engineering proceedings

MDPI

Proceeding Paper

Tackling the Data Sourcing Problem in Construction Procurement with File Scraping Algorithms

Luís Jacques de Sousa [1,2]*, João Poças Martins [1,2,3] and Luís Sanhudo [2,3]

[1] Faculty of Engineering of the University of Porto (FEUP), Department of Civil Engineering (DEC), 4200-465 Porto, Portugal;
[2] CONSTRUCT/GEQUALTEC, FEUP–DEC, 4200-465 Porto, Portugal;
[3] BUILT CoLAB–Collaborative Laboratory for the Future Built Environment, 4150-003 Porto, Portugal;
* Correspondence: luisjsousa@fe.up.pt;
† Presented at the 1st International Online Conference on Buildings, online, 24-26 October 2023.

Abstract: The Architecture Engineering and Construction (AEC) sector has a lower adoption rate of machine learning (ML) tools than other industries with similar characteristics. A significant contributing factor to this lower adoption rate is the limited availability of data, as ML techniques rely on large datasets to train algorithms effectively. However, the Construction process generates substantial data that provides a detailed characterisation of the project. In this sense, this paper presents a data-scraping algorithm to search construction procurement repositories systematically to develop an ML-ready dataset for training data for ML and Natural Language Processing (NLP) algorithms focused on Construction's procurement phase. This tool automatically scrapes procurement repositories, developing a procurement file dataset comprised of bills of quantities (BoQ) and project specifications.

Keywords: Construction; Public Procurement; Contract Awarding; Scraping Algorithm; Database; Artificial Intelligence; Machine Learning; Natural Language Processing.

Citation: de Sousa, L.; Martins, J.P.; Sanhudo, L. Tackling the Data Sourcing Problem in Construction Procurement with File Scraping Algorithms. **2023**, *5*, x. https://doi.org/10.3390/xxxxx
Published: 24 October

1. Introduction

In recent years, there has been an increased interest in implementing ML and NLP tools in the AEC sector [1,2]. Still, the adoption rate of these tools is low compared to other industries with similar characteristics [3].

A significant contributing factor to this lower adoption rate is the limited availability of data, as ML techniques rely on large datasets to train algorithms effectively [4,5]. This difficulty in sourcing abundant data in the Construction sector presents one of the main challenges for ML developers within the AEC industry [6,7]. Nonetheless, this paradigm clashes with the inherent workings of the construction process since it generates a significant amount of data that offers a comprehensive characterisation of the project [4,8].

In the specific case of Portuguese Construction Procurement, public construction projects are mandatorily submitted to online, open-source repositories [9,10]. However, the consultation and extraction of procurement files is decentralised and not automated, making data agglomeration difficult and time-consuming [11]. Previous studies have tackled this difficulty by scraping procurement data in these repositories to a tabular dataset to be used in ML applications [11,12]. Thus, if the necessary diligence is ensured, the procurement phase represents a great opportunity for data aggregation.

In light of this, this paper presents a data-scraping algorithm capable of extracting data from construction procurement repositories. This tool automatically scrapes procurement repositories, developing a procurement file dataset comprised of BoQ and project specifications. Future studies will use the gathered data to develop an ML-ready dataset

for training data for ML, and NLP algorithms focused on Construction's procurement phase.

The remaining document is organised into three Sections: Section 2, presents the methods and codes developed to scrape open-source data to a semi-strutured format; Section 3 describes the gathered data and briefly highlights the framework where scraped data will be used; and Section 4 presents conclusions and final remarks.

2. Methods

Following previous work [11,12], a reengineered version of the PPPData algorithm was developed. As highlighted in Figure 1, this new algorithm focused on scraping procurement files from the open-source online repository Portal Base [9] using Selenium [13] and Chrome Driver [14] Python libraries.

