**Developing software for web scraping: the Italian experience on portals offering tourist accommodation**

**Keywords:** Web Scraping, Experimental Statistics, Tourism Statistics, Accommodation establishments

# Introduction

This work stems from the experience conducted by Italy in the context of the Essnet - Big Data II project. The overall objective of the ESSnet Big Data II is to prepare the ESS for the integration of big data sources into the production of official statistics. The project involves 23 NSIs as partners, plus five other ESS members, and includes new pilot projects, identified by five different work-packages. This paper relates to the work done by Italy for WPJ (Innovative Tourism Statistics), whose main objective is to address the need for a conceptual framework and to set up a smart pilot Tourism Information System to support statistical production in the field of tourism. In this context, IT is extremely important with web scraping as one of the fundamental methods of obtaining data from web portals [1]. As part of the WPJ, Italy had the opportunity to experiment, for the first time, web scraping on some of the portals offering accommodations for tourism purposes, specifically Hotels.com and Booking.com. The National Institute of Statistics (Istat) had already developed web scraping procedures to be used in other sectors, but never in the field of tourism. The web scraping had the objective of downloading from the portals both information on tourist facilities and their availability for certain specific dates, to compare them with the information derived from the census surveys on “Capacity of collective accommodation establishments” and “Occupancy in collective Accommodation establishments”, conducted by Istat.

The combining of data provided from the traditional sources and those provided from web scraping has two main goals. Firstly, completing and enriching information on the tourist accommodation establishments already surveyed in the context of the traditional statistical surveys; this means an integration to the official statistics required by the European Regulation on tourism statistics [1]. Furthermore, understanding the degree of coverage of these surveys, particularly in the case of those typologies of accommodation not well or not yet covered by the official statistics on tourism.

After an initial attempt on the Hotels.com portal, web scraping was performed on the Booking.com portal, as it was verified that it contains more accommodation establishments on the Italian territory[[1]](#footnote-1). The first web scraping was carried out in November 2019 at the national level. The high number of accommodation establishments in the country required limiting the search to a region considered particularly interesting from a tourist point of view and which shows a high degree of reliability in the statistical coverage of the census survey: Emilia-Romagna. Web scraping was performed in two different periods: in November 2019 and in May 2020.

# Methods

The development of the software started from the reuse and adaptation of the web scraping programs already implemented in Istat. In particular, the experience of some previous projects was a good starting-point: web scraping and text mining to integrate and validate the information of the ASIA statistical register of active companies [3], web scraping and job application (Online Job Vacancies) and web scraping in the context of the survey of consumer prices for international and national holiday packages.

The process includes three main steps:

1. ***Design and development of the java application for the acquisition of the accommodation*** ***establishment information.***

The Java programming language was used with the extension of the Selenium web scraping libraries within the Netbeans IDE development suite. The Selenium library associated with the drivers of the Internet browser allows the use of the browser as if it were a puppet. The use of a real browser allows navigation, without any limitation, of any site, even those that run Javascript code.

1. ***Inspection and analysis of the Booking's site***

The web development tools installed as an extension in the Firefox browser were used to carry out the preventive analysis of the Booking.com site. Initially, a single municipality as a search parameter was set. The results page was studied for the optimization of the extraction of the available hotel information.

To get all the facilities listed on the portal, the number of people and the start and end dates of the stay were omitted in the search criteria. In this way, all the rooms with their type, services, and direct connection to them can be collected. The software executes two scans of the portal simultaneously: one for the facilities and rooms and the other for their availability and price. The frequency of the two scans is planned to be different (the former can be much less frequent than the latter).

To exceed the limit imposed by the site, which displays up to a maximum of 1,000 listings, an attempt was made to use a different paging algorithm by increasing the offset, but without success. Booking shows only the first 1,000 listings regardless of the offset set and continues to show, on subsequent pages, the same listings obtained on the last page. The final solution was to set more selective criteria, such as the type of accommodation establishment and the number of stars, thus obtaining all available results.

To limit the possibility of being intercepted as a bot, random delays were inserted to simulated human interaction. Despite these attentions, a popup window occasionally appears indicating whether the interaction is taking place through a human user. This obstacle was solved with the implementation of specific programming code.

1. ***Generation of output files***

The application produces three output files with different contents:

1. information on the accommodation establishments: internal identification to Booking, internal identification to the database, scan date, detail URL, the HTML content of the page, accommodation establishment name, type of accommodation establishment, address, latitude and longitude Atlas, extended description and services.

2. information on the rooms of the accommodation establishment: room identification inside Booking, scan date, type of room, size, description of the room, dimensions in square meters, occupancy, services.

3. information on availability and price for the number of people and the specified dates: room ID inside Booking, scan date, current price, tourist taxes, breakfast price, breakfast included option, availability.

In addition to the output files, the application stores the data to a NoSQL database (without schema) MongoDB to carry out more in-depth analytical reports.

# Results

The results of web scraping were compared with the microdata contained in the register of the accommodation establishments sent by the Emilia-Romagna region in compliance with the mandatory transmission of microdata on accommodation establishments started from 2018, concerning all Italian regions. The Emilia-Romagna register has a good quality in terms of exhaustiveness and completeness of the information contained, as there is a complete coincidence of its contents with the results of the Capacity survey carried out by Istat.

