Statistics Finland

Memo

14 November 2019

TK-00-1597-19

Statistics Finland's general operating principles concerning web scraping

When Statistics Finland is web scraping data from the Internet for statistics production, the following general principles are applied:

A data collection decision is made on the data to be acquired by web scraping and the data are described in the data collection register. In connection with web scraping based on the data supply obligation, the normal data provision and consulting obligations of statistical authorities are applied. Risks related to the quality of data acquired by web scraping are identified and assessed before decisions are made concerning web scraping. The use of web scraping as a data collection method is reported in connection with the publication of the statistics.

Lawfulness. Legislation is taken into consideration starting from the planning stage, and it is complied with in full. Possible changes in the legal state are monitored.

Transparency. Information on web scraping is provided publicly on Statistics Finland's website, and the following details are given at the same time: purpose of web scraping, information types targeted by web scraping and contact information which the website administrator can use for contacting. In data collection based on the data supply obligation the communication obligation according to the Statistics Act is taken into consideration. If the data contain personal data, information concerning processing of personal data is published openly and is easily available on Statistics Finland's homepages.

Principle of least harm. Web scraping is implemented in all respects so that scraping will cause as little harm and costs as possible to the operation of the website and its owners.

Right to prohibit. The website administrator is given the right to prohibit web scraping by contacting Statistics Finland. The requests for prohibition are respected, and they are listed.

Compliance with statistical ethics. The procedures and principles applied to statistics compilation and codes of professional ethics in statistics and research are complied with.

Review of terms and conditions. Web scraping is directed to only such websites in which the terms and conditions do not specifically prohibit web scraping, or the prohibition is clearly restricted to apply to commercial activity. In uncertain cases the website administrator can be contacted. If no answer is received within a reasonable time, the website can be scraped.

In addition to the general principles, Statistics Finland complies with the following operating principles when implementing web scraping:

Necessity of data. The scraped data must be justifiably necessary for the compilation of statistics and bring added value for statistics production.

Intended use. Data collected by means of web scraping can be released only for purposes intended in Section 13 of the Statistics Act.

Identity disclosure (user agent string). Statistics Finland's identity, contact point for contacting and a link to the notification concerning web scraping on Statistics Finland's web pages are reported to the website.

Minimisation of burden. Websites are not burdened by excessive inquiries, but inquiries are made at reasonable intervals. If possible, web scraping is scheduled at a time of day during which the website is not expected to be under heavy load. Additional inquiries are not made.

Advance hearing in exceptional cases. The website administrator is heard in cases where web scraping would be exceptionally extensive or burdening.

Case-specific consideration. Case-specific appropriateness of web scraping is examined before scraping is started.

Robots.txt. If the website has a robots.txt file prohibiting web scraping, it is respected. An advance written permission from the website administrator is required for deviating from the robots.txt file.

If data obtained by means of web scraping are acquired from a third party, the following principles are followed:

Statistics Finland can acquire web scraped data only in case the procedures of the supplier having produced the data are not in conflict with Statistics Finland's operating principles concerning web scraping. Before an agreement is made, it must be ensured that the acquisition will not even indirectly lead to breach of the operating principles defined in these guidelines or acquisition of unlawfully scraped data.

The supplier must prove that the data are lawfully acquired either so that they were not scarped contrary to the terms and conditions of the websites, or that agreements have been made on web scraping and selling rights of data with the administrators of websites prohibiting web scraping. In addition, the data must in other respects be acquired in an ethically sustainable manner, and they must not be copied from a database protected by copyright, for example.

As a rule, data containing personal data cannot be acquired complete, but web scraping should take place on Statistics Finland's specific assignment so that an agreement about the processing of personal data is made between Statistics Finland and the supplier.

Marjo Bruun
Director
General

Timo Koskimäki Deputy
Director General of
Statistics Production

Appendix: Operating principles of web scraping at Statistics Finland

Operating principles of web scraping at Statistics Finland

Introduction

Digitalisation and growth in data volumes challenge and also enable statistics production and its development in a novel fashion. Data on enterprises and persons are increasingly available on the Internet for utilisation in statistics production as well.

Although the use of data extracted from the Internet challenges both statistics production systems and the used methods, there are also many good reasons for their use.

According to the Statistics Act, data required for the compilation of statistics should be collected as efficiently as possible, minimising the burden to data suppliers: what would be a better way than use data already available on the Internet. Data scraped from the web could also be a solution for the falling response rates of data collections, and they would make it possible to reduce labour-intensive and thus expensive direct data collections. Statistics production could also be intensified in this way. For example, web scraping of price data from enterprises' websites has been tested with good results.

