ethics Code.

In addition, Istat is also defining a Netiquette protocol for accessing Enterprises web sites.

**2.5 ONS - Office for National Statistics of United Kingdom**

The operation of the UK’s statistical system is governed by the Statistics and Registration Service Act 2007. Other relevant legislation includes the 1947 Statistics of Trade Act.

The Statistics and Registration Services Act empowered the UK Statistics Authority to create a National Statistics Code of Practice. This Code sets out common practices to be followed by National Statistics – for example, on proportionate burden and confidentiality.

The UK Government has a set of Information Security Protocols; the key points from these items of legislation, codes of practice and protocols with respect to web-scraping of enterprise data are:

- **Section 26 of the Statistics and Registration Service Act** gives the UK Statistics Authority (of which the ONS is the executive office) a statutory right to produce and publish statistics relating to any matter relating to the UK or any part of it, and allows for the gathering of information within this function, although information gathering via the internet is not mentioned explicitly. However, in gathering information the Authority must act within the law. In the context of web-scraping, the ‘within the law’ stipulation is particularly relevant with respect to copyright law, as discussed in section 3.6.

- **The 1947 Statistics of Trade Act** gives the ONS the legal authority to compel business enterprises to provide information for statistical purposes. However, in general the wording of the Act is framed in a way that is not relevant to the modern digital age, and it is very prescriptive about issuing notices to businesses. As such, we do not have a clear picture and considerable uncertainty remains about the legal basis for web-scraping.

- **Principle 6 of the National Statistics Code of Practice** mandates practices on proportionate burden, including to minimise the cost burden on data suppliers relative to the benefits arising from the use of the statistics. This implies that scraping should be carried out in a way which minimises any cost to the website owners (for example, by introducing a delay in the scraping), but it also forms an argument in favour of scraping, as it may lead to a reduction in business compliance costs through filling in surveys. Principal 5 of the National Statistics Code of practice mandates practices to follow on Confidentiality, including to keep confidential business information secure.

**Copyright**

The legal guidance on copyright issues with respect to web-scraping prices from supermarket websites (at present the main application of web-scraping at the ONS) suggests that in this
context, it is unlikely that any relevant copyright subsists in scraped data. Anyway, if copyright does subsist, ONS would have a strong case for copyright exemption since the actions of the ONS are carried out under statutory authority.

ONS have been advised that, while the status of copyright with respect to web-scraping is unclear, the overriding issue in copyright protection is to prevent someone using someone else’s data in a way that has a detrimental financial impact on the actual owner of the data. We can therefore mitigate some of the risk of copyright infringement by ensuring that there is no financial impact.

In summary, there is still considerable uncertainty with respect to copyright protection and web-scraped data, and in the absence of test cases this uncertainty is likely to persist in the near future. There are several lines of argument which imply that the risk financial risk of copyright infringement in web-scraping business data may be relatively low, however, there is also a significant reputational risk involved in any suggestion of copyright infringement.

2.6 SCB – Statistics Sweden

The legal framework of official statistics

According to the directives, Statistics Sweden is responsible for producing and disseminating official statistics on number of (but not all) subject matter areas. The recently established Strategy 2020 states that “Statistics Sweden meets the current and future needs for reliable statistics as a basis for analysis, discussion and decision-making.” It is further declared that “To ensure input data for statistical purposes, Statistics Sweden steadily seeks and examines new potential sources of information.”

Important laws and regulations at the national level for production of official statistics are:

- The Personal Data Act (Personuppgiftslag 1998:204, https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/personuppgiftslag-1998204_sfs-1998-204, will be replaced by Regulation EU 2016/679 on data protection);
The legal framework provides Statistics Sweden with the necessary legal rights to collect and process data, and to disseminate aggregated information, while at the same time protecting the rights of individuals, households, businesses, and organizations.

**Laws and regulations relevant for web scraping for general purposes**

Web scraping from Internet is a new potential data source for official statistics, and the collection of such data must follow the general framework given by bullet point 1 to 4 above. In addition, regulations on copyright are relevant (see below).


Databases are protected as a related right to copyright, for the efforts of collecting and organizing data. For example, sales and exhibition’s catalogues, timetables, and address calendars are results of investments in information collection and presentation. They are protected as databases. The European Database Directive (96/9) gives similar protection.

Many enterprise websites state that the content of the web pages is under the protection of the Copyright law and only allow conditional use of the web pages. To respect their copyrights and the database rights, prior agreements with the owners of the web sites can be negotiated.

With respect to a general description of privacy protection of enterprises while data are collected, processed and stored, Statistics Sweden must take proper measures to protect data during processing and storing. The Official Statistics Act states that data cannot be disseminated unless it is made certain that data can be disclosed without causing any damage to businesses or organizations.

