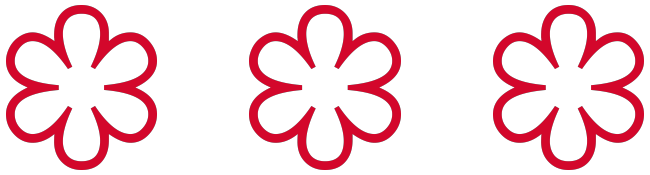
French Michelin-Starred Restaurants and Chefs Databases

Mathilde Kretz, Alexandre Ngau, Thomas Boudras[1]

[1]PSL Research University - Data Acquisition, Extraction and Storage Course Project

January 2024



1 Introduction

This project focuses on the acquisition, extraction, restructuring and cleaning of data, with the main aim of creating a meticulously organised and refined database.

With this in mind, we focused on the topic of French Michelin-starred restaurants. The aim was to create a database providing information related to Michelin-starred restaurants on the one hand, and chefs on the other. In order to restrict the size of the datasets, the acquired data has been limited to French Michelin-starred restaurants. To build this database, we used the Michelin Guide website, as well as Wikidata's SPARQL query service and Wikipedia, as data sources.

2 Acquiring the Data

In this section, we will detail how we acquired the data from the data sources to create our database. We will describe the method we used and the problems we encountered.

2.1 Scraping the Michelin Guide Website

After making sure the scraping of the website was authorized by referring to the `robots.txt` file [1], we created a web-scraping spider using the Scrapy library in Python, and applied to its settings the basic ethics of web-scraping (scraping delay, etc.).

At first, the goal was to scrape all the search pages (the 31 of them) containing the French Michelin-starred restaurants [2] (there were 617 of them as of January 3rd, 2024) and then follow each link to each restaurant's page to scrape the information needed to build the dataset.

However, upon scraping the html of the 31 pages in question, and storing them locally, we discovered that each restaurant not only had a link to its own page on the website, but also had a full set of information describing it, among which we extracted the following keys :

- restaurant-name
- restaurant-region
- restaurant-city
- restaurant-distinction
- restaurant-chef
- restaurant-cooking-type
- restaurant-menu-price

This discovery alleviated the difficulty of following links to scrape the desired information. The spider's parsing method was then modified using CSS selectors, in order to save a collection of 31 dictionary elements, representing the 31 scraped search pages, into a JSON file named `restaurant_scraped.json`. Each dictionary element comprised seven lists, corresponding to the seven above-mentioned keys.

In the end, running the `preprocess_json.py` script formatted the contents of the `restaurant_scraped.json` file into the `restaurant_cleaned.json` file. This last file housed a list of 617 dictionary elements, each representing a restaurant, and containing the seven keys mentioned earlier.

Please find the code and the instructions to reproduce the Michelin Guide's French-starred restaurants dataset in the `michelin_guide_webscraping` subdirectory of the associated GitHub [3].

[1] `https://guide.michelin.com/robots.txt`
[2] `https://guide.michelin.com/fr/fr/selection/france/restaurants/restaurants-etoiles`
[3] `https://github.com/alexandrengau/French_Michelin_Starred_Restaurant_Database/tree/main`

2.2 Scraping Wikipedia

Unfortunately, in the Michelin Guide's French-starred restaurants dataset created in paragraph 2.1, some restaurants were missing their chef's names. These values, when missing, were missing in the html of the scraped search pages, as well as in the html of the restaurant's page, but could sometimes be found in the restaurant's description on the latter page, among other unrelated names.

To tackle this issue, the first idea we had consisted in following each link in the html of the scraped search pages that lead to a restaurant page, and scrape their description. Then, each description would be analyzed using NER (Name Entity Recognition), and the chef's name would then be added when missing. Sadly, the chef's name was sometimes never mentioned, or mentioned with other unrelated chefs names.

For that matter, this solution was not carried on with, and instead we decided to scrape the Wikipedia page hosting the lists of all the two and three Michelin-starred restaurants [4], in order to use the data collected to fill in the missing values in the dataset sourced from the Michelin Guide website. Unfortunately, the chosen solution would not be enough to fill in the missing chef names for the one-starred French restaurants.

As in paragraph 2.1, we made sure the scraping of Wikipedia was authorized by referring to the `robots.txt` file [5], then created a web-scraping spider.

