Web Intelligence Hub – the case for a shared platform for using Web data in official statistics

**Keywords:** Web data, web scrapping, big data, official statistics.

# Web data

There are two types of information exhausted from the use of the Web: the content (websites and Web APIs) and traffic (when websites are accessed). This information is located in two different types of places, the web servers and the web clients (web browsers). The content of the Web is located in the web servers and is for the most part openly available or available upon a registration (and eventually the payment of a fee). The traffic is normally not openly available and is located both in the web servers (in the web logs) and in the web clients (i.e. web browsers). Sometimes, web servers make aggregated traffic data produced from their web logs openly available.

The content of the Web is expressed as HTML (HyperText Markup Language) which is stored (or generated) and transmitted by web servers and interpreted and displayed by web browsers. HTML includes both structured and unstructured information. Structured information normally follow a regular structure between web pages of the website, but it rarely follow standards across websites. Sometimes structured data is tagged within the HTML source code that makes it easier to identify and extract. Unstructured information is for the most part expressed as natural language textual data.

Traditional statistical methods are not immediately applicable to natural language textual data (NLTD). Natural Language Processing (NLP) is a sub-field of Artificial Intelligence (AI) dedicated to the application of statistical methods to the analysis of this type of data or at least to its transformation into a format to which traditional statistical methods can be applied. There are several approaches to the analysis of NLTD. The final result is the structured representation of the content of the NLTD or the (automatic) classification of the “document”. The automatic classification of the “documents” is done via machine learning where a sample of human classified cases are used to “train” an algorithm (more than a hundred different possible algorithms exist nowadays), which is then used to automatically classify other non-classified cases.

# The use of Web data for statistical purposes

On one hand, the ESSnet Big Data I launched in 2016 two pilots dedicated to the exploration of Web data, in particular content data. The first one approached job advertisements in Web portals for enhancing job vacancies statistics and the second one attempted to extract business data and enterprises’ characteristics from their websites for enhancing business registers and business statistics (e.g. ICT usage statistics). At the time of finalisation of the ESSnet Big Data I, there was a call for a refocus on the implementation of the most successful pilots towards statistical production. Answering to this call, the ESSnet Big Data II, launched at the end of 2018, included a track on implementation.

Three use cases were selected for implementation work, two of which were the ones exploring Web data.

The Big Data ESSnet apply mostly a national approach to the introduction of the exploration of Web data in official statistics, where ESS partners explore either national data sources or global sources restricting their scope to specific countries, and develop parallel and to some extent complementary research and development activities.

On the other hand, Eurostat has carried out its own exploration of Web data. Traffic data provided by the Wikimedia Foundation for the Wikipedia (Wikipedia page views) has been used to produce experimental statistics in the domains of culture statistics and urban statistics . This data source has also been explored to be used on the temporal disaggregation of tourism indicators.

Eurostat has also developed a prototype on the extraction of business data from dbpedia and Wikidata, knowledge graphs extracted from the Wikipedia, as a complement to increase the amount and timeliness of information feeding the Euro-Group Register.

Finally, it should be noted that at the same time, Web content data has been explored in other statistical domains such as online prices for prices statistics.

# The challenges of using Web data in official statistics

Web content data is normally openly accessible, but systematic access to this data source is not necessarily easy. Despite of the enabling legal situation, many websites try to prevent the exhaustive scraping of their content (by blocking access). Reasons might be posing high loads on their web servers or, in the case of Web platforms4, protection of information against automatic extraction by third parties including competitors. For this reason, the sustainable use of Web content data requires the establishment of agreements with the website owners. From the point of view of statistical offices , agreements are preferable as they may include the direct access to the database in the backend facilitating the process of acquiring the data.

The ability to process the amounts of data acquired from the Web requires specialised big data infrastructure as necessary condition to implement web data sources into statistical production. Despite the efforts of the last 5 years in making such infrastructure available to the ESS partners for the purposes of the big data pilots (Big Data Sandbox, BDTI), to this date very few NSIs have their own big data infrastructure.

Transforming Web data into official statistics requires specialised skills to develop and maintain web scrapping and machine learning algorithms. Despite the efforts of Eurostat in the last 5 years in providing big data training via the ESTP, only few NSIs have acquired sufficient skills to run a full fledged production system.

