Big Data in Official Statistics: Estimation of job vacancies by using web scraping techniques

Martina Rengers[1]

[1] *Federal Statistical Office, Wiesbaden, Germany, martina.rengers@destatis.de*

Abstract
In many definitions Big Data is characterized at least by 3 V's: high volume, velocity and variety. In the last years, Big Data has become more and more of high interest; this extraordinary interest or possible hype also affects official statistics. The relevance of Big Data for the European Statistical System (ESS) is examined within an ESSnet action on Big Data. Different pilot studies should explore the opportunities and challenges of selected big data sources. One of them is the pilot study regarding web scraping job vacancies. The aim is to explore the possibility of this data source to complement job vacancy statistics, as well as labour market statistics in general.

Keywords: big data, web scraping, job vacancy, labour market statistics

1. Big Data in Official Statistics

In a solicitation dated 2012 the American National Science Foundation (NSF) state that "big data" refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future. Big Data often consist of semi-structured or unstructured data.

In many other definitions the letter 'V' plays an important role. It starts very early in 2001 with the leading information technology research and advisory company Gardner (former META Group), where the analyst Laney defined data growth challenges and opportunities by using "3Vs": (1) increasing volume (amount of data), (2) velocity (speed of data in and out), and (3) variety (range of data types and sources) (Laney, 2001). Gartner updated its definition in 2012 as follows: "Big Data is high volume, high velocity, and/or high variety information

assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Laney, 2012).

Gartner, and now much of the industry, continue to use this "3Vs" model for describing Big Data. Some other analysts or researchers add supplementary "V"s. For example IBM with the following four V's (4V) namely volume, velocity, variety, and veracity (LaValle et al, 2011), whereas Ramasamy, provide a consolidated 5V Big Data Model Framework: (1) volume, (2) velocity, (3) variety, (4) veracity and (5) value (Ramasamy, 2015, Fig 2.).

In the last years, Big Data has become more and more of high interest. 2012, through the work of the UNECE, there was a request for "a document explaining the issues surrounding the use of big data in the official statistics community". In response to this requirement the report "What does 'Big data' mean for official statistics?" (UNECE, 2013) was prepared. This report outlines the opportunities and challenges that Big Data poses for official statistics.

Various conferences, workshops, and other events during the recent years provide another evidence of the enormous interest of official statistics in Big Data. At the DGINS conference the heads of the European National Statistical Institutes stated in the Scheveningen Memorandum of September 2013 (DGINS, 2013) the relevance of Big Data for the European Statistical System (ESS) and the need for adopting a related action plan.

Such an action plan was endorsed in September 2014 (Eurostat, 2014a); based on this, in May 2015 the ESS committee (ESSC) approved an ESSnet action on Big Data. Within the ESSnet Big Data at least three pilot studies should explore the potential of selected big data sources for producing or contributing to the production of official statistics, with all its attendant challenges.

2. ESSnet Big Data

A consortium of 22 partners, consisting of 20 National Statistical Institutes and 2 Statistical Authorities, including the countries Austria, Belgium, Bulgaria, Denmark, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Romania,

Sweden, Slovenia, Spain, United Kingdom, signed on 18 November 2015 a first so-called Framework Partnership Agreement (FPA) with Eurostat to take part on the Essnet Project Big Data. According to this agreement between the consortium and Eurostat, the project runs from February 2016 to May 2018.

The consortium has subdivided its work into work packages. The main activity of the consortium will be the execution of pilot projects for examining the potential of big data sources for official statistics in the ESS context. There are five pilot studies as independent work packages (WP); these are (WP1) web scraping job vacancies, (WP2) web scraping enterprise characteristics, (WP3) using smart meters for measuring electricity consumption of each building, (WP4) using AIS data (AIS – Automatic Identification System to measure real-time ship positions), (WP5) using mobile phone data (geo-localisation).

