Albrecht Wirthmann

The Web Intelligence Hub – A tool for integrating web data in Official Statistics

Albrecht Wirthmann[1]; Fernando Reis[1]

[1] Eurostat, European Commission

Keywords:
Official Statistics; Trusted Smart Statistics; Big Data; Web Data; Online Job Advertisements

Abstract:
The Web Intelligence Hub (WIH) is a European Statistical System (ESS) statistical infrastructure built to use data from the world wide web to regularly produce statistics complying with the quality requirements of official statistics.

The WIH is being developed in the context of the Trusted Smart Statistics initiative. The initiative is organised in hubs, with each hub specialising on data sources with similar characteristics and processing similar types of data. The WIH is the pillar of the TSS initiative that provides the fundamental building blocks for harvesting information from the Web to be used in official statistics.

The WIH follows a modular architecture in order to be easily evolvable to the technological changes occurring on the Web and to re-use components for different processes and outputs. The processes running in the hub and the corresponding outputs are open, transparent and auditable. The WIH will provide several services to the ESS, such as commonly agreed partnership models with web portals, commonly agreed data gathering, processing and statistical methodologies, algorithms and ready to run scripts. The data and scripts running in the WIH will also be available to the partners as a service.

The creation of the WIH follows the steps of the developments within the ESS, in particular, the piloting of the use of online job advertisements (OJA). In parallel, several initiatives on the collection and use of OJA data have been launched at European level. Therefore, the starting point of the creation of the WIH is the OJA use case.

1. Introduction

The use of the World Wide Web as a source of big data has become quite common both in research and in the production of statistics ([4]). Web scrapping is a quite easily implementable activity given the amount of tools freely available, in particular if it is done at a small scale. However, the use of web data in regular statistical production following the quality standards expected in official statistics more complex. In that case, the activity needs to be automated, robust, methodologically sound, transparent, reproducible, efficient, consistent and providing comparability over time. The Web Intelligence Hub (WIH) is a European Statistical System (ESS) statistical infrastructure built to address those requirements.

The WIH is being developed in the context of the Trusted Smart Statistics. Following the exploration of the use of new types of data and new data sources for official statistics during

2015-2019, the ESSC at its meeting on 16 May 2019, discussed principles and strategic orientations to guide the path towards their regular implementation as so-called trusted smart statistics ([1]). New data types and sources require new methodological approaches to transform the raw data into statistical information, to address quality issues and to be jointly processed with traditional ones. At the same time, the integration of non-traditional data sources requires new IT capabilities, enabling the processing of huge amount of structured and unstructured data and deployment of new statistical methods. As many of the new sources are held in the private sector, alternative scenarios for sustainable and secure use need to be established. Finally, ESS skills and communication strategies need to match these requirements. To provide a common umbrella for these developments and cater for synergies and economies of scale, the Trusted Smart Statistics initiative (TSS) was launched.

The initiative is organised in hubs, with each hub specialising in a specific group of data sources with similar characteristics and processing similar types of data. Each hub can serve multiple statistical domains and each statistical domain can be served by several hubs. Construction of the hubs is output-driven, centring on concrete use cases, which are defined so that they generate new or contribute to existing European statistics in line with users' needs.

The WIH is the pillar of the TSS initiative that provides the fundamental building blocks for harvesting information from the Web to produce statistics. The TSS and the WIH represent a long-term commitment, effectively launching a new process in the European Statistics production chain. The WIH's ambition is to become a high quality web data source for the production of European and national official statistics. The WIH's capabilities are gradually constructed, based on the set-up of building blocks for the collection and processing of data for specific use cases.

2. Principles and services

The WIH implements the principles adopted for the trusted smart statistics concept ([2], [3]). It is a hub providing multiple services to the several members of the ESS and to which these members contribute. It is multi-purpose serving needs at European and also national level. It follows a modular architecture in order to be easily evolvable to the technological changes occurring on the Web. The processes running in the hub and the corresponding outputs are open, transparent and auditable, with the lineage of the data and the processes being identifiable and traceable. The priority is on having the partners working together in the hub with shared processes and data, but the members should have the possibility to run their activities independently and still take advantage of the methodologies and algorithms developed in the context of the hub.