**3. Challenges and recommendations from NSIs’ legal departments**

In this chapter alternative possible approaches to achieve the agreement on websites web scraping with enterprises on a large scale (that is thousands of enterprises) are discussed on the NSIs’ legal departments point of view basis. We try to understand if different legal issues are highlighted according to different use cases and we collect recommendations on possible solution in short and medium term.

**3.1 BNSI – Bulgaria**

With respect to the question “How to achieve a general agreement with enterprises on a big scale?”, first best is to work with e-traders on voluntary base. In order to establish a good cooperation with respondents we will submit to them aggregated data. In addition the aggregated data we will provide
an analysis of e-commerce: dynamic, average priced across products, etc. We hope this information is valuable for e-traders and will motivate them to cooperate with the NSI.

A good cooperation between NSI and e-traders is a must to have a reasonable estimation of the e-trade in Bulgaria. In order to reach thousands of enterprises, we will organise a series of meetings with organisations that represent the interest of e-traders and have a saying on the e-trade legal framework in Bulgaria. The meeting format could be seminars, conferences, private meetings, etc.

Finally, the BNSI’s legal department considers that there are no legal constraints on the web-scrapping activities (even for different use-cases) at present.

3.2 CBS – Statistics Netherlands

Statistics Netherlands has the policy to operate as transparent as possible and to minimize the response burden. Using data from the internet instead of using surveys is a logical consequence of this. Performing scraping in such way that it agrees to common practices on internet (Netiquette) as well as the policies to take care of intellectual property rights, data protection and to communicate with enterprises if considerable amounts of data are retrieved on a regular basis were developed in close cooperation with the legal department of Statistics Netherlands. The application of these policies were refined over the last years from the early scraping projects back in 2010 up to the larger bulk scraping as it is performed in 2016. Up to now, these policies have proven to be effective for the tasks performed.

For achieving agreement with enterprises on a big scale the same policies should be applied but some practical implications might be different. In this case the statistical law in combination with the communication of the best practices should be our main instruments as well, but if the number of sites to be observed increases, it will become impossible to communicate with the enterprises on an individual basis. A more global communication strategy (website) should be put in place.

Scraping is either useful and feasible and should be performed as transparent as possible or there may be reasons to avoid it in which case Statistics Netherlands can use more traditional means to get the data. Up to now we did not see many cases of the latter.

3.3 GUS - Główny Urząd Statystyczny / Central Statistical Office of Poland

Current law or provisions in agreements/rules of websites concerns only commercial use of data. We can recommend to think of introducing a new law for non-commercial use of the data and further processing of this data with Big Data tools.

Based on art. 5 p. 1 written in the Law issued on 29 June 1995 on official statistics, it can be concluded that other data sources and data tools can be used to gather statistical data by official statistics. However it is necessary to formulate those additional forms in a legal act.
3.4 Istat - Italian National Institute of Statistics

As mentioned, Istat has the obligation to notify the enterprises about the scraping activity that Istat carries out on their web sites. In addition, Istat has to notify them for which specific purpose the scraping activity will be accomplished.

To this scope, Istat has the strategy of publishing on its official web site an informative note. This method turns to be a simple but effective one to reach a huge number of enterprises.

As far as specific recommendations by Istat’s legal department, they can be summarized as follows:

- The current regulation framework allows to access enterprises web sites via scraping activities;
- A general «iter» must be followed by web scraping projects that must be part of the national Statistical Program;
- Specific initiatives to inform enterprises must be put in place;
- Both the Privacy Code and the Ethics Code must be taken into account if personal data should be involved in the web scraping activities.

3.5 Office for National Statistics (ONS) - United Kingdom

Current practice in web scraping at the ONS is that any data obtained from the web is subject to the terms and conditions of the relevant website or web service. This covers any form of web scraping as well as data obtained through web services, such as the Twitter Public API. ONS does not currently scrape web sites that prohibit web scraping in their Terms & Conditions.

The legal position surrounding web scraping in the UK is unclear, partly due to a lack of case law. However, the material legal risks focus mainly on what is done with data once it is collected - in particular with respect to copyright law. As discussed in section 3.6, there are a number of arguments to suggest that there may be relatively little risk related to copyright infringement in web-scraping data at the ONS. However, our current practice of checking terms and conditions ensures that ONS remains a good partner organisation, and offers some protection against reputational risk.

While this approach generally works, website terms and conditions are not always easy to interpret, and there have been several cases at the ONS where data have been collected from the Internet having checked terms and conditions, only to encounter uncertainty later on. These issues have usually been resolved by obtaining consent for web scraping directly from the website and ceasing web scraping if the website raised an objection. We also seek to inform website owners before doing any large-scale web-scraping on their website, even if web-scraping is not prohibited in their terms and conditions.