The challenges faced by European statistics in the use of Web data have to be put in the context of the data market in which official statistics now finds itself. Large players on the Web use the data generated in their activities to offer statistical/analytical services, including to public institutions and for the purpose of policy making. That is the case of, for example, Google, which uses the data generated by the use of their search engine to offer Google Trends (GT). Google Trends has been used at least by the ECB for nowcasting official statistics indicators, so far at an experimental level, but one cannot exclude the possibility of it guiding or at least influencing monetary policy in the future. Another example is LinkedIn that has been willing to offer analytical services to at least DG-EMPL to guide policy making in particular in the domain of employability of higher education graduates. Eurostat has been involved in this initiative. However, the continuation of its involvement into non-experimental, more sustainable use of their data would require capabilities beyond those currently available at Eurostat.

The situation concerning the move towards implementing the use of Web data into statistical production can be summarised as the lack of capabilities of Eurostat and of the vast majority of the National Statistical Institutes to use Web data for the production of official statistics. At the same time, there is the opportunity to take advantage of the development work of Cedefop for online job advertisements, and extend such system to other Web data sources and leap forward the use of Web data in official statistics in the domain of skills.

The description above demonstrates that the use cases based on Web data for improving the Official Statistics portfolio are numerous and technology and techniques to collect and process Web data are mature and already used by some forerunners from Official Statistics and from policy makers while methodological challenges are still important.

# The gains in efficiency from using a shared platform

The cost of establishing the WIHP including the implementation of the two use cases OJA and MNE are estimated in the table below.

The amounts for 2023 and 2024 are estimates for operations and evolution and could be projected further out in time based on actual experience.

The use cases costs include:

* development cost of the use case application (UI, workflow manager, dashboard, tailoring of components) onto the platform not including the methodological development
* the continuous and corrective maintenance of the application
* the operation of the workflow up to the release of structured/output data sets (for integration in production and/or statistical products)
* porting OJA application to the WIHP architecture

The platform costs include:

* the WIHP services i.e. the cost of the support for the use and integration of the infrastructure components (ML models, scraping and ingestion, workflow manager, data lake and catalogue) for the different uses cases
* the WIHP Minimum Viable Product (MVP) development i.e. the cost of development of the minimum set of services and component to host the 2 use cases
* the WIHP maintenance and lifecycle management i.e. the cost of continuous maintenance and gradual upgrade of platform components
* the WIHP infrastructure include the fixed and variable cost of mobilisation and use of the infrastructure and technology stack (storage, licences, network traffic...)

The estimated costs are based on a full externalisation/outsourcing for the development and operation. Eurostat staff currently attached to the project can be leveraged to fill some of the tasks below, fostering internalisation and sustainability. For more discussion on sourcing scenarios see specific deliverable.

The raison d’être for the WIHP is to achieve synergies in the form of reusing investments in technology and infrastructure as well as the reuse of data and code. The graph below quantify the implications of different investment strategies, based on the cost model developed for the WIHP.

The results of the model show that the level of reuse has a significant impact on the cost of supporting the use cases and should be encouraged by the platform.

# References

1. ESSC2018 Bucharest Memorandum on ‘Official Statistics in a datafied society (Trusted Smart Statistics)’ https://ec.europa.eu/eurostat/web/ess/-/dgins2018-bucharest-memorandum-adopted
2. ESSC2019 Implementation of the Bucharest Memorandum on ‘Official Statistics in a datafied society (Trusted Smart Statistics)’ – Trusted Smart Statistics Strategy and Roadmap, Document ESSC 2019/40/7 https://europa.eu/!bG33tw
3. DM2019 Trusted Smart Statistics Action plan and Roadmap (TSSAR 1.0), DM document Doc. 1015\_2.2
4. EC2019 Political Guidelines for the next Commission 2019 – 2024 https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission\_en.pdf
5. EC2016 A New Skills Agenda for Europe https://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-381-EN-F1-1.PDF
6. ESSNET2018 Web scrapping / Job vacancies – Strategy for ongoing engagement https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/WP1\_SGA2\_Deliverable\_1\_1\_1.0docx.pdf