There are constantly new information needs which are considered to be filled by acquiring data. Complete registers cannot always be found, or collecting of data with direct collections would be too expensive or laborious. Web scraping could offer solutions to such situations, too.

Even though many things are in favour of utilisation of data scraped from the web, using of the data in statistics production is not without problems. In addition to deficiencies related to data quality, web scraping is faced with both ethical and legal problems. Can data found on web pages be used in statistics production, in case web scraping of the data is prohibited according to the website's terms and conditions? Should permission be asked from data subjects to use data produced by them? How to view a situation where the website administrator allows web scraping only against payment?

The same challenges are considered in European statistical cooperation. The first policy concerning the use of web scraped data (ESS Web scraping policy template) was published in July 2019, and it is consistent with the practices presented in this document.

These guidelines include Statistics Finland's general operating principles concerning the use of web scraped data in statistics production. The document is updated if the legislation changes or new information and instructions concerning web scraping become otherwise available.

1. Legislation with an effect on web scraping

The compilation of statistics is guided by the Statistics Act (280/2004). The Statistics Act or other legislation does not include any actual regulations relating to web scraping. The only exception is the act on storage and retention of cultural materials (1433/2007) in which the National Library of Finland is assigned the task of retrieving and storing web data available to the public. There are no established interpretations or legal practice related to web scraping in statistics production, and practices are only being formed on the international level. At the moment, web scraping must be viewed based on general legislation and legislation concerning compilation of statistics. When evaluating the lawfulness of web scraping, three viewpoints should be taken into account: copyrights, data protection and terms and conditions.

Copyrights possibly directed to the content of websites do not as such restrict web scraping, because Statistics Finland's centre of interest is the information transmitted from the website instead of the published works, information or data. For example, copyrights do not have an effect on the web scraping of information concerning the enterprise on the enterprise’s home page. However, some web platforms consider they form a database. According to the Copyright Act (404/1961), a producer of a database requiring substantial investment has in accordance with the EU Database Directive the exclusive right to the database and the right to prevent copying or re-use of the database (sui generis right). However, in the legal practice of the Court of Justice of the European Union, acquiring the sui generis protection has required such substantial monetary or temporal resources in making the database that extremely few web platforms meet these criteria. It should be noted then that if the database is open to public, the producer of the database cannot prohibit others from looking for information there according to the Court of Justice.

If pages to be scraped contain personal data, the **data protection legislation** must also be taken into consideration before starting web scraping. The EU's General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council) and its supplementing Data Protection Act (1050/2018) provide for the processing of personal data. The person to whom the personal data pertain is called a data subject. Particular attention should be paid to the personal data included in the data to be scraped being appropriate, essential and necessary from the viewpoint of statistics production. However, downloading a website in connection with web scraping may lead to other data than those needed in statistics production being caught in web scraping. If such temporary data stored as a by-product of the technical process inadvertently contain personal data, they must be destroyed or anonymised without delay.

With respect to statistics production, web scraping of additional personal data must be minimised, and scraping must as a rule be planned so that additional personal data are not saved.

Data subjects need not be informed personally on web scraping directed to their data, as that would require unreasonable inconvenience and would prevent the attainment of the statistical objective in accordance with the exception allowed by the Data Protection Regulation. Collective provision of information is sufficient in these cases, and it is made by informing about the web scraping and the related processing of personal data on the home page of Statistics Finland. The information must cover the requirements of the Data Protection

Regulation and they must be clearly expressed and easily found.

If the data to be scraped contain personal data, information concerning processing of personal data is published openly and is easily available on Statistics Finland's homepages.

The data subjects' right to rectify, investigate, restrict and object can be exceptionally restricted by virtue of the Data Protection Act if statistics production or information need so requires. Deviating from data subjects' rights always requires hearing of Statistics Finland's employee responsible for data protection and the implementation of impact assessment.

The use of information on the website can be restricted with **terms and conditions**. To make terms and conditions applicable, the website user must accept them in connection with registration or visit on the website. Some web platforms prohibit web scraping in their terms and conditions by stating that "user has no right to use automatic systems to make copies" or that "use of automated software for copying data on the Internet pages is prohibited". Because the main ratio for such terms and conditions has been to restrict commercial utilisation of data, varying interpretations have been given on their significance for web scraping by statistical authorities. The strong main rule is, however, that no stand is taken on web scraping in the terms and conditions of websites.