Three other work packages are not dealing with pilot projects, but instead with horizontal topics, like using big data sources for early estimates (WP6), using big data sources in multi statistical domains (WP7), general methodological aspects (methodological framework, quality framework, metadata framework, IT infrastructure, skills, partnerships with data providers and scientific community – WP8). Two other work packages WP0 and WP9 deal with the co-ordination of the action and the dissemination of the results, respectively.

The overall objective of the project is to prepare the ESS for integration of big data sources into the production of official statistics. The project has to focus on running pilot projects exploring the potential of selected big data sources for producing or contributing to the production of official statistics. It is the aim of these pilots is to undertake concrete action in the domain of big data and obtain hands-on experience in the use of big data for official statistics.

3. Web scraping job vacancies (work package 1 – WP1)

3.1. Major Tasks and biggest Challenges

Germany is one of six countries which take part in the pilot project "web scraping job vacancies", which is coordinated by the Office for National Statistics of the United Kingdom. The general intention is to explore a mix of sources including job portals, job adverts on enterprise websites, and job vacancy data from third party sources. This pilot should demonstrate by concrete estimates which approaches (techniques, methodology etc.) are most suitable to produce statistical estimates in the domain of job vacancies and under which conditions these approaches can be used in the ESS.

For each above mentioned data source the study consist of four main tasks. Task one address data access, task two data handling, task three methodology for output production and task four – at a later date – future perspectives.

The work package focused in a first step the source of job portals. In this case data can be obtained by directly buying from portal owner(s) or by web scraping. Web scraping is a technique employed to automatically extract (often large amounts of) data from websites.

Regarding the *conceptual* aspects, it will have to be analysed which kind of information is available at different job portals and this information fits the existing statistical standards. Regarding the *technical* implementation, a suitable web scraping tool needs to selected and tested in practice. Additionally, there are a number of *legal* aspects relating to data access, e.g. copyright laws and terms of use of the job portals. Only on the basis of these findings, the quality of the data can be assessed and the potential future role of these data can be outlined. A (non-exhaustive) list of the most important challenges look like follows:

- Legal aspects: Is it legally allowed to do web scraping? Who does the data belong to?
- Internet Security: How can data collection using web scraping be implemented in the IT environment of statistical offices with its high requirements for IT security?

- Assessment of job portals: How many need to be selected and which ones are suitable? How stable is the information that can be obtained from job portals over time?
- Technical implementation: Web scraping or provision of data directly by the portal owners (or third parties)?
- Data quality: Reconciliation with job vacancy statistics from surveys and registers

Due to the fact that there have already been studies and experiences on big data and web scraping in some countries a few of the above mentioned challenges have been examined.

The document Eurostat (2014b), for example, contains an examination on *legal aspects* of the automatic data collection methods from the internet (see pp. 46 of Annex D2). In the case of web scraping job portals the sui generis database right is of particular importance. As Abbott, 2014 describes, that *internet security* problems can be solved with a separate IT net. For example, UK's national statistical institute ONS (Office for National Statistics) installed so-called Innovation Laboratories. "The Innovation Labs are a resource for learning, research and innovation which provide technologies which are not available on the standard ONS secure network. They are stand-alone networks of high specification computers which are not connected to the ONS network, but have full internet access." (Abbott, 2014).

In contrast, the *assessment of job portals* is something that has to be done again and again at regular intervals and for each country specific labour market. The following paragraphs will present a general approach on the assessment of job portals and some results for the German labour market.

3.2. Assessment of job portals

In a first action you can try to find job portals by using web search engines such as Google or Bing (general information about the site search market can be found in Eurostat, 2014b, p. 36 ff. of Annex D2.). This is a rudimentary approach which can be helpful for the assessment and is necessary to get an overview of URLs. The keywords and phrases, that have to be used in the web search engines, are – on one hand – used to provide direct information about job portals and – on the other hand – to find other web sites which maybe have rankings or assessment analysis of job

portals. Relevant keywords and phrases are ‘online job portals’, ‘ranking of online job portals’, ‘assessment of online job portals’ and, respectively, ‘Competitive Recruiting’, ‘Jobcoach’, ‘Jobmarketing’, ‘HR Reporting’, ‘HR recruitment’, ‘trade fair for human resources management’.