This approach clearly does not work for large-scale web-scraping of enterprise websites, where checking terms and conditions or contacting websites directly to obtain consent is infeasible. We have sought legal advice on this issue, and in summary it appears that while the financial risk related
to copyright infringement may be ‘fairly low’ (as discussed in section 3.6), there is a ‘very high’ reputational risk related to large-scale web scraping as envisaged in Work Package 2, although we believe there are steps we can take to mitigate this risk. At present, we are not conducting any large-scale web-scraping of enterprise websites, although we are doing exploratory work on the methods and tools that this kind of web-scraping would utilise.

ONS is currently drafting a revised web-scraping policy at the ONS, and hope to send this for ethical and senior manager sign-off in early 2017 – in particular, we are sending a paper to our National Statistics Data Ethics Committee. We hope that this revised policy will permit large-scale web scraping without checking terms and conditions in some specific circumstances and following a set of good practices. We may seek to mitigate the reputational risk in large-scale web-scraping by going out to a public consultation on this revised policy.

However, web-scraping without checking the terms of conditions of websites is not current ONS policy, and we require further legal advice, senior management sign-off, ethical approval, and potentially public consultation before proceeding with web-scraping along these lines.

3.6 SCB – Statistics Sweden

The legal office of Statistics Sweden has investigated the general prerequisites for alternative data sources. It can be concluded that in addition to the framework given by the general laws and regulations for official statistics production, at least two legal aspects are of major importance for web scraping: collection of individual data must be avoided, and copyrights must not be violated. As a first step, we need to be allowed to test web scraping methods and tools on web pages. In order to meet the above requirements, we will contact a small number (about 150-200) of web site owners (preferably government institutions or businesses with which we already have a good communication and understanding) and ask for permission to test web scraping on their web pages. We will avoid collecting any personal data, and we will not collect entire web pages (i.e., images or pictures), thereby not violating any copyrights.

In the future, it is impossible to carry out web scraping on a larger scale if we must first receive the consent of all web page owners. It must be further investigated if general information, for example an announcement at Statistics Sweden’s web site, is enough, and for example if we can gain understanding and consent by stressing the possibility to reduce response burden.

4 The access to Job Portals: a case study

The Big Data ESSNet WP1 pilot (web scraping for job vacancy statistics) has considered legal issues as part of the wider task of gaining access to job portal data. The participating countries have received different advice from their legal departments:
• The advice for Germany (DESTATIS) is that there is no specific law against web scraping and that the terms and conditions of websites around web scraping are not legally enforceable. However, it is recommended to get the consent of the portal owners before using data on a larger scale;

• The current UK (ONS) position is that web scraping is permitted as long as the terms and conditions of the relevant website do not prohibit it. However, there may be separate ethical and reputational issues that need to be considered;

• Statistics Sweden has not yet been given clearance from their legal department to do web scraping. This is not based on any specific prohibition, but on a view that web scraping should not proceed while the legal situation remains unclear. This has led Sweden to focus on gaining access to job advertisement data from their Government Employment Agency and a few private job portal owners.

There is no evidence that these varying stances are due to different laws operating within each country. Instead there seems to be different interpretations of the law and possibly different risk appetites within the different NSIs.

A recent EU Commission report\(^\text{14}\) into the use of the internet as a source of data collection provides a legal opinion on the legal feasibility of web scraping more generally. The main issue focuses on the “sui generis” database right set out in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. This database right is considered a property right that exists to recognise the investment that is made in compiling a database, even when this does not involve the creative aspect that is reflected by copyright. It suggests that the act of creating a database from web scraped data could breach database rights, at least if essential parts of the database are scraped.

Williams (2006\(^\text{15}\)) makes a distinction between “soft scraping” and “hard scraping” of job boards. Soft scraping involves scraping and publishing summary information about a job vacancy with a link to the original advertisement. Hard scraping involves scraping job advertisements and then purporting to be the original website. The latter is clearly a breach of database rights, but even soft scraping could be considered problematic. In 2001, the German courts ruled that providing “deep links” infringed a major job board’s exclusive rights in its database of job vacancies\(^\text{16}\).