Table 1 shows the listings extracted from web scraping carried out on Booking.com compared to the accommodations of the Emilia-Romagna region in 2019. The number of listings found by web scraping (6,419) corresponds to about half of the accommodation establishments contained in the regional archive (49.9%).

As shown, there is a critical issue regarding the different strings of the names of the accommodation establishments, which must be addressed to understand the outcome of the extraction from Booking. The variety of denominations present in the regional archive is vast, however, it probably responds to administrative classification needs. A similar great variety is present on Booking, which however shows names that are probably more oriented to attract tourists, therefore more characterized and more descriptive of the type of accommodation. Some denominations are the same between the two sources (e.g Hotel, bed & breakfast, camping grounds, etc.), however some categories, for example, other private accommodations, do not appear at all in the Booking listings for the Emilia-Romagna region, but probably they are included in other types.

On the other hand, accommodations like alpine refuges and hiking refuges are not covered at all and it seems difficult that they can be traced in other categories. A series of types appear only on Booking. Among them, chalets, boats, inns, lodges, motels, resorts, villas, homestay accommodations, and only a more in-depth analysis, based on a match or link procedures between the accommodations coming from the two sources, will be able to ascertain whether they are included in other denominations.

Based on the identified categories, it emerges that the potential coverage of Booking on the Emilia-Romagna region at the time of the scraping concerned all the touristic villages (100%) and almost all the rented houses (78.7%), and a high percentage of residences (67.7%). Instead, the less visible categories on Booking in the Emilia-Romagna region were campsites (13.0%), hostels (24.3%), and rented rooms (29.9%).

Table 1. Listings on Booking.com and accommodations of the Emilia-Romagna   
(E-R) regional archive (2019)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type of accommodation on E-R regional archive** | **Type of**  **accommodation on Booking.com** | **SOURCE** | | **Difference (absolute values)** | **% potential coverage** |
| **E-R archive** | **Booking.com** |
| **Nr of accommodations** | **Nr of listings** |
| Hotel | Hotel | 3,982 | 1,810 | 2,172 | 45.5 |
| Rented house | Rented house | 2,917 | 2,296 | 621 | 78.7 |
| Bed & Breakfast | Bed & Breakfast | 2,349 | 948 | 1,401 | 40.4 |
| Rented room | Rented room | 1,115 | 333 | 782 | 29.9 |
| Farmhouse | Farmhouse | 791 | 244 | 547 | 30.8 |
| Holiday house | Holiday house + Country house | 846 | 354 | 492 | 41.8 |
| Other private accommodation (not enterprise) | Not present | 349 | - | 349 | - |
| Residence | Residence | 294 | 199 | 95 | 67.7 |
| Camping sites | Camping sites | 100 | 13 | 87 | 13.0 |
| Hostel | Hostel | 70 | 17 | 53 | 24.3 |
| Alpine refuge | Not present | 22 | - | 22 | - |
| Hiking refuge | Not present | 11 | - | 11 | - |
| Holiday village | Holiday village | 11 | 11 | 0 | 100.0 |
| Others | Not present | 5 | - | 5 | - |
| Not present | Chalet, Boat, Inn, Lodge, Motel,  Resort, Villa,  Homestay accommodation | - | 194 | - | - |
| **TOTAL** |  | **12, 862** | **6,419** | **6,637** | **49.9** |

Source: Emilia-Romagna regional archive and Booking.com

# Conclusions

Although the timeframe of the ESSnet is too short to have full implementation of new statistical processes, the development of production prototypes by NSI has allowed us to share knowledge and to have earlier research results. For Italy, it was very important to build and consolidate a standard web scraping method, as all portals have more or less the same problems, and the solutions can be replicated. Nevertheless, a great issue is the frequency of updates of the portals, which frequently change their layout (CSS style) with the consequence of a continuous need for customization of the software. Another relevant issue is the significant initial effort for setting all the specific search parameters to cover the entire territory. The preliminary identification and specification of all provinces and municipalities to be searched is essential to be almost sure that some of the available accommodation data are not missing in the results. Now, our main challenge is to move from experimentation to production, by populating incrementally the database for the Emilia-Romagna region and then for the whole country. Also, other ways to obtain data at the micro-level could be tried with the portals, like for example the use of APIs that Booking makes available for its partners to present the accommodations on its site.

# References

1. WPJ (2020), [Deliverable J1: EssNet Methods for webscraping data processing and analyses](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4f/WPJ_Deliverable_J1_ESSnet_Methods_for_webscraping_data_processing_and_analyses_2019_07_23.pdf), 23 July 2019
2. Regulation (EU) no 692/2011 of the European Parliament and of the Council of 6 July 2011 concerning European statistics on tourism <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:192:0017:0032:EN:PDF>
3. <https://www.istat.it/en/experimental-statistics>

1. <https://www.similarweb.com/top-websites/Italy/category/travel-and-tourism> [↑](#footnote-ref-1)