Work Platforms

Feasibility Report

Lorenzo Malandri, Fabio Mercorio

Web Intelligence Hub

Feasibility study on work platforms web data retrieval

Deliverable D3 - Work platforms scraping feasibility report

Authors: Lorenzo Malandri, Fabio Mercorio

Approved by: Emilio Colombo, Mario Mezzanzanica

Version: 1

Date of Release: 2022-07-20

Conducted for Eurostat under Specific Contract No 2020.0406

Framework Contract 2020-FWC7-AO-DSL-VKVET-JBRAN-WIH-OJA002/20 between Cedefop and the Universitá Degli Studi di Milano-Bicocca, Burning Glass Europe S.R.L. and GOPA Luxembourg SARL - Towards the European Web Intelligence Hub - European system for collection and analysis of online job advertisement data (WIH-OJA)

Table of Contents

1.0 Scope of the analysis

The report describes the analysis implemented to assess the feasibility of collecting data from Digital Labour Platforms (DLPs) through web scraping. The DLPs analysed have been collected and selected within the activities of the "Deliverable D1 – work platform and scoping report" of the Request for Service No ESTAT04, Specific Contract No2020.0406 "Feasibility study on work platforms web data retrieval".

Table 1 List of selected platforms

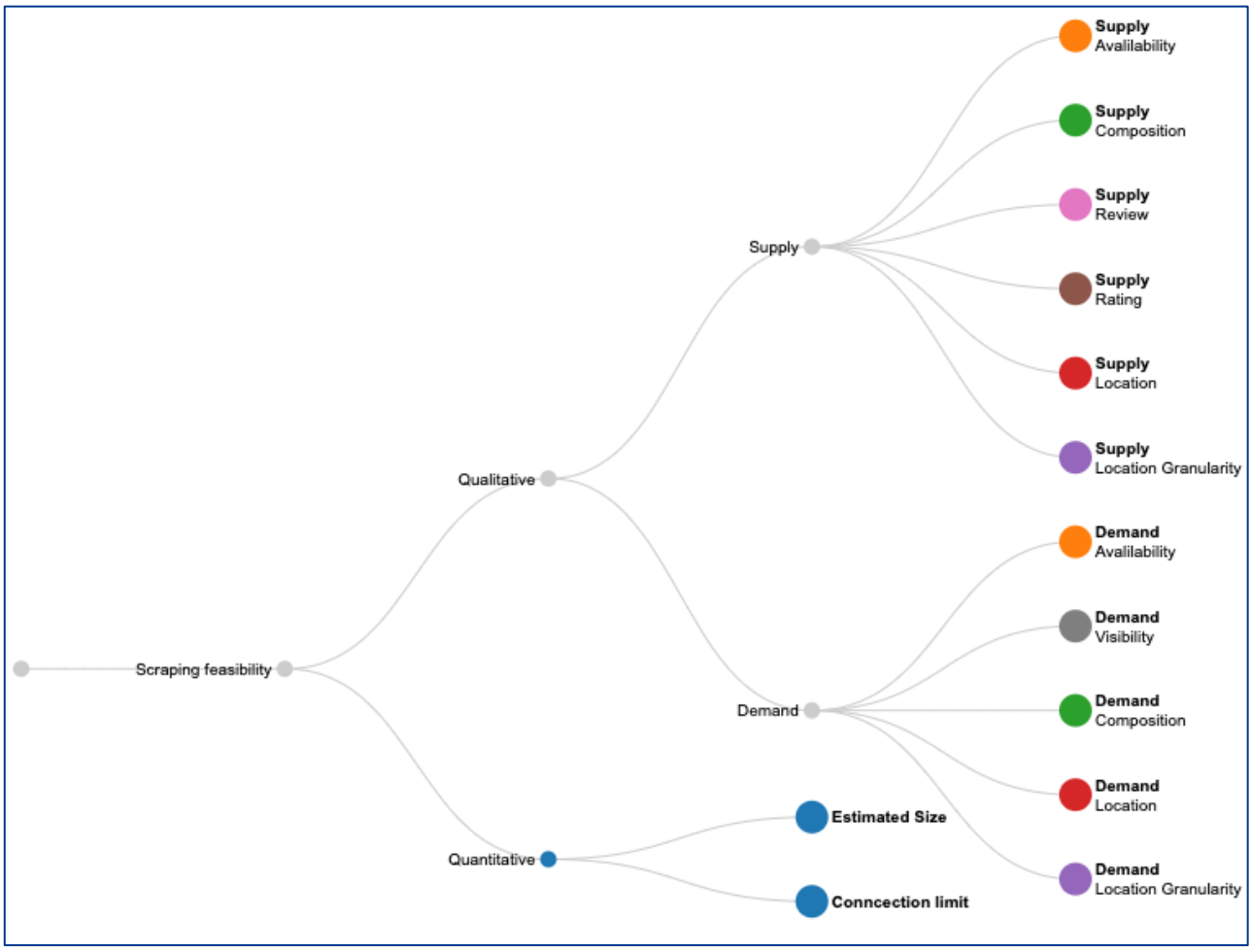
Platform Name	Website	Country	Estimated Size
Cronoshare	https://www.cronoshare.com	ES, IT, USA, CH, AU	336,990
khdemti	https://www.khdemti.com/	Global	223,861
freelancermap	https://www.freelancermap.com	Global	12,000
yoopies	https://yoopies.it	IT, BE, FR, DE, ES	7,744
Go lance	https://golance.com/hire/	DE, ES, FR, IT	na
Care	care.com	BE, UK, FR, CA	33,000,000
Starofservice	https://www.starofservice.it/	ES, FR, BE, IT, ES, DE, BG, PL	9,000,000
upwork	upwork.com	Global	1,800,000
MeineStadt	https://www.jobs.meinestadt.de/	DE	836,683
gojob	https://gojob.com/candidat/	FR	350,000
Taskia	https://www.taskia.es/	ES	253,000
chose your boss	https://www.chooseyourboss.com/	FR	178,000
Freelance-informatique	https://www.freelance-informatique.fr/	FR	114,915
Prontopto	https://www.prontopro.it/	IT	660,000
freelance.com	https://www.freelance.com/	FR	370,000
toptata	https://toptata.it/	IT	330,000
BlauArbeit	https://www.blauarbeit.de/	DE	150,000
Sitterlandia	https://www.sitterlandia.it/	IT	128,869