However, it remains unclear whether these database rights can (or at least would) be claimed if data from job portals were scraped for official statistical purposes. For a claim to be upheld in the courts a job portal would need to demonstrate evidence of damages. This seems extremely unlikely if the data were being used solely for statistical purposes. The business models of most job portals are geared mostly around the services of matching job seekers and employers rather than the intellectual property of the underlying data. The content of most vacancies advertised on job portals


\(^{16}\) Stepstone vs. Ofir (via http://www.linksandlaw.com/linkingcases-deeplinks.htm#Stepstone v. Ofir)
originates from material uploaded directly by employers - job portals will often claim no responsibility for the accuracy of information contained within the job advertisement. Further, many job portals allow crawling and reposting of job advertisements and may even provide APIs to facilitate this. Therefore, it is very difficult to see how the web scraping of job advertisements for statistical purposes could be considered a breach of database rights.

In early 2016, CEDEFOP (European Centre for the Development of Vocational Training17) completed a study on web scraping job vacancy data, which considered legal and ethical issues (CEDEFOP, 201618). This concluded that there was no specific legal framework for web scraping and that the only possible legal consideration was around issues of copyright. However, the conclusion was that there “are no legal issues in simply running a web crawler over publicly accessible content on the Internet for researcher purposes.”

However, the CEDEFOP study also covered ethical considerations, which were considered more relevant. The study referred to existing ethics codes, such as those developed by the Ethics Committee of Association of Internet Researchers19. Key factors included within this framework include: considering the subjects of the research, the purpose of the research, the potential benefits versus the risks of harm, recognising the autonomy of others, and how the results will be presented. In the context of job board, some ethical issues such as the rights of individuals are not relevant due to the nature of the data. Relevant ethical considerations are associated almost exclusively to the owners of job portal websites with the specific ethical considerations being around what data is collected and how, and then what use is made of it.

The CEDEFOP study determined that the easiest way of addressing these kinds of issues was to simply contact job portals directly while applying rules of copyright. This involved explaining the purpose of the research and to obtain their consent, or at least give them an opportunity to opt out. All 16 of the job portals scraped as part of this initial study were contacted. A number of job portals gave permission to web scrape, with some even entering into formal agreements and none explicitly refusing.

In conclusion, it is recommended that NSIs embarking on projects to web scrape job portals should adopt a similar approach and contact targeted job portals directly to explain that nature of the web scraping, how it will be done and how the data will be used. If permission is granted (or at least not refused) then this should remove all practical legal, ethical and reputational concerns around accessing job portal data.

There may also be other benefits of building closer relationships with job portal owners. This could include improved access to data through use of preferential APIs, access to more complete information (e.g. full job description rather than just a “snippet”) and historical data. A direct relationship may also lead to a better understanding of how the data is compiled and of the job portal’s business strategy. This in turn would inform how the data might change in future and help identify the basis for longer-term collaboration.

17 http://www.cedefop.europa.eu/
19 https://aoir.org/reports/ethics2.pdf
However, until international protocols are agreed, NSIs should also get appropriate legal advice to ensure that planned web scraping activities are not prohibited by national laws and are aligned with the prevailing organisational risk appetite. It is also worth mentioning that the principle of contacting website owners could be adopted for web scraping for other domains (e.g. Prices). Specifically, it could be used where the number of target websites is limited and where it is feasible to contact websites directly. It would not be feasible to do this for web scraping that involves scraping a large number of websites, such as the approach described by WP2, hence specific initiatives to inform enterprises must be put in place.
ANNEX 1

Netiquette for web scraping for official statistics: a proposal

M. Greenaway and O. ten Bosch

Background

The aim of this proposal is to set out key principles for good practice in web-scraping (‘Netiquette’) at ESS National Statistics Institutes. These principles are intended as a ‘lowest common denominator’, in fact they are relevant for all ESS NSIs carrying out all types of web-scraping. More stringent principles, particularly around informing website owners, may be appropriate in some cases; for example, some NSIs may adopt more stringent principles due to the legal or cultural context in their country, or when scraping sensitive or personal data.

Principles

- Respect the ‘robots.txt’ robots exclusion protocol\(^20\) and nofollow links;
- Identify yourself in the user-agent string, and provide a means for the website to contact you, which could be via a link to a web-page;
- Be transparent about your web-scraping activities, possibly by providing information on your website;
- Inform website owners if a considerable amount of data is collected on a regular basis.
- Seek to minimise burden on website owners, for example:
  - By adding idle time between requests
  - Scrape at a time of day during which the website is not under heavy load
  - When crawling multiple domains, consider ‘parallelising’ the crawl to avoid repeated requests to the same domain – submit a request to domain A, then domain B, then submit another request domain A;
  - Do not crawl sensitive areas of a website – URLs with /admin/ in, for example:
    - Only scrape data for the production of official statistics within the scope of your mandate, and do not re-use or distribute the data for any other purpose;
    - Handle web-scraped data securely according to all relevant protocols and laws.

Example web-scraping policies


ONS – currently unpublished, available on request