Date of issue: 01.04.2022

Web content retrieval guidelines

Authors: Fernando Reis,
Kostas Giannakouris,
Albrecht Wirthmann
Head of Unit: Albrecht Wirthmann

Director: Emanuel Baldacci

Preamble

These guidelines are designed to support European Statistical System (ESS) members, namely Eurostat, the National Statistical Institutes (NSIs) and Other National Authorities (ONAs), that the data retrieval from web data sources carried out by is performed transparently, consistently, ethically, and in line with EU and national legislation. Recognising the necessity that source data of adequate quality is available for the development, production or dissemination of official statistics, content from World Wide Web sources should be retrieved and used in an appropriate and ethical manner that limits the burden on website owners and survey respondents as much as possible.

For the purpose of these guidelines, web content retrieval activities, including the use of Application Programming Interfaces (APIs) and web scraping, are defined as the automated extraction of content available on the World Wide Web. The data extracted from web content complement traditional data collection through surveys and the use of administrative sources to compile European and national statistics.

ESS partners using this technique are required to operate in accordance with:

- Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics;
- Respective national statistical legislation;
- The European Statistics Code of Practice,
- The General Data Protection Regulation (GDPR), or Regulation (EU) No 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices, and
- Intellectual property legislation (EU copyright law and respective national copyright law and the Database Directive)
- All other applicable EU and national legislation;

This document outlines key principles for good practice of retrieving content from the World Wide Web for the purposes of developing, producing or disseminating official statistics. These principles are applicable to ESS partners using techniques to collect data from Web sources. Individual ESS partners may adjust these guidelines to comply with national legislation, conforming to the national legal or cultural context, or when retrieving sensitive or personal data.

Background

The ESS and its partners are committed to using new data sources to produce statistics and analysis across the EU. Application programming interfaces and web scraping techniques enable statistical offices to collect new and more up-to-date data to produce statistical information. They provide opportunities to increase and enhance the information based on more traditional data collection methods, while reducing response burden and increasing sample size.

Simply put, rather than, for example, requiring staff to answer questions about their employers, techniques for retrieving data from the web automate the collection of data that companies publish about themselves and their goods and services.

Scope

The aim of these guidelines is to support the ESS partners in their web content retrieval work. This includes the extraction, processing and use of web data for statistical purposes.

It covers web content retrieval by ESS partners, and by third parties or intermediaries acting on behalf of an ESS partner. The ESS partners will seek to ensure that principles outlined in these guidelines are met when procuring web content retrieval services or web content retrieved by third parties.

Principles

The principles guiding web content retrieval activities aim to maximise the benefits, e.g. in terms of quality and resources, while minimising burden, risks and negative impacts arising from these activities, within the boundaries of the legal framework surrounding official statistics and the retrieval of content from the Web. To this end, the ESS partners should:

- use the web content retrieved solely for statistical purposes, as laid down in Regulation (EC) No 223/2009 on European statistics and in applicable national statistical legislation, including for the purpose of developing, producing or disseminating official statistics;
- ensure that statistical information is developed in a professionally independent manner according to the principles of the European Statistics Code of Practice, e.g. when preparing training datasets for machine learning.
- be transparent about the methods used to retrieve web content;
- process personal data in compliance with the GDPR, or with Regulation (EU) No 2018/1725 in the case of European institutions, as appropriate;

When applying web content retrieval techniques, the ESS partners should specifically:

- seek to minimise the impact on the web servers;
- inform in general about the web content retrieval activities and policies of the ESS partner, e.g. in the website of the ESS partner;
- inform website owners directly and individually when the content retrieval from the website is expected to have a significant impact on the web server, e.g. when a website is scraped with a high frequency;
- be open to making agreements with websites owners, establishing alternative content retrieval channels, such as API and file transfer and preferential retrieval time scheduling; identify themselves to the website, except in very specific cases where the web content retrieval needs to be done anonymously, e.g. for quality control or audit purposes;
- be open about the tools, methods and processes used for web data retrieval;

- abide by the website's scraping policies, in accordance with the statistical principles laid down in the Regulation (EC) 223/2009 on European statistics and the European Statistics Code of Practice;

Practices

When retrieving web content, the ESS partner:

- identifies itself via the user-agent string[1] of the web bot and provides contact channels. This should include a link to a webpage setting out the retrieval purpose and what content it collects, the contact details of the team responsible, and information on how to share data alternative to web scraping;
- follows standard internet conventions, such as standards established by the World Wide Web consortium (W3C), specifically the hypertext transfer protocol and the recommendations "data on web best practices";
- is transparent about its web-scraping activities, preferably by providing information on the associated website;
- informs website owners on the basis of institutional information (e.g., letter, or email) if a substantial amount of data is extracted on a regular basis. This should include information on the purpose and scope of web scraping, how to identify the web bot, contact details of the team responsible, a weblink to the ESS web content retrieval guidelines, inform about other possibilities to share data. The website owner should be given sufficient time to react to the communication. If no reaction is received within a specified time, ESS partners will take this as no objection and commence with web scraping activities;
- informs website owners by means of general information, such as publication on the ESS partner's website, if a website is scraped incidentally, or not substantially; seeks to minimise the impact on web servers applying measures, such as:

- adding idle time between requests,
- retrieving content at times when the web server is not expected to be subject to a heavy load (e.g. day of the week and a time of day),
- optimising the retrieval strategy in order to minimise the number of requests made to a given domain (e.g. skipping inline elements such as images or style sheets if those are not necessary);

- scrapes data only within the scope of the statistical office's legal mandate; handles web-scraped data securely.

[1] User-agent string — When a browser or web scraper accesses a webpage it provides a 'user-agent string' to the server hosting the website, and this string is then viewable by the website owner. It is possible, when building a web scraper, to modify this user-agent string so that it contains custom text — for example, to identify the operator or purpose of the web scraper

In case web scraping activities are performed without entering into an explicit agreement with the website owners, the ESS partner in addition:

- complies with the robots exclusion protocol[2] and follows links only to the extent necessary to maintain the quality of statistics;
- complies with the wishes of website owners as set out in terms and conditions, insofar as it is feasible to check those terms;

Roles and responsibilities

Staff of the statistical offices is informed when involved in web content retrieval activities of these guidelines to ensure compliance in any web content retrieval activities. The statistical office provides the infrastructure necessary for compliance.

Evaluation and Review

Eurostat shall carry out an evaluation of these guidelines one year after their implementation and submit a report to the DIME/ITDG.

[2] Robots exclusion protocol — A widely-used protocol that allows website owners to prevent any web scraping, to limit web scraping to search engines only, or to shield parts of their website from web scraping. For more details, see the robots.txt website.