Different DLPs present different richness of information and issues when trying to automatically get data from them. The scope of this study is to evaluate what information is contained in DLP and how difficult is to get it, both in terms of access policies and policies to block tools that automatically harvest data from the websites.

Therefore, the information content of DLPs will be assessed having in mind the task of information retrieval and the technical issues related to it.

Each DLP will be evaluated according to the qualitative and quantitative criteria displayed in Figure 1 and described below.

Figure 1 Taxonomy of evaluation criteria for scraping feasibility



2.0 Qualitative criteria

DLPs are built to match **demand** and **supply**. For this reason, they usually make available both professional resumes and job descriptions. The qualitative variables are divided accordingly.

2.1 Supply-side

1. **Availability**: Is the professional cv present?
   Open, free registration, match, payment
   When job offers are present, they usually are expressed through the CV or a brief resume of the professional. The resumes might be (i) *open* and accessible to anyone, (ii) available only upon registration (with a free or paid account), (iii) made available by the work-platform that proposes one or more professionals when an application for a job/task is made.
2. **Composition:** How are resumes composed?
   only description, description + skill, description + structured data
   In some cases, the resume might contain structured fields, like the professional skills or competencies owned or additional information such as working time, age, etc.
3. **Review:** is there a qualitative review of the professional?
   yes, no
   Some DLPs provide user reviews for professionals.
4. **Rating:** is there a quantitative rating of the professional (Likert scale or similar)
   yes, no
   In addition, or as alternative to user reviews, some DLPs display a numeric rating for each professional, assigned by previous customers. Those ratings are not uniform among different websites, though they are usually expressed through a 5-values or 10-values Likert scale, thus they are comparable.
5. **Location:** Is the working location mentioned?
   mandatory, optional, possibility of remote working
   In most DLPs, the professionals' resumes contain the city or province where they are based or, in the alternative, if they work remotely. In some cases, professionals can propose themselves as both remote or in-person workers.
6. **Location granularity**: Is it possible to filter professionals by location?
   country, state, province, remote, etc.
   On many yet not all the websites it is possible to select resumes by zone. In other cases, is the platform that automatically proposes professionals areas when applications are made.