The WIH will provide several services to the ESS. In terms of data gathering, it will offer commonly agreed partnership models with web portals to be used for trans-national data agreements set by Eurostat and for national data agreements set by National Statistical Institutes. It will also offer commonly agreed data gathering, processing and statistical methodologies, as well as their implementation as algorithms and ready to run software code. In order to complement these, the hub will provide capacity building (training) material on the use of the WIH and of web data in general for statistical purposes. For those use cases implemented in the WIH, it will gather data from the Web and process it to a state ready to be used for statistical purposes. The data collected, as well as the one processed, will be available to the ESS partners. The software code running in the WIH for the processing of data in the context of the use cases implemented will also be available to the partners as a service. The WIH will provide an IT infrastructure to host the several services of the hub, the Web Intelligence Platform. Finally, the WIH should provide the means to

facilitate R&D collaboration and should identify those regulatory aspects which should be addressed by the ESS.

3. The use of web data for statistical purposes

Through a series of initiatives, in particular the ESSnet Big Data I [6][7] and ESSnet Big Data II [8][9][10], the ESS launched in 2016 two pilots dedicated to the exploration of Web 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.

There have been also European level initiatives exploring Web data. The European Commission has launched in 2013 "MOVIP - Monitor of Online Vacancies for ICT Practitioners", followed by a series of follow up initiatives based on OJA data with a particular focus on ICT practitioners.

Similarly but with a much larger scope, Cedefop launched in 2014 the "Real-time labour market information on skill requirements: feasibility study and working prototype" with the goal to assess the potential of online job advertisements data for skills intelligence ([ref]). Following the conclusion that such data source had potential and could feasibly be explored, Cedefop launched "Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis" with the goal of setting up a fully fledge system to carry out analysis of vacancies and skills based on online job advertisements

4. Online job advertisements

The ESS has followed closely the development of the Cedefop system. The data collected in the scope of the Cedefop feasibility study was used for the first European Big Data Hackathon, run in March 2017. The Hackathon had the purpose of exploring the data source for building innovative statistical products which would help a particular policy question, how to design policies for reducing mismatch between jobs and skills at regional level in the EU. The Hackathon follow-up workshop has then put together participants of the Hackathon, the development team of the Cedefop system and the members of the ESSnet Big Data I working on OJA data to discuss synergies. Until the end of its lifetime, the ESSnet Big Data I

organised several joint workshops with Cedefop and has proposed a strategy for ongoing engagement ([ESSNET2018]) where it recommends:

"It is expected that the first OJV data from the Cedefop vacancy scraping system from selected countries will start becoming available towards the end of 2018. It is hoped that data from the Cedefop will become available for use by NSIs within the ESS and will become an important, and possibly, the main source of OJV data for the ESS. It is also expected that NSIs will contribute by providing statistical expertise as well as other data sources to validate data from the Cedefop system. These data source would include Job Vacancy Survey (JVS) data and OJV data obtained from other sources."

The experience of the ESSnet Big Data in the exploration of OJA data and the development of pan-European systems with that purpose allows us to have a good understanding of what can be produced with such data. The ESSnet Big Data I has successfully used OJA data to nowcast JVS, demonstrating that it may be possible to enhance current job vacancies statistics by providing increased timeliness and granularity.

In order to establish the WIH, Eurostat will build on the methodological developments of ESSnet Big Data I web-scrapping work packages (online job advertisements and enterprises websites) and on the system already under development by Cedefop. It will generalise the architecture of the system developed by Cedefop, making it extendable to all web data sources (and to other hubs of the Trusted Smart Statistics). Finally, it will port Cedefop system to the new TSSC/WIH architecture. Concerning business data on multinational enterprises, the project will enhance the WikiData prototype and extend it to additional web data sources, with all the new developments already following the TSSC/WIH architecture.

