Proceedings of the Federated Conference on Computer Science and Information Systems pp. 599–603
DOI: 10.15439/2019F69
ISSN 2300-5963 ACSIS, Vol. 18

Mapping of Dental Care in the Czech Republic: Case Study of Graduates Distribution in Practice

Matěj Karolyi
Faculty of Informatics, Masaryk University, Botanická 68a, Brno, 602 00, Czech Republic
and
Institute of Health Information and Statistics of the Czech Republic
Palackého nám. 4, 128 01, Praha 2, Czech Republic
Email: karolyi@iba.muni.cz

Jakub Ščavnický, Martin Komenda
Institute of Health Information and Statistics of the Czech Republic and Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University - joint workplace,
Palackého nám. 4, 128 01, Praha 2, Czech Republic
Email: {scavnicky, komenda}@iba.muni.cz

Jan Buďa, Tereza Jurková, Monika Mazalová
Faculty of Science, Masaryk University, Kotlářská 2, Brno, 625 00, Czech Republic
Email: {451441, 451627, 451455}@muni.cz

Abstract—**Online registers contain a large amount of data about healthcare providers in the Czech Republic. Information is available to all citizens and can be useful to patients, governmental organisations or employers. Based on these data, we are able to create a high-quality snapshot of the current state of healthcare providers. Interconnecting data from more data sources together is an interesting task, and accomplishing it enables us to ask more complex questions. This paper focuses on answering several questions about dentists in our country. A dataset from one online database was created, using automated data mining methods and a subsequent analysis. Results are presented via an online tool, which was provided to owners of the data. They reviewed our results and decided to use our findings for the presentation to the Czech government and subsequent negotiation processes. Our paper describes used methods, shows some results and outlines possibilities for further work.**

I. Introduction

Dental care coordinated by the Czech Dental Chamber is an integral part of the healthcare system in the Czech Republic. Various dental services are provided by more than 8,000 dentists working at university clinics or municipal health centres as well as by private dentists and dental laboratories[1]. More than half of these dental practices are based in big cities, including the Capital of Prague. Universities provide dental study programmes fully in accordance with standards of the European Union. Dentistry curricula are completely separated and independent from general medicine study programmes. In general, dentistry programmes cover all basic requirements to practice dentistry in terms of prevention, diagnosis, treatment, medical opinion and monitoring. Students achieve knowledge and skills of all the activities and interventions as regards prevention, diagnosis and treatment of anomalies and diseases of the teeth, gums, jaws, and surrounding tissues[2]. From the government and higher education perspective, many online data sources describing the field of dentistry on both national and regional levels are available. This information can provide interesting inputs for further analyses which explore new relations and patterns between graduates and practices.

A. Portals presenting national dental care professionals

The network of healthcare providers in the Czech Republic is complex and well-described. The Ministry of Health of the Czech Republic and its departments provide searchable databases of all registered providers. These databases, which are available to all citizens, were included in a complex national review [1] of publicly available web portals together with others mainstream websites providing information from healthcare and medicine. The major online databases of individual healthcare providers and organisations are listed in Table I.

Table I
Major Czech databases of healthcare providers

Name of the portal	**Reference**
Czech Medical Chamber	www.lkcr.cz
Czech Dental Chamber	www.dent.cz
National Register of Healthcare Providers	nrpzs.uzis.cz
Open Data of the Ministry of Health of the Czech Republic	opendata.mzcr.cz
Open Data of the State Institute for Drug Control	opendata.sukl.cz
Portal of Advisory Bodies, Working Groups and Expert Committees of the Ministry of Health of the Czech Republic	ppo.mzcr.cz
Portal for Patients and Patient Organisations	pacientskeorganizace.mzcr.cz
ZnamyLekar	www.znamylekar.cz

In this paper, we focus on the second of the above-mentioned databases, which is guaranteed by the Czech

[1] https://www.dent.cz/

[2] https://www.muni.cz/en

Dental Chamber. These data fit most conveniently to our further investigation because they contain information only about dentists, not about other health professionals. The obtained dataset from the publicly available database is therefore as relevant as it can be for further examination.

B. Motivation and exploratory questions

On the one hand, a lot of data describing dental care in the Czech Republic are freely available. In theory, there are no limitations and borders to mine and to process those data. On the other hand, a huge amount of records on individual dental care providers make a global overview and orientation in the particular domain of dentistry on the national level quite complicated and unclear. Moreover, the manual process of data extraction and local database construction is very time-consuming.