Table 2 Availability of the platforms - supply side

Platform_name	Professional resume	Professional skills	Reviews	Rating	loc filter	loc mentioned
Cronoshare	open	no	yes	yes	province	yes/remote
khdemti	open	yes	yes	yes	city	yes/remote
freelancermap	open	yes	no	no	country	yes/remote
yoopies	register	no	no	no	30km	yes
Go lance	open	yes	no	yes	country	yes
Care	register+pay+match	/	/	/	/	/
Starofservice	match by email	/	/	/	/	/
upwork	register	yes	yes	yes	country	yes
MeineStadt	pay	/	/	/	/	/
gojob	no	/	/	/	/	/
Taskia	4 upon request	/	/	/	/	/
chose your boss	no	/	/	/	/	/
Freelance-informatique	open	yes	no	no	region	yes/remote

Prontopto	match (by email)	/	/	/		
freelance.com	available profiles (20k)	yes	yes	no	city	yes
toptata	register	no	yes	yes	50km	yes
BlauArbeit	register	yes	yes	yes	city+distance	yes
sitterlandia	open	yes	no	no	country	yes

2.2 Demand-side

1. Job descriptions: Which job/task descriptions are available?
   Open, free registration, match, payment
   The job descriptions might be (i) *open* and accessible to anyone, (ii) available only upon registration (with a free or paid account) or, in some cases, (iii) jobs/missions are presented by the DLP when a professional professional profile is created. In this last case, it can be directly provided on the website or by email/SMS.
2. Visibility: Are all the job descriptions visible?
   All, in a distance range
   When job offers are present, they usually are provided through job descriptions. In this context, they are also called missions or projects, to indicate that usually are single tasks or services limited in time. Depending on the DLP, the user might have visibility on the total amount of tasks/jobs on the website or just the ones in a certain area around its position.
3. Composition: How are job descriptions composed?
   only description, description + skill, description + structured data
   In some cases, the job description might contain structured fields, like the skills required or additional information on the task/project, like working time, age, etc.
4. Location: Is the working location mentioned in the description?
   mandatory, optional, possibility of remote working
   In most DLP, the job description contains the city or region where they are based or, alternatively, if the work can be done remotely. In some cases, professionals can propose themselves as both remote and in-person workers.
5. Location granularity: Is it possible to filter job descriptions by location?
   country, state, region, remote, etc.
   On several DLP it is possible to select job descriptions by area. In other cases, is the platform that automatically displays professionals in the area when a profile is created.

Table 3 Availability of the platforms - demand side

Platform_name	job description	description visibility	composition	loc filter2	loc mentioned2
Cronoshare	all	open	description+structured	state	yes/remote
khdemti	all	open	description+skills	no	sometimes+remote
freelancermap	all	open	description+skills	country	yes/remote
yoopies	30km from you	register	description	no	yes
Go lance	no	/	/	country	yes
Care	match, 50 miles from you	register+pay	/	/	/
Starofservice	all	register+pay	/	/	/
upwork	all	register	description+skills+structured	country	yes

MeineStadt	all	open	description+structured	city	yes
gojob	all	open	description+qualifications	city	yes
Taskia	all	open	description	yes	yes
chose your boss	all	open	description+skills	yes	yes/remote
Freelance-informatique	all	register	description+skills	region	yes/remote
Prontopto	by email	/	/	/	/
freelance.com	match (by email)	/	/	/	/
toptata	match, 50km from you	register	description + structured	your zone	yes
BlauArbeit	all	register	description	city+distance	yes
sitterlandia	match (by email)	register	description	your zone	yes

3.0 Quantitative criteria

These refer to the use of quantitative measures or tests to assess automatic information retrieval from DLPs.

Estimated size

The first quantitative measure is the estimated size of DLPs (i.e. the number of tasks/jobs posted) collected by experts during the landscaping phase

Loading and connection limits.

Web scraping refers to a method for automatically extracting data from web pages by means of ad-hoc tools. A web scraper can be used manually by a user, but when the number of pages to be harvested increases, it is necessary to use a specific software (or bot), called *web crawler*, that automatically visits all or a part of the pages of a website and applies the scraper to all of them. Making several attempts in a short period of time, the web crawler can be blocked by some websites. Therefore, it is important to understand how many requests can be made by the software before is being blocked. In order to assess this, for all the DLPs identified above we constructed a bot that automatically visits the pages of the target websites, increasing the number of pages visited per second until the bot is blocked.

In particular, two tests have been applied.