2. General principles of web scraping

A data collection decision is always made on the data acquired by web scraping and the data are described in the data collection register similarly as other acquired data. In connection with web scraping based on the data supply obligation, the normal data provision and consulting obligations of statistical authorities are applied.

Risks related to the quality of data acquired by web scraping are identified and assessed before decisions are made concerning web scraping. The accuracy and timeliness of data cannot often be verified in all respects. The use of web scraping as a data acquisition method must therefore be reported in connection with the publication of statistics formed on the basis of web scraped data. If scraped data are to be used in machine learning, the impartiality of data must also be verified.

To ensure ethical sustainability, the following principles must be complied with in all web scraping done by Statistics Finland:

Lawfulness. Legislation, including data protection, is taken into consideration starting from the planning stage, and it is complied with in full. Possible changes in the legal state (legislation, legal practice, established interpretations) are monitored.

Transparency. Information on web scraping is provided publicly on Statistics Finland's website, The following details are given on web scraping at the same time: purpose of web scraping, information types targeted by web scraping and contact information with which the website administrator can get in contact to request additional information or restrict web scraping. If it is a question of data collecting based on the data supply obligation, the communication obligation according to the Statistics Act is also taken into account when collecting data.

The information obligation related to the processing of personal data is followed.

Principle of least harm. Web scraping is implemented in all respects so that scraping will cause as little harm and costs as possible to the operation of the website and its owners.

Right to prohibit. The website administrator is given the right to prohibit web scraping (opt-out) by contacting Statistics Finland. The requests for prohibition are respected, and they are reported in a common list (so-called black list).

Compliance with statistical ethics. The procedures and principles concerning data collection and designing and compilation of statistics applied to statistics production are also complied with in web scraping. The same applies to the codes of professional ethics in statistics and research that are the basis of Statistics Finland's activity.

Review of terms and conditions. Web scraping is directed only to such websites whose terms and conditions have been checked. Web scraping is considered to be allowed if it is not specifically prohibited or the prohibition is clearly restricted to apply to commercial activity. In uncertain cases the website administrator can be contacted. If no answer is received within a reasonable time, the website can be scraped.

3. Practices of web scraping

In addition to general principles, Statistics Finland complies with the following operating principles when implementing web scraping:

Necessity of data. Web scraping is directed only to such data that are justifiably necessary for the compilation of statistics. The data must bring added value for statistics production.

Intended use. Data collected by means of web scraping can be released only for purposes intended in Section 13 of the Statistics Act.

Identity disclosure (user agent string). Statistics Finland's identity, contact point for contacting and a link to the notification concerning web scraping on Statistics Finland's web pages are reported to the website.

Minimisation of burden. Websites are not burdened by excessive inquiries, but inquiries are made at reasonable intervals. If possible, web scraping is scheduled at a time of day during which the website is not expected to be under heavy load (e.g. night time). Additional inquiries are not made, but web scraping is implemented so that only necessary data are extracted.

Advance hearing in exceptional cases. The website administrator is heard in advance in cases where web scraping would be exceptionally extensive or burdening.

Case-specific consideration. Case-specific appropriateness of web scraping is examined before scraping is started. Data can also be extracted through API if that is offered.

Robots.txt. If the website has a robots.txt file prohibiting web scraping, it is respected. Where necessary, a permission may be requested from the website administrator (in writing) for deviating from the robots.txt file. Web scraping must not be started before a positive response has been obtained.

4. Acquisition of data web scraped by a third party

In the market there are enterprises whose business is based on selling and utilisation of web scraped data. Acquisition of data of interest direct from a commercial operator engaged in web scraping may be an enticing alternative. In addition to establishing cost-efficiency, it must also be made sure that the same principles have been observed in web scraping regardless of who – Statistics Finland or a third actor – has acquired the data.

Before an agreement is made, it must be ensured that the acquisition will not even indirectly lead to breach of the operating principles defined in these guidelines. Particular attention should be paid to the data not including unlawfully scraped data. The supplier must either prove that the data do not include data scraped contrary to the terms and conditions, or that agreements have been made on web scraping and selling rights of data with the administrators of websites prohibiting web scraping. In addition, the data must in other respects be acquired in an ethically sustainable manner (minimisation of burden, robots.txt, transparent identity), and they must not be copied from a database protected by copyrights.

As a rule, data containing personal data cannot be acquired completed, but web scraping should take place on Statistics Finland's specific assignment so that an agreement about the processing of personal data is made between Statistics Finland and the supplier.