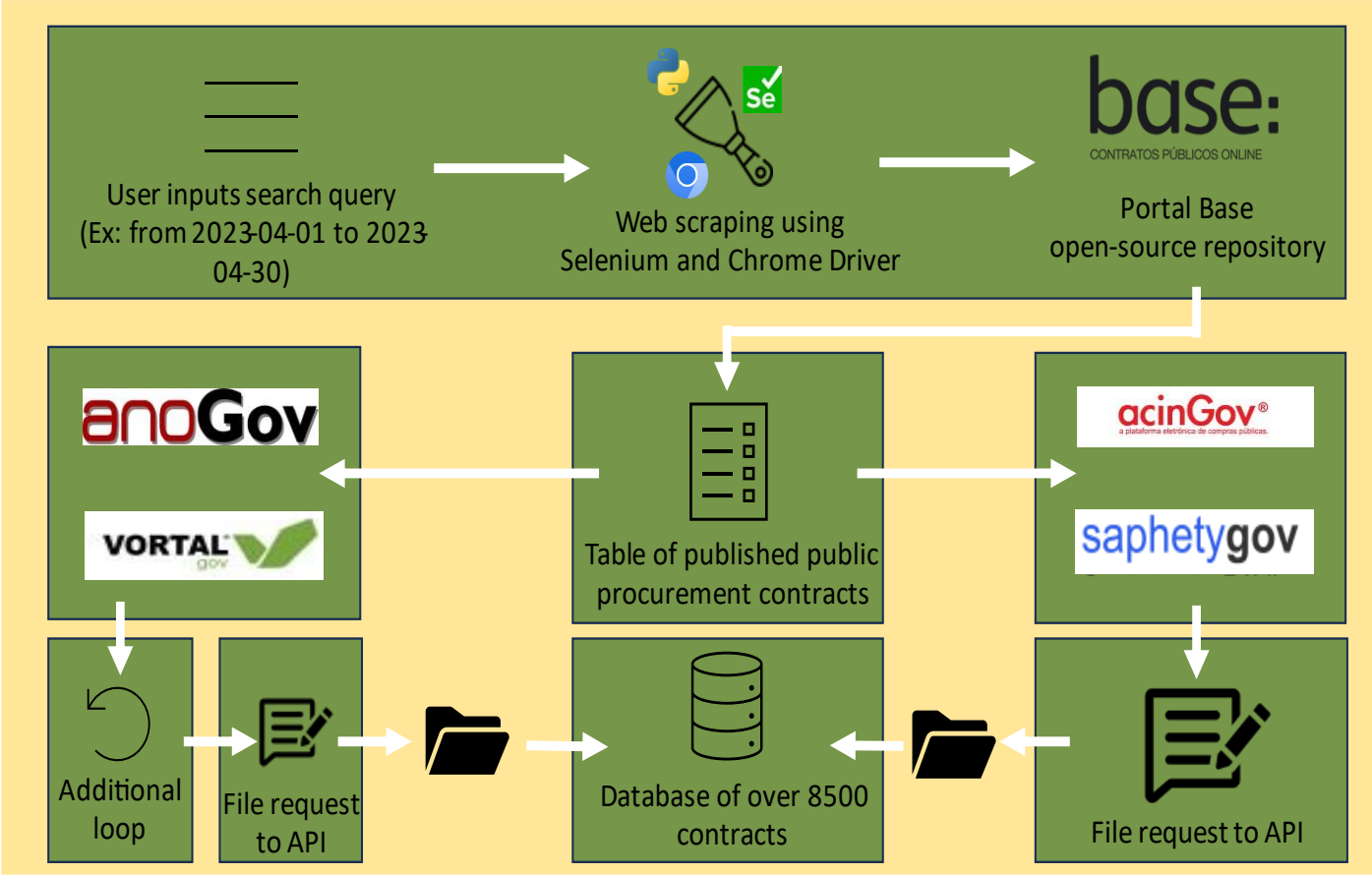

Figure 1. Data scraping methodology

Although bulk download is possible using this algorithm, a month-by-month method was used for scraping where a search query was inputted to the algorithm stating the month and year the user intended to scrape. This method allowed for easier database organisation in later phases of data processing. Next, the algorithm would open google chrome and load the page with the results of Portal Base for that specific month. The Portal Base platform organises its information in a table where each line is a contract. Each contract has a detailed page from which procurement files can be downloaded. The algorithm looped through all the tables on each page and the lines in each table to open the detailed contract page and download the procurement files.

The procurement files could be located in 4 different online platforms: (1) Acingov [15]; (2) Saphetygov [16]; (3) Vortalgov [17]; (4) Anogov [18]. For the first two platforms, a simple request to the platform API using the reference located in the procurement files

download link was sufficient to obtain a compressed folder with the procurement files of that contract.

In the case of Vortal, a new web page was opened from Vortal's website. This page had all the information associated with the contract in question, including the procurement files, in a table. The algorithm had to loop through all the lines in the table and individually download each file, which was then associated into a single folder.