The difficulty of the scraping did not reside in the number of pages to scrape – because there was only one – but in the parsing of its html. Indeed, the html tables containing the desired information did not have the same structure from top to bottom (e.g. sometimes a restaurant would have its name between <td> elements, and sometimes between <td><a> elements). That is why the spider's parsing method was updated using CSS and XPATH selectors, and various other methods to be able to deal with the changing html table formatting.

Thanks to this intricate parsing method, the spider scraped, and saved – in a fairly processed fashion – the data in a JSON file named `two_three_stars_restaurant_scraped.json`. This JSON file contained a collection of 104 dictionary elements, representing the 104 two or three Michelin-starred restaurants in France, each of which had the following key information :

- restaurant-name
- restaurant-region
- restaurant-distinction
- restaurant-city
- restaurant-chef

The `preprocess_json.py` script, contrary to paragraph 2.1, did not have a formatting action when run, but rather an aesthetic and simplifying action, in order to obtain the `two_three_stars_restaurant_cleaned.json` file.

Please find the code and the instructions to reproduce Wikipedia's Two and Three French Michelin-starred restaurants dataset in the `wikipedia_michelin_stars` subdirectory of the associated GitHub.

2.3 Querying Wikidata

After gathering a decent amount of information about Michelin-starred restaurants, we opted to focus on the chefs and owners of these establishments. Wikipedia served as a valuable source, but due to inconsistencies in information across pages, creating a dataset directly from the site proved challenging. Conversely, certain standardized details, found in a banner on the page, are sourced from the Wikidata database. This database is accessible and can be utilized through its query platform [6].

Wikidata is well-organized, and we can use SPARQL queries to access different details about instances saved in the database. We decided to include chefs from all around the world, regardless of their nationality, to account for foreign chefs working in France. In the end, we put together a table with – when available – the following information about each chef:

- chefLablel (for their name)
- countryLablel (for their country of citizenship)

[4] `https://fr.wikipedia.org/wiki/Liste_des_restaurants_deux_et_trois_toiles_du_Guide_Michelin`
[5] `https://fr.wikipedia.org/robots.txt`
[6] `https://query.wikidata.org`

- birthdate

- deathdate

- distinctionLable (eg. "Chevalier de la légion d'honneur")

Please find the code and the instructions to reproduce the queried Wikidata chefs dataset in the `wikidata_chefs` subdirectory of the associated GitHub.

3 Structuring and Cleaning the Database

Once we had these three datasets at our disposal, we needed to build two well-organized, high-quality tables to meet our initial expectations. From the original JSON files, we built two `pandas.DataFrame` tables on which various queries could then be performed.

Please find the final tables of the database in the `final_tables` subdirectory of the associated GitHub.

3.1 The restaurants table

Information for Michelin-starred restaurants is sourced from two places: the Michelin Guide website and the Wikipedia page that lists two and three-starred restaurants. Our approach was to consider the information from the Michelin Guide as the primary source (given its role in rating the restaurants) while supplementing it with information from Wikipedia whenever possible. To ensure consistency and prevent spelling confusions, we standardized all strings by converting them to lowercase and removing accents and special characters.

In practice, we transformed the database extracted from the Michelin Guide website into a dictionary in which the restaurant names served as keys, each of which referring to another dictionary element with the following attributes: `region`, `city`, `distinction`, `cooking-type`, `menu-price`, `current-chef`, `starred-chef`, `current-chef-id`, `starred-chef-id`.

The keys `region, city, distinction, cooking-type, menu-price`, and `current-chef` have been populated with data obtained from the database extracted from the Michelin Guide website, as this data constituted the originally scraped information. We have also chosen to treat them as empirically perfect data. This implies that any alternative data related to the same restaurant, containing different information, has not been considered. Additionally, any restaurant not present in the original Michelin Guide website database has not been regarded as starred.

The constructed table was then supplemented with data sourced from the Wikipedia page. The key information differed significantly from that of the Michelin Guide. We were particularly interested in the chefs, whose names were sometimes absent in the Michelin database. However, the chefs listed in the Wikipedia database were the ones who *originally* received the stars for the restaurant, not necessarily the *current* chef overseeing the restaurant's kitchens. For instance, the restaurant "Le Louis XV - Alain Ducasse à l'Hôtel de Paris" received three stars under the direction of Alain Ducasse, as mentioned by Wikipedia, but it is managed by Dominique Lory, as stated by the Michelin Guide. This is why we decided to introduce the key `starred-chef`, corresponding to the chef who initially received the stars for the restaurant.