This paper aims to find an effective way of extracting data automatically from freely accessible online sources using a machine-based – instead of a human-based – approach. With respect to our other research activities [2]–[4], we decided to explore the domain of dental care from two different perspectives: (i) Czech higher education institutions, which guarantee various dental medicine study programmes, (ii) real distribution of dental professionals in everyday clinical practice. The process of mapping of dental care in the Czech Republic in terms of graduates' distribution across the country was a challenge from the very beginning. A student project devoted to this particular topic was solved at the Faculty of Science of the Masaryk University. Based on data from the Czech Dental Chamber portal, a pilot automated mapping between graduates and dental professionals was done. Finally, a web-based application presenting the achieved results in the form of an interactive visualisation has been designed, developed and implemented.

II. METHODS

The preparation of a final output (i.e. the online visualisation tool in this case) had several stages: obtaining the dataset, data preparation for further analysis, development of multiple interactive views and a final evaluation. All activities were carried out by a team of three students under the supervision of mentors from the Web Design Department[3] of the Institute of Biostatistics and Analyses at the Faculty of Medicine of the Masaryk University (IBA FM MU). During the process of data mining, we followed the standardised and proven methodology called the cross-industry standard process for data mining (CRISP-DM). It helped us to avoid the common mistakes and to work efficiently as a team [5]. We have distributed our activities in this case study, too. The next sections describe our steps in the context of CRISP-DM.

A. Business and data understanding

The web portal dent.cz provides information for members of the Czech Dental Chamber (CDC) as well as for the general public.

The portal consists of the following sections:

- list of dentists,
- education – a calendar of events, recommended literature and other study materials,
- LKS journal – information about their periodical,
- news – current events and news in the dentistry,
- about us – general information about CDC,
- for members – accessible only to CDC members,
- contacts – contact to the CDC office.

In particular, we were interested in the very first item, i.e. the list of dentists, for further investigation.

Records on particular dentists are available through records on individual healthcare facilities, and each dentist can be registered at none, one or more of these facilities. All records have a clearly defined common structure consisting of the dentist's name, information about his/her workplace, education and regional dental chamber. The section about healthcare facility where the dentist works is the key part of the record. It consists of the name of the healthcare facility, its address and contact. Three ways of filling this section are distinguished. In the first case, the dentist works only in one facility. In the second case, the dentist works in more than one facility. Finally, no healthcare facility is mentioned. Information about the dentist's workplace(s) is supplemented by a map.

The education section was another important part of this study. As was the case of healthcare facilities, it was filled in three different ways (one university, more than one university, no university mentioned).

B. Data preparation

Data preparation consisted of two main steps. The first one involved web mining methods and the insertion of gathered pieces of information about dentists in a structured form into the database. The subsequent phase focused on data cleaning, which meant extracting useful analytical information by regular expressions from HTML codes into a new table in the database. All steps in this section were created in the Python programming environment using libraries that are described in Section 2.2.1. In the following text, more detailed information about the algorithm we have designed will be provided.

Web mining is generally called crawling [6] because the algorithm goes gradually through the web portal hierarchy. The crawling algorithm has two functionally different parts. The first part consists of many functions which work with URL links. The function for extracting all URLs from a specific web page is the most important segment of the code, in which several key conditions are defined. For instance, we had to select only unique URLs from the list of all URLs on the page, and we needed to ensure the algorithm would be terminated when the 'offset' was detected in the crawled URL. After we obtained the final list of URLs, we

[3] http://www.iba.muni.cz/index-en.php?pg=contract-research--web-design

created the function for scraping a specific URL. This is represented by extracting the HTML code, which includes all the information about dentists in the free-form text, from the web page.

In the second part of the proposed crawling algorithm, the obtained information was inserted into a newly created database. Firstly, the connection to a SQLite Database Server was created, then a table for crawled URLs was created and the first record was added to the database. Secondly, all URLs from the first web page were inserted into the database table and the table was updated with the HTML code of this page. This process was iterative until the HTML codes of all records in the table were filled.

The proposed methods of the extraction algorithm mentioned in the introduction of this section have two parts as well. In the first one, functions for extraction of information about dentists were created, using various regular expressions from the HTML codes. In this part of the code, names of schools had to be unified because there was an enormous inconsistency in foreign school names. Firstly, a new table for the extraction was created, then only the records about dentists (not about healthcare facilities) were selected from the primary table. Secondly, all extracted attributes were inserted into the database table at the same time. In this key step, the issue with more healthcare facilities per dentist was resolved by a uniform distribution of the Full-time equivalent (FTE) among the workplaces. For example, if a dentist worked in three workplaces, then the weight of each record about this dentist was 0.33. For further interactive data analyses and visualisations, it was crucial to extract the names of healthcare facilities as well, since each healthcare facility was defined with respect to its geographical location as a unique combination of the name of the relevant facility, its latitude and longitude. Subsequently, it was necessary to obtain further information about a given region and a district workplace using the postal code of that workplace, using data from the web portal http://www.psc.cz. This information was automatically extracted from the HTML code of this web portal using the above-mentioned method. The postal code of each workplace was primarily used to assign both the region and the district to each healthcare facility.