A similar process had to be done to the Anogov-based contracts. However, the algorithm had to open new rows in the table of files by clicking a hidden button, only visible if the mouse hovered over a symbol in the table. Each file was downloaded through a request to Anogov API using the reference in the hidden row. Finally, all the files were associated in a folder.

At the time of writing, all available procurement files from April 2023 to January 2020 have been gathered in a raw dataset comprising 8612 folders from as many public venture construction contracts.

All code used in this paper's methodology can be accessed on GitHub using the following link: https://github.com/LuisJSousa/ScrapeProcurementFiles.

3. The Data

As previously mentioned, the algorithm successfully scraped over 8500 folders of files from as many contracts, each containing text-based documents in Microsoft Excel, Microsoft Word, and PDF formats. These files represent various procurement documents, including BOQs, Project specifications, and other legally required files essential for procurement processes.

The existing dataset is in a raw format. Its structure follows a hierarchical order, with folders organised by year and month, further divided by the name or number of each contract. All the documents associated with each specific contract are stored within these final folders.

In future endeavours, multiple rounds of data treatment will be necessary to classify the different types of documents into standardised groups, making them suitable for machine learning applications.

The primary objective of these future efforts is to create a substantial dataset of BOQs which will be instrumental in automating the generation of these documents for budget proposal purposes, as established in the framework shown in Figure 2, presented in [19].

The framework involves data aggregation using the web scraping algorithm presented in this paper. Subsequently, a "master" BoQ is selected, preferably the one most frequently used by enterprises selected to participate in this study. In case this "master" BoQ is not available, an arbitrated BoQ will be chosen.

Next, different algorithm architectures will be developed, employing various architectures and different Python libraries focused on ML and NLP, using the scraped data to train algorithms to classify BoQ tasks. This training phase is followed by a testing phase where the accuracy of the algorithms will be evaluated, testing their ability to classify BOQ tasks effectively. Moreover, the efficiency of the algorithms will be compared with the manual classification typically performed by technicians.

BoQs used during budgeting will be uploaded to the database to enable continuous learning, thereby increasing the volume of historical data that contributes to the algorithm's classification capabilities. This iterative learning process will enhance the tool's performance and effectiveness over time.

In this sense, the scraping algorithm developed in this communication is a crucial step in achievieng future goals because establishing a well-organised and extensive dataset will significantly enhance the potential for accurate and efficient automation of these processes.