The first one (Loading test) starts from one page visited per second and with every step visit one more page of the website. This test is called the incremental RPS test, where RPS is for responses per second, and defines the maximum number of requests per second that the website allows from a single connection. The results of this test are presented in the white-background plots in the examples below. In the plots, the green line represents the number of requests (page visited) per second and the red one the number of requests that failed.

The result of the loading test is then employed for the second test, the connection limit test which estimates the maximum number of pages that can be visited before being blocked. In this test, estimated the maximum number of requests per second obtained using a test of 20 minutes.

In Table. 4 we present the maximum number of pages that can be visited before being blocked for each DLP, in the column *pages visited per second*. When the value is not applicable (na), it means that the connection has been blocked immediately because the website detects immediately and blocks bot agents. In the first two columns are reported the name and the total number of job descriptions for each DLP. In the column *% of failures*, we reported the ratio of failures over the number of requests in a timespan of 20 minutes. The fact that those numbers are low for almost all the websites suggests that many of them block bots visiting several pages in a small time, but if this number is lowered, the requests may be repeated over time without incurring in failures. The only exceptions are *Starofservice*, which allows 24.2 requests per second but after 30 seconds blocks about half of them, and *chose your boss*, which after one minute accepts only 0.8 requests per second. Moreover, is worth noticing that *khdemti* allows only up to 0.3 requests per second from a single machine. Following the time allowed for scraping before being blocked and the estimated size of the DLP, in the last column, we calculated how many days it would take to visit all the pages of the website, assuming the same performances of the simulation implemented.

Table 4 Analysis of the automatic collection of web pages from DLPs

Platform name	Estimated size	Pages visted per second	% failures	Pages visited (20 min)	Days to scrape all the website
Cronoshare	336,990	3.21	0.05%	3,855	1.22
khdemti	223,861	0.33	0.00%	391	8.64
freelancermap	12,000	na	na	na	na
yoopies	7,744	5.7	0.63%	6,800	0.02
Go lance		127.82	0.00%	153,387	
Care	33,000,000	19.03	0.00%	22,841	20.10
Starofservice	9,000,000	24.18	49.97%	14,512	2.15
upwork	1,800,000	na	na	na	na
MeineStadt	836,683	na	na	na	na
gojob	350,000	42.52	0.00%	51,017	0.10
Taskia	253,000	2.12	0.00%	2,550	1.39
chose your boss	178,000	9.67	92.16%	910	0.02
Freelance-informatique	114,915	34.48	0.00%	41,382	0.04
Prontopro	660,000	17.86	0.17%	21,395	0.43
freelance.com	370,000	45.41	0.00%	54,512	0.09
toptata	330,000	164.04	0.00%	196,814	0.02
BlauArbeit	150,000	47.72	0.00%	57,278	0.04
sitterlandia	128,869	14.7	0.00%	17,642	0.10

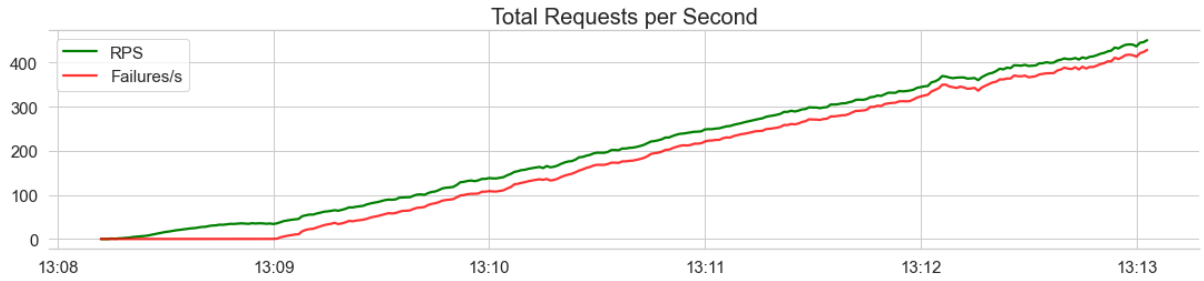
4.0 Some useful cases

Below, we report the loading test for some of the platforms, to show the tendency of the failures for some representative cases.

4.1 Case 1: all the requests below a certain threshold are successful

For **sitterlandia**, after reaching about 16 requests per second, the system starts refusing all the additional ones. The behaviour of the website is the same increasing the number of requests.