5. The layered view of the WIH

The WIH can be seen as having a layered architecture (Figure 1). The base of the WIH is the data that is located in the World Wide Web. In the process of transforming those data into official statistics, the WIH relies on a platform (the Web Intelligence Platform) offering data acquisition, data processing and data analytics services. These follow state-of-the-art methodologies commonly agreed in official statistics implemented in the form of algorithms and open source code. The WIH is put to work with use cases. Use cases have a particular scope, reflected in a set of sources of Web data selected based on their quality, and particular outputs serving the production of specific statistics.

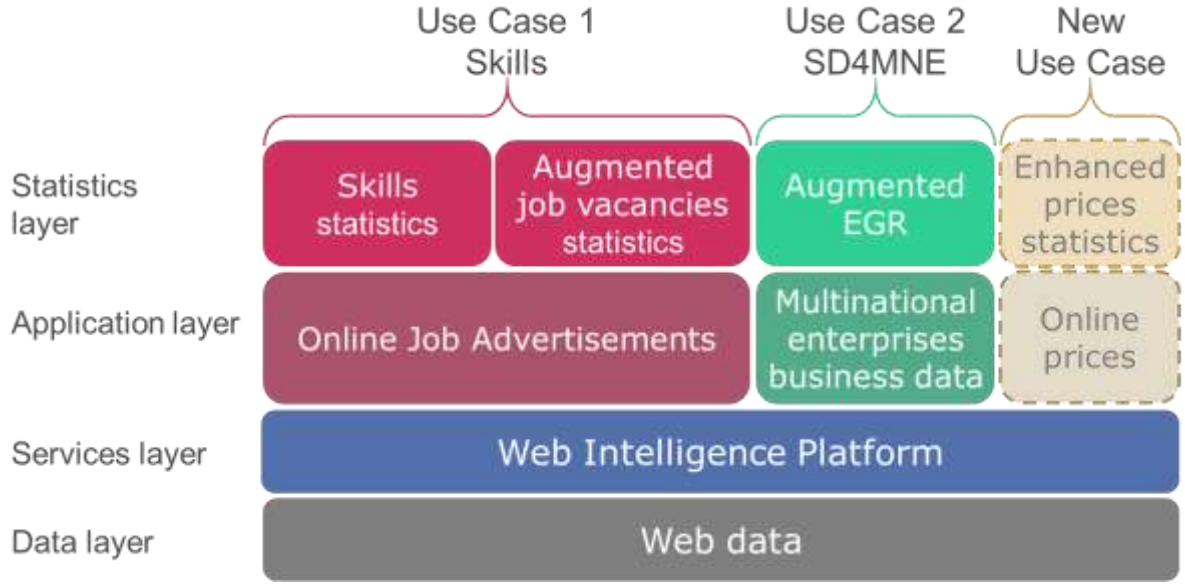

Figure 1. Layered architecture of the WIH

These use cases can be seen themselves as composed of two layers. The first one consists of the application of particular services offered by the platform to the sources of Web data selected in order to produce microdatasets ready to be used to produce statistics. The second and last layer consists of the statistics produced based on those datasets. The particular statistical outputs are instrumental in the creation of the use cases. They guide the development of the use cases by providing specific needs in terms of quality and definitions. However, afterwards, the datasets resulting from the application layer become available for the production of additional statistics not necessarily foreseen initially.

The data layer is obviously a fundamental part of the WIH, as without the Web data the upper layers are irrelevant to produce statistics. However, not only the Web data needs to be available, but also its quality has great influence on the quality of the statistical outputs. Availability and quality are properties that depend greatly on the data space on which the Web sources are involved.