Using this procedure, however, we were not able to assign all regions and districts, so we subsequently decided to use the municipality where a given healthcare facility was located for search on the http://www.psc.cz portal. In this manner, the number of healthcare facilities with unknown regions and districts was significantly reduced. The proposed extraction algorithm was conducted with a SQL update of the table and thus the final version of the dataset for further data analysis was obtained.

C. Modeling and Evaluation

Extracted and cleaned data saved in the SQLite Database Server was connected with the R programming language. Afterwards, SQL queries were executed over the database using the R programming environment. These data aggregations were used for the creation of R Shiny application, especially for descriptive statistics and visualisations that were represented for example by textual descriptions donut charts or cartograms.

The application was independently evaluated twice in the work team: within the students' team and then in the mentors' team. The final output was also presented to other teams of the subject and their mentors. The whole auditorium had an opportunity to participate in the discussion. Subsequently, the application was presented to representatives of the CDC. They commented on factual accuracy and usability of the presented outputs. Finally, all remarks collected during the review process were incorporated into the application.

D. Technological background and deployment

Various technologies, tools and packages were used during the deployment process. We are able to divide the technologies we used into two categories by their purpose within the whole project: (i) data retrieval group – tools and libraries which were used to obtain data from the web portal of the Czech Dental Chamber, (ii) data visualisation group – tools used in the final presented application for computing and rendering the user views with graphs and text information.

Data retrieval group

The technological group of data retrieval consisted of scraping [7], database operations, data cleaning and data parsing. Each of the related procedures and methods were performed using the Python 3 language. Packages like urllib, requests, BeautifulSoup and sqlite3. SQLite was used as the application's database layer.

Data visualisation group

The technological group of data visualisation consisted of data aggregation, charts plotting and app deployment. Each of the related procedures and methods were performed in the R software environment. The RSQLite package was used for communication with the SQLite database engine in order to aggregate data effectively. Packages plotly and networkD3 were used to create interactive pie charts and a map of the Czech Republic. The whole application was built as an interactive R application using the packages Shiny and shinythemes. Furthermore, it was deployed on our Open CPU server. Therefore, our R Shiny applications allow real-time user interaction and data filtering in a simple online environment.

III. RESULTS

We have created a publicly available tool which shows the state of distribution of dental care graduates all over the Czech Republic. The dataset was collected on 22 October 2018. The tool is available on a public URL[4] with the user interface translated to English.