Figure 2 Case 1: all the requests below a certain threshold are successful

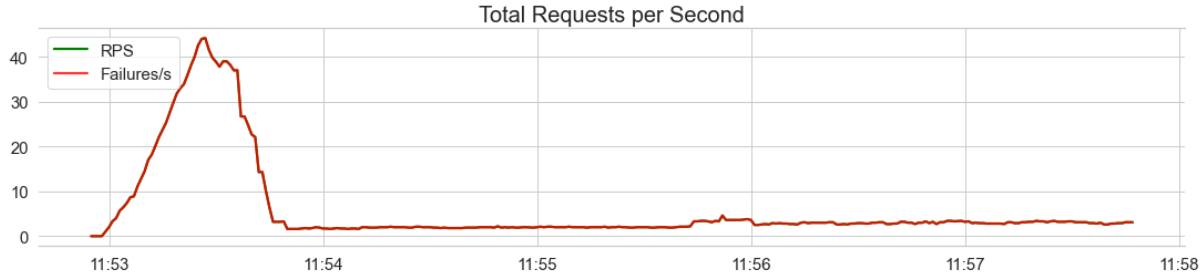


As shown in Table 4, using a fixed number of 14.7 requests per second, all the pages are visited without any error.

4.2 Case 2: no request can be made

For instance, for **freelancemap** every single request returns the error "*403 forbidden*".

Figure 3 Case 2: no request can be made

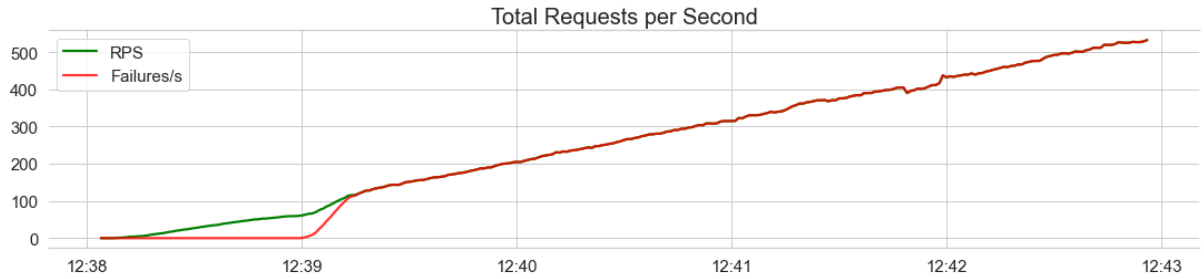


In this case, we did not pursue the connection limit test since no request has been successful.

4.3 Case 3: after some time, the system start blocking the requests over a certain threshold

For instance, for choose your boss, the RPS tests show the platform blocking the agent only when it visits more than 100 pages per second. However, this behaviour of the website is not the same over time.

Figure 4 Case 3: after some time, the system start blocking the requests over a certain threshold



Visiting about 10 pages per second, the website allows it for a period of one minute, after which it limits the number of requests to 0.8 per second, as presented in Table 4.

5.0 Contacting DLPs for agreements

Albeit feasible, automatic scraping of DLP contents is certainly not the first best option for data quality. Data quality is generally maximised by direct access to the data usually provided through API. These are usually the outcome of agreements undertaken with websites/data providers.

In order to explore this possibility, the research team selected a list of the most prominent DLPs (chosen considering the size and country coverage) to explore formal agreements.

ICEs have been asked to provide contact detail for those websites, giving precedence to direct contacts.

In most cases the contact has been provided either through email or through a LinkdIn contact. The message contained a formal letter explaining the scope of the exercise and the use of the data.

In the annex we provide both the agreement letter and the Excel file which lists the platforms contacted and the contact mean used.

DLPs have been contacted twice, the first at the end of June and then in a follow-up message two weeks later. In most cases there has been no answer, probably because contact access points are too general and do not allow the message to reach the right person in the organisation of the DLP.

6.0 Conclusions

In principle the technical analysis conducted on the DLPs shows that automatic scraping is feasible, albeit with some restrictions and constraints. However, the fact that several DLP are characterised by filters that select the availability and visibility of tasks/jobs and of professionals to the location and to the login preferences.

In practice the possibility of having access to complete information contained in DLPs rests on the possibility of striking formal agreements with them that would provide the data directly instead of indirectly through automated scraping activities.