As a result of this first action (state: January 2016) there were found some important URLs of other websites which have rankings or assessment analysis of job portals. For Germany worth mentioning are (i) *deutschlandsbestejobportale.de*, (ii) c*rosswater-job-guide.com*, (iii) *online-recruiting.net* and (iv) *jobboersen-im-test.de*.

(i) *deutschlandsbestejobportale.de*
This web site contains ranking lists of job portals or job search machines from 2010-2015. The initiators of the test called “Deutschlands Beste Jobportale" (best job portals of Germany) are ICR, Institute für Competitive Recruiting (competitiverecruiting.de) and the joint project CrossPro Research (crosspro-research.com). The latter project is a corporation project of Cross Water Systems and PROFILO Rating GmbH.

(ii) c*rosswater-job-guide.com*
The web sites c*rosswater-job-guide.com* and *crosswater-systems.com* belong to the company Crosswater Systems. According to the company's own information it is dedicated to assist job searcher by providing web guides on a selected range of topics, including annotated links and web resources, to allow the surfer a pre-selection of his next destination. They provide, among others things an own assessment of online job portals on the sub web site: *jobbörsen-kompass.de.*

(iii) *online-recruiting.net*
This web site contains many free available research results, e.g. description of job portals and job portal rankings. It also names job portal URLs of the following 27 countries: Australia, Austria, Belgium, Bosnia Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, France, Germany, Great Britain, Greece, Hungary, India, Italy, Japan, Jordan, Netherlands, Norway, Poland, Portugal, Romania, Russia, South Africa, Spain, Switzerland.

(iv) *jobboersen-im-test.de*
This web sites is done by an private individual interested in the topics of online job-placements and -recruiting.

According to the information from *deutschlandsbestejobportale.de* in 2015 there were more than **1600** job portals for the German labour market. *Jobboersen-im-test.de*/ („Jobbörsen A-Z“) lists altogether **781** job portals (only classified by „target group, profession or sector“, but not by „kind of job portal“). Of these 781 job portals **34** are ranked among the best job portals *Online-recruiting.net* lists **99** job portals; some of them are not included in the above mentioned 781-list.

Figure 1 Compilation of ranking lists

Nr.	Name der Jobbörse	Zielgruppe	Anzahl der Stellenanzeigen (April 2011)	Reichweite Alexa-Ranking (April 2011)	Anzahl der Stellenanzeigen (April 2010)	Arbeitgeber-Zufriedenheit laut Profilo-Ranking (März 2011) Skala: 7 = sehr gut bis 1 = überhaupt nicht gut	Nutzer-Zufriedenheit laut Crosspro-Research (März 2011) Skala: 1 = sehr gut bis 4 = überhaupt nicht gut	Suchqualität laut Crosspro-Research (März 2011) Skala: 1 = sehr gut bis 4 = überhaupt nicht gut
Nr.	Allgemeine Jobbörsen							
1	Meinestadt.de	Allgemein	428.813	1.571	265.222		1.87	2.07
2	Arbeitsagentur	Allgemein	373.192	2.255	202.697	4.91	2.13	2.26
3	Jobmonitor	Allgemein	364.001	59.139	124.355			
4	Rekruter.de	Allgemein	279.606	87.295	152.423			
5	Arbeit-Regional	Allgemein	267.749	943.057	288.324			
6	Jobinfo24	Allgemein	263.066	643.503	102.681			
7	Gigajob	Allgemein	195.450	14.072	166.995		1.93	2.14
8	Jobomat.de	Allgemein	94.500	88.972	65.405			
9	Monster Deutschland	Allgemein	68.300	4.915	49.800	4.98	1.96	2.17
10	StepStone	Allgemein	55.282	3.379	36.650	5.62	1.73	1.94