[4] https://vis.iba.muni.cz/apps/dent-en/

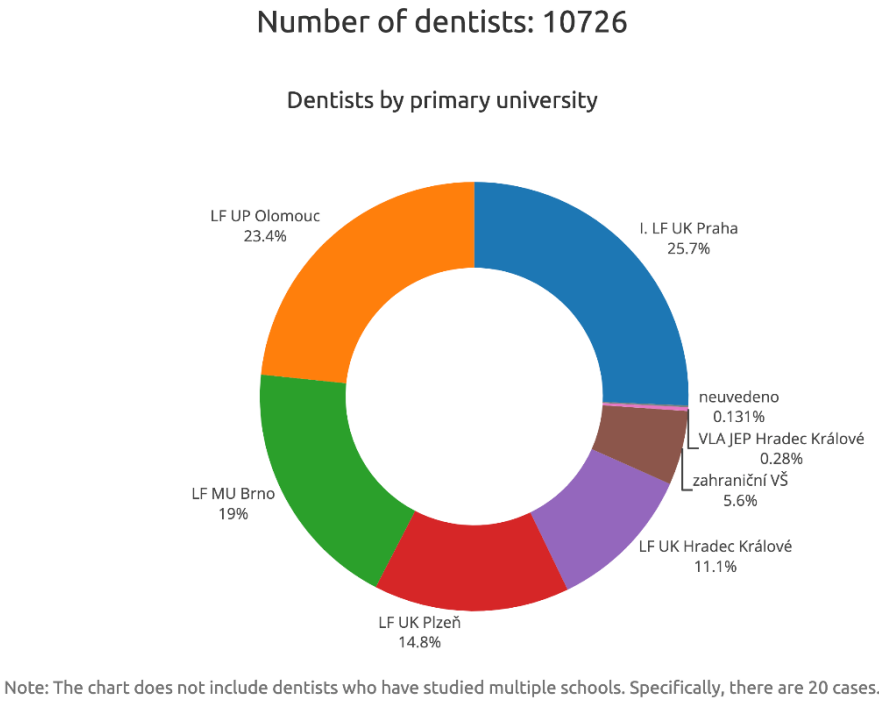

Fig. 1 Percentage of dentists by universities at which they studied primarily

A. Basic description and overview of studies

The application was created as an output of a student project. On the initial tab of the web application, the objective and basic information are mentioned.

The second tab deals with the education of dentists, precisely with their primary school in relation to their dental practice. Six faculties of medicine in the Czech Republic were distinguished:

- First Faculty of Medicine of the Charles University in Prague,
- Faculty of Medicine and Dentistry of the Palacký University in Olomouc,
- Faculty of Medicine of the Masaryk University in Brno,
- Faculty of Medicine in Plzeň of the Charles University,
- Faculty of Medicine in Hradec Králové of the Charles University,
- Jan Evangelista Purkyně Military Medical Academy in Hradec Králové.

All foreign universities were united into one category and another category was created by merging dentists with missing information on university at which they studied.

The total number of dentists registered on the website of the Czech Dental Chamber was 10,726. Twenty of them mentioned two different universities at which they had studied. It was not possible to identify which of these universities was the one at which they had studied primarily, therefore these dentists were not included in further analyses.

Fig. 1 shows the percentage of dental practitioners by universities at which they studied primarily. Most dentists graduated from the First Faculty of Medicine of the Charles University in Prague (25.70%), the Faculty of Medicine and Dentistry of the Palacký University in Olomouc was the second most frequently mentioned one (23.40%), and the Faculty of Medicine of the Masaryk University in Brno was the third one (19.00%). The proportion of dentists who studied abroad was 5.60%. In fourteen cases, the dentist's education was unknown (0.13%).

B. Dental offices in the Czech Republic

Dental offices in the Czech Republic are displayed on two tabs (the third one and the fourth one). The first of them describes only dental offices regardless of information on university graduates. About a fifth (21.00%) of dentists did not mention the healthcare facility in which they worked. More than two thirds /70.10%) of dentists worked in just one office and 8.90% of them worked in several offices. Therefore, a new variable was created – work time. Connection of work time to the FTE is described in more detail in chapter Data preparation. The third tab also displays maps containing information about numbers of healthcare facilities (6,579 in total) and work time of dentists by region. The cartogram allows user interactivity in the form of radio buttons. The user can choose from two options: the first of them shows the numbers of work times, whereas the second option displays the numbers of healthcare facilities by region. The first option can be seen in Fig. 2. This image clearly shows that the majority of dentists work in the capital (Prague) and its vicinity (Central Bohemian Region), followed by the South Moravian Region and the Moravian-Silesian Region. It is also obvious that the Karlovy Vary Region, has the lowest work time in the Czech Republic.