6. The OJA data space

The OJA landscaping run in the context of the OJA system developed by Cedefop identified around 500 sources of OJA in the EU. These sources include job search portals, websites of recruitment agencies, public employment offices, websites of companies and other. It is a rather diverse data space.

Job portals play several important roles in this data space. Firstly, they are the collectors of the OJAs and promoters of the online publishing of job advertisements. They contribute this way for having a labour market panorama on the Web that is representative of the overall labour market. Secondly, they contribute to the harmonisation of the several job advertisements made available on the Web. Without the structure required to provide these advertisements in a platform it would be more difficult to re-use this data for statistical and analytical purposes. Thirdly, they are the curators of the quality of those job advertisements in particular when they are in direct contact with the advertisers and potential employers.

Official statistics producers can also play an important role in this data space. Official statistics offices enjoy of an institutional setup providing them with statistical authority (can legally require the provision of data for statistical purposes), independence and the tools to protect the statistical confidentiality of the data used to produce official statistics. This institutional setup puts these organisations is a privileged position to act as an independent third party in a data space with many players. Statistical offices have the responsibility to produce statistics in almost all domains of people's lives and produces a very large mix of statistics and have access to many data sources. This puts them in a privileged position to act as data integrators opening up the possibility to increase the quality of the statistics produced with a single data source, including those based on OJA data. Finally, official statistics producers have the responsibility to set standards, from definitions to methodologies, guaranteeing the comparability with the remaining statistics produced and can play a role as a standard setter also in the OJA data space.

Both job portals and official statistics producers have the possibility to collaborate to contribute to a data space providing data of higher quality for the benefit of those in the labour market and for society in general.

References:

1. European Statistical System Committee. (2018). Bucharest memorandum on Official Statistics in a datafied society (Trusted Smart Statistics). DGINS 2018.

2. Ricciato, F., Wirthmann, A., Giannakouris, K., & Skaliotis, M. (2019). Trusted smart statistics: Motivations and principles. Statistical Journal of the IAOS, 35(4), 589-603.
3. Ricciato, F., Wirthmann, A., & Hahn, M. (2020). Trusted Smart Statistics: How new data will change official statistics. Data & Policy, 2.
4. Kühnemann, H. (2021). Anwendungen des Web Scraping in der amtlichen Statistik. AStA Wirtschafts-und Sozialstatistisches Archiv, 15(1), 5-25.
5. Descy, P., Kvetan, V., Wirthmann, A., & Reis, F. (2019). Towards a shared infrastructure for online job advertisement data. Statistical Journal of the IAOS, 35(4), 669-675.
6. ESSnet Big Data I (2017): Work Package 1 Web Scraping / Job Vacancies. Final technical report (SGA 1), https://ec.europa.eu/eurostat/cros/sites/default/files/WP1_Deliverable_1.3_Final_technical_report.pdf
7. ESSnet Big Data I (2018): Work Package 1 Web Scraping / Job Vacancies, Final technical report (SGA 2), https://ec.europa.eu/eurostat/cros/sites/default/files/SGA2_WP1_Deliverable_2_2_main_report_with_annexes_final.pdf
8. ESSnet Big Data II (2021): Work Package B Online Job Vacancies. Methodological framework for processing online job adverts data for official statistics, https://ec.europa.eu/eurostat/cros/sites/default/files/ESSNet_II_WPB_OJV_Methodological_framework_V2.pdf
9. ESSnet Big Data II (2021): Work Package B Online Job Vacancies. Report on the statistical output, required quality and definition of the necessary metadata at European and national level, https://ec.europa.eu/eurostat/cros/sites/default/files/DRAFT_ESSNet_II_WPB_OJV_Report_on_output.pdf
10. ESSnet Big Data II (2021): Work Package B Online Job Vacancies. Technical report on the implementation requirements of prototypes in the relevant statistical production processes at European and national level, https://ec.europa.eu/eurostat/cros/sites/default/files/ESSNet_II_WPB_OJV_Report_on_software.pdf