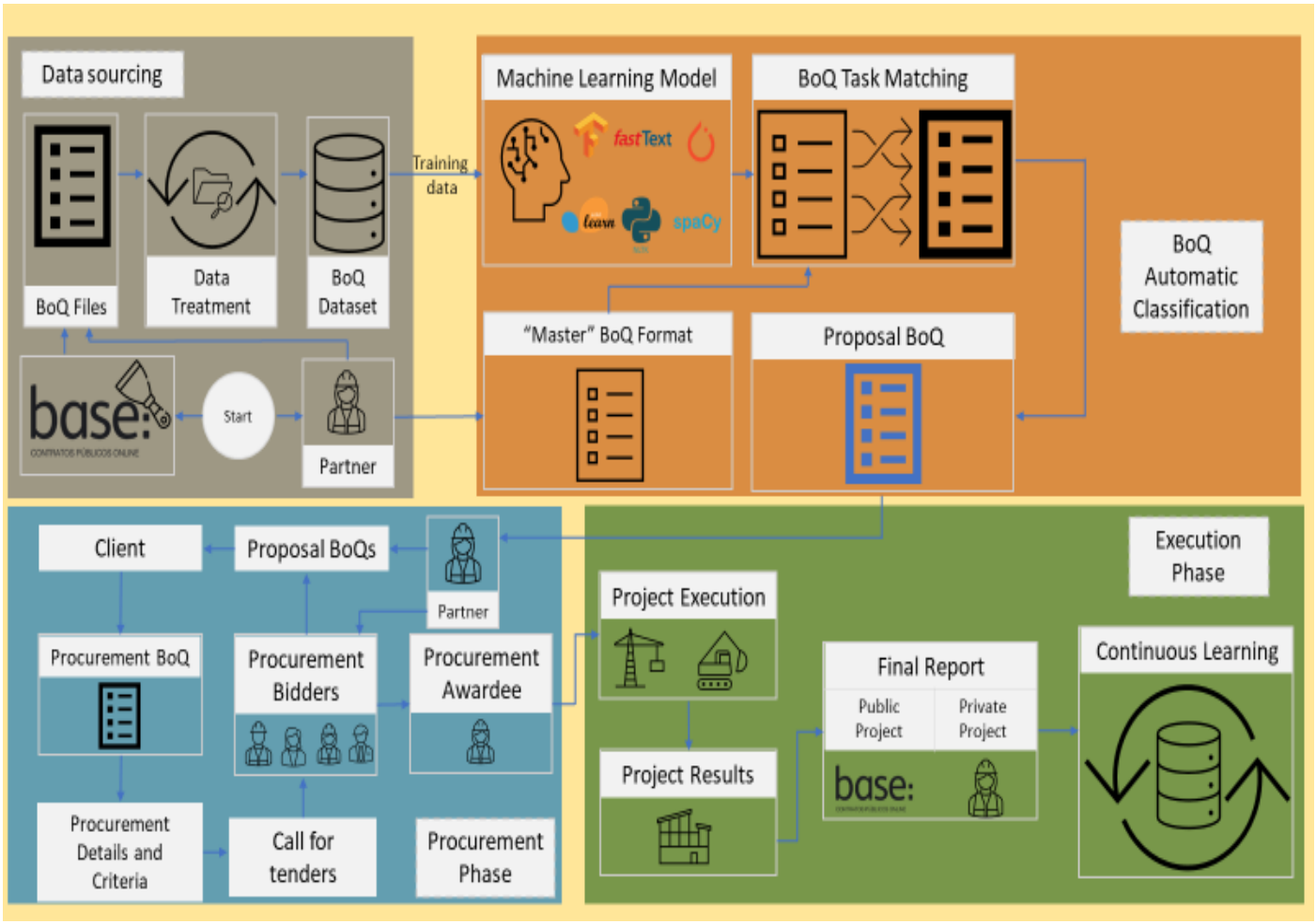

Figure 2. Implementation framework, adapted from [19]

4. Conclusions

The implementation of ML and NLP applications in the AEC sector is still in its early stages compared to other industries with similar characteristics. A major obstacle to progress in this area is sourcing relevant and reliable data. However, the construction industry itself generates a vast amount of data during its operations, presenting a unique opportunity to tackle this issue and enabling the use of ML tools.

For the specific case of the Portuguese AEC sector, procurement files are mandatorily submitted to online repositories presenting a significant opportunity for data agglomeration.

In this sense, this research proposes a solution to the data-sourcing problem through the use of data-scraping algorithms. By employing an automated approach, the presented algorithm can extract information from online open-source repositories containing procurement files suitable for ML applications in the AEC domain.

Notably, the algorithm was capable of scraping more than 8500 file folders from as many public procurement contracts. This led to the creation of a significantly large and diverse raw dataset comprising procurement documents such as BOQs and project specifications, laying the groundwork for future advancements in ML and NLP within the construction industry. Future studies will focus on processing and organising the gathered data to create a well-structured dataset. This critical step will pave the way for the development of complex ML applications aimed at automating the creation of BOQs for procurement purposes. Its final goal is to transition the budget-making process from a laborious classification task to a more efficient verification-based approach. By streamlining the BoQ generation process, this technology can accelerate budget proposal

development in the construction sector, saving time and resources while improving accuracy and efficiency.

Supplementary Materials: The following supporting information can be downloaded at: Code: https://github.com/LuisJSousa/ScrapeProcurementFiles

Author Contributions: Conceptualisation, J.P.M.; methodology, L.J.d.S.; software, L.J.d.S.; validation, J.P.M. and L.S.; formal analysis, L.J.d.S., J.P.M. and L.S.; investigation, L.J.d.S; resources, L.J.d.S. and J.P.M; data curation, L.J.d.S., J.P.M. and L.S.; writing—original draft preparation, L.J.d.S.; writing—review and editing, J.P.M and L.S.; visualisation, L.J.d.S; supervision, J.P.M. and L.S.; project administration, J.P.M. and L.S.; funding acquisition, J.P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the European Regional Development Fund (ERDF) through the Operational Competitiveness and Internationalisation Programme (COMPETE 2020) (funding reference: POCI-01-0247-FEDER-046123) and by Base Funding—UIDB/04708/2020 of the Research Unit CONSTRUCT—Institute of R&D in Structures and Constructions—funded by national funds through FCT/MCTES (PIDDAC). This work was also co-financed by the European SocialFund (ESF) through the Northern Regional Operational Programme (Norte 2020) (funding reference: NORTE-06-3