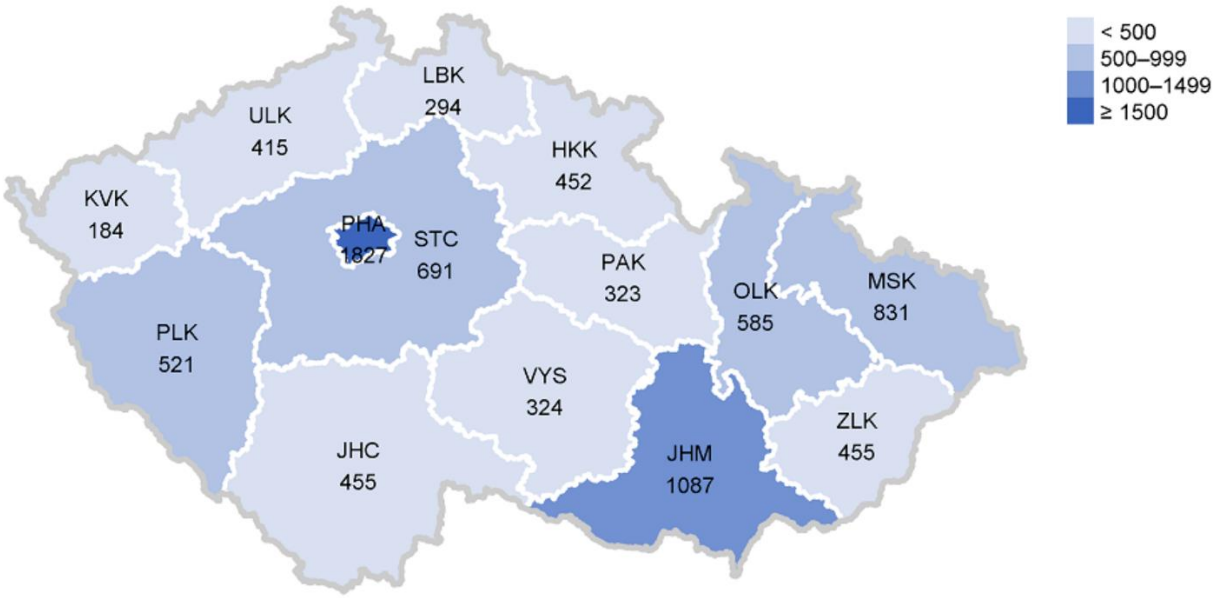

Fig. 2 Numbers of work time by region.

IV. Discussion

This paper represents a student project solved under the supervision of senior mentors, where a proof of concept in online data crawling and scraping was carried out. Real data from a guaranteed online source on dental care were automatically mined and processed by a set of Python algorithms and stored in a relational database running on our own servers. The challenging task of complex mapping between data describing the distribution of dental professionals in practice was successfully solved. The final application is one from a group of similar projects [8], [9] being solved at the Institute of Biostatistics and Analyses at the Faculty of Medicine of the Masaryk University. It provides an original overview of data stored in the portal of the Czech Dental Chamber because it reveals hidden relations between graduates' and dental professionals' distribution across the Czech Republic.

The public R Shiny application and its outputs were subsequently consulted with representatives of the Czech Dental Chamber. Conclusions of the review were considered and implemented in the application. We believe that information obtained in this way will serve to increase the transparency of healthcare in the Czech Republic and will be an interesting source of knowledge for the entire community associated with the Czech Dental Chamber.

The proposed application is still open to changes and improvements. Updating the underlying dataset at different times would also be worth considering. It would then be possible to compare the evolution of migration over time and monitor the increment / decline of registered dentists across certain periods. In the future, it would be interesting to include demographic data from individual regions of the Czech Republic in the analysis. It would then be possible to estimate the number of citizens in a certain region per one dental practitioner and whether a certain region is lacking this type of healthcare providers.

References

[1] M. Karolyi and M. Komenda, 'PŘEHLED ELEKTRONICKÝCH INFORMAČNÍCH ZDROJŮ VE ZDRAVOTNICTVÍ ČR', *MEDSOFT 2019*, p. 5.

[2] L. Dušek, J. Mužík, M. Karolyi, M. Šalko, D. Malúšková, and M. Komenda, 'A Pilot Interactive Data Viewer for Cancer Screening', in *Environmental Software Systems. Computer Science for Environmental Protection: 12th IFIP WG 5.11 International Symposium, ISESS 2017, Zadar, Croatia, May 10-12, 2017, Proceedings 12*, 2017, pp. 173–183.

[3] C. Vaitsis *et al.*, 'Standardization in medical education: review, collection and selection of standards to address', *MEFANET J.*, vol. 5, no. 1, pp. 28–39, Nov. 2017.

[4] M. Komenda, M. Karolyi, C. Vaitsis, D. Spachos, and L. Woodham, 'A Pilot Medical Curriculum Analysis and Visualization According to Medbiquitous Standards', in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 2017, pp. 144–149.

[5] R. Wirth, 'CRISP-DM: Towards a standard process model for data mining', in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.

[6] S. vanden Broucke and B. Baesens, *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. Apress, 2018.

[7] R. Mitchell, *Web scraping with Python: collecting data from the modern web*, First edition. Sebastopol, CA: O'Reilly Media, 2015.

[8] L. Woodham, J. Ščavnický, M. Karolyi, and M. Komenda, 'Interactive presentation of evaluation data in training against medical errors', *Masarykova univerzita*, 2018. [Online]. Available: https://www.muni.cz/vyzkum/publikace/1476359. [Accessed: 30-Apr-2019].

[9] M. Komenda, J. Ščavnický, P. Růžičková, M. Karolyi, P. Štourač, and D. Schwarz, 'Similarity Detection Between Virtual Patients and Medical Curriculum Using R', *Stud. Health Technol. Inform.*, vol. 255, pp. 222–226, 2018.