To establish a connection between both sets of information, we needed to identify a restaurant from Wikipedia as one listed in the Michelin Guide (and consequently, in our database). The restaurant names serve as keys, supposed to be unique identifiers; however, this is not the case in practice. For example, the names "Le Meurice" and "Restaurant le Meurice Alain Ducasse" both refer to the same restaurant but are keys in the two databases. Using a RegEx search, we were able to compare whether a string is present within another, considering it *as a complete word*. This was done to avoid, for instance, recognizing the restaurant "Pic" within the name "Jean Sulpice." In addition to this initial recognition, we decided to impose an additional condition for matching distinctions. This was done to prevent, for example, the restaurant "Le Louis XV Alain Ducasse," which holds three stars, from being recognized as "Le Louis," which has 'only' two stars. This decision is somewhat arbitrary and may not be applicable in a more generalized setting with more restaurants, but it works effectively in this case. All matches have been manually checked to ensure maximum accuracy.

However, it's important to note that eleven restaurants from Wikipedia were unable to find a match in the database. Among them, there are restaurants whose names contain parts found within the Michelin Guide's restaurant names (e.g., "Regis et Jacques Marcon" compared to "Restaurant Marcon"), or restaurants with spelling mistakes (e.g., "le 1947 à cheval blanc" not recognized as "le 1947 au cheval blanc"). These issues might be addressed by excluding 'general words' such as *restaurant, à, au, le ...*, but caution must be exercised to avoid unintended

matches. Even more challenging, some names are entirely different, like "L'auberge du pont Collonges," which is now simply called "Paul Bocuse". Surprisingly, certain restaurants listed on Wikipedia as having two stars (verified on the internet) are not on the Michelin Guide's website, or are listed as having no stars at all, and therefore do not appear in the final table.

3.2 The chefs table

After constructing the restaurants table, we aimed to create a list of chefs with external references. This list would allow us to access information about a chef associated with a specific starred restaurant, such as whether they held the distinction *Meilleur Ouvrier de France* for instance. This was made possible by utilizing the Wikidata-sourced database, which included all distinctions given to individuals whose declared occupation was chef or cook (properties of Wikidata). Initially, we restructured this table so that each chef, identified by their name, served as the key for a dictionary. Each chef's entry then linked to a dictionary containing general information (date of birth, country of citizenship, etc...) and various distinctions represented as boolean variables indicating whether the chef received each distinction. Due to the original database lacking significant constraints and consequently being quite extensive, we compiled a list of over 250 different distinctions, although not all of them are relevant.

To construct the chefs table, we utilized this information and assigned a unique identifier to each chef. Each chef and the different distinctions were intended to be unique. However, the quality of this initial table relies on that of the Wikidata database, as we directly queried it. Subsequently, we scanned the restaurant table with the chefs listed in the Michelin Guide website scraped table (and the Wikipedia page) to check if they were already in the chefs table. If they were, the restaurant table keys `current-chef-id` (or `starred-chef-id`) were assigned the associated id from the chefs table. If not, we added the chef to the chefs table without setting their information, created an identifier, and established the reference in the restaurants table as previously described.

Searching for chefs in the table was relatively straightforward compared to the process performed for restaurants, as the names were standardized. However, it is important to note that we did not consider some specificities of the Michelin Guide website's formatting when dealing with several names. For example, Paul and Marc Haeberlin each have their own Wikipedia page and, consequently, their own entry in the chefs database. However, in the Michelin Guide, they are referred to as "Paul and Marc Haeberlin," and for the key `current-chef` of their restaurant, we will find an additional entry for "Paul and Marc Haeberlin" in our chefs table.

4 Conclusion

This project aimed to create a refined and organized database of French Michelin-starred restaurants and chefs through data acquisition, extraction, restructuring, and cleaning. We efficiently scraped comprehensive information from the Michelin Guide website, while Wikipedia was utilized to fill in missing data, particularly chef names. The Wikidata database was also a key resource for creating a comprehensive chefs table.

The final database consists of two well-organized tables: one for Michelin-starred restaurants and another for chefs. Challenges included handling inconsistencies in names and specificities of the Michelin Guide website formatting.

For future improvements, addressing discrepancies in names and exploring advanced methods for linking information across databases could be considered. Overall, this project provides a foundation for diverse analytical queries related to Michelin-starred restaurants and chefs.