Translation of Column Labels:

(1)	Nr.	No.	(1)
(2)	Name der Jobbörse	Name of job portal	(2)
(3)	Zielgruppe	Target group	(3)
(4)	Anzahl der Stellenanzeigen (April 2011)	Number of job vacancies (April 2011)	(4)
(5)	Reichweite Alexa-Ranking (April 2011)	Alexa popularity ranking (April 2011) Note: This ranking is according to the Internet statistics provider Alexa (www.alexa.com), a subsidiary of Amazon.	(5)
(6)	Anzahl der Stellenanzeigen (April 2010)	Number of job vacancies (April 2010)	(6)
(7)	Arbeitgeber-Zufriedenheit laut Profilo Ranking (März 2011) Skala: 7=sehr gut bis 1 = überhaupt nicht gut	Employer satisfaction according to Profilo Ranking (March 2011) Scale: 7= very good to 1 = absolutely not good	(7)
(8)	Nutzer-Zufriedenheit laut Crosspro-Research (März 2011) Skala: 1 = sehr gut bis 4 = überhaupt nicht gut	User satisfaction according to Crosspro-Research (March 2011) Scale: 1 = very good to 4 = absolutely not good	(8)
(9)	Suchqualität laut Crosspro-Research (März 2011) Skala: 1 = sehr gut bis 4 = überhaupt nicht gut	Quality of search according to Crosspro-Research (March 2011) Scale: 1 = very good to 4 = absolutely not good	(9)

Source: personalmagazin 06 / 11

At the end of this action neither the precise number of job portals nor their importance for the job market was really clear. Figure 1 shows ambiguous rankings due to different ranking

criteria and/or different dates of the process. Besides the number of job vacancies, there are four other ranking criteria, namely the Alexa popularity ranking, the employer satisfaction according to Profilo ranking, the user satisfaction according to Crosspro-Research and the quality of search according to Crosspro-Research.

These ambiguous situation leads to the requirement of a second action with the URLs known from the first action. On the basis on the number of job vacancies as indicated in the results of the first action **55** general job portals, job search machines and specialized online job portals were selected and examined in more detail (see the first 14 job portals in the field of general job portals in Figure 2). The job search quality in the sense of accuracy of job vacancy description was of high interest to do an additional assessment. For this aspect search criteria according to input fields for job search (simple search and advanced search) were compared with necessary or desired criteria to classify and describe job vacancies. That is, the number of job vacancies is only the first step in the iterative process of assessing job portals.

Figure 2: A selection of important general job portals; ranked by number of job vacancies

Nummer	Art	Name	Link	Stellen
1	Allgemeine Jobportale	StepStone	https://www.stepstone.de/	59.880
2	Allgemeine Jobportale	Monster	http://www.monster.de/	
3	Allgemeine Jobportale	Meine Stadt.de	http://jobs.meinestadt.de/deutschland/stellen	366.524
4	Allgemeine Jobportale	Rekruter	http://www.rekruter.de/	277.886
5	Allgemeine Jobportale	Bundesagentur für Arbeit	http://jobboerse.arbeitsagentur.de/vamJB/startseite.ht	1.083.929 303.493 Ausbildungsstellen
6	Allgemeine Jobportale	Jobmonitor	http://www.jobmonitor.com/	296.419
7	Allgemeine Jobportale	Kalaydo	http://www.kalaydo.de/jobboerse/	
8	Allgemeine Jobportale	Jobcluster	https://www.jobcluster.de/	43.741
9	Allgemeine Jobportale	Stelleneinzeigen.de	http://www.stellenanzeigen.de/	
10	Allgemeine Jobportale	Süddeutsche Zeitung	http://stellenmarkt.sueddeutsche.de/	8.489
11	Allgemeine Jobportale	LinkedIn	lädt nicht	
12	Allgemeine Jobportale	JobScout24 / Jobs.de	http://www.jobs.de/?cbRecursionCnt=1	105.332
13	Allgemeine Jobportale	JobStairs	https://www.jobstairs.de/	25.900
14	Allgemeine Jobportale	Gigajob	http://de.gigajob.com/index.html	634.735
15	Metasuchmaschine	Icjobs	https://www.jobbörse.com/	2.500.000

Necessary or desired criteria to classify and describe job vacancies are mainly arising from current job vacancy statistics and a logical point that a description of a job vacancy at least should give answers to the questions: What? – When? – Where? – Who? Figure 3 lists ten criteria who give a response. More due to analytical reasons as for the needs of a job searcher, the job vacancies should be broken down by NACE at section level in aggregation level as shown in Figure 4 (see EU (2008), Article 1).

Figure 3: Ten criteria to classify and describe job vacancies

What?	
1	position
2	occupation
3	education (required education of the candidate)
4	type of job (permanent- temporary employment, full time-part time job)
When?	
5	date of the advertised job vacancy
6	data of application deadline
7	date to fill a vacancy
Where	
8	location of the job
Who?	
9	Direct employer or agency
10	Economic activity of the employer (NACE groups)

Figure 4: Required NACE groups (NACE Rev. 2 aggregation level)

NACE Rev. 2 sections	Description
A	Agriculture, forestry and fishing
B, C, D and E	Mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply; water supply, sewerage, waste management and remediation activities
F	Construction
G, H and I	Wholesale and retail trade; repair of motor vehicles and motorcycles; transportation and storage; accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M and N	Professional, scientific and technical activities; administrative and support service activities
O, P and Q	Public administration and defence; compulsory social security; education; human health and social work activities
R and S	Arts, entertainment, recreation and other service activities

4. References

Abbott, Owen (2014), ONS Innovation Laboratories, ONS Methodology Working Paper Series No 1. *http://webarchive.nationalarchives.gov.uk/20160105160709/ http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/ons-working-paper-series/mwp1-ons-innovation-laboratories.pdf; date: 2016-04-27*

DGINS (2013), Scheveningen Memorandum, Big Data and Official Statistics, *http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13*

EU (2008), COMMISSION REGULATION (EC) No 1062/2008 of 28 October 2008 implementing Regulation (EC) No 453/2008 of the European Parliament and of the Council on quarterly statistics on Community job vacancies, as regards seasonal adjustment procedures and quality reports.

Eurostat (2014a), 22nd Meeting of the European Statistical System Committee, Item 8 of agenda, ESS Big Data Action Plan and Roadmap 1.0, Work Programme Objective 11.1, *https://ec.europa.eu/eurostat/cros/sites/crosportal/files/ESSC%20doc%2022_8_2014_EN_Final%20with%20ESSC%20opinion.pdf; date: 2016-04-27*

Eurostat (2014b), Analysis of methodologies for using the Internet for the collection of information society and other statistics, D0.4 Final technical report, *http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D0.4_Final%20technical%20report_20140430_1.pdf; date: 2016-04-27*

Laney, Douglas (2012), "The Importance of 'Big Data': A Definition". Gartner. Retrieved 21 June 2012

Laney, Douglas (2001), "3D Data Management: Controlling Data Volume, Velocity and Variety". Gartner. Retrieved 6 February 2001.

LaValle, S., /Hopkins, M. /Lesser, E. / Shockley/, R. and Kruschwitz, N. (2011), Analytics: The new path to value: How the smartest organizations are embedding analytics to transform insights into action, IBM Global Business Services, 2011.

Ramasamy, Ramachandran (2015): The production of salary profiles of ICT professionals: Moving from structured database to big data analytics, in: Statistical Journal of the IAOS vol. 31, no. 2, pp. 177-191, 2015.

UNECE (2013), What does “big data” mean for official statistics? Report of the High-Level Group for the Modernisation of Statistical Production and Services (HLG). *http://www1.unece.org/stat/platform/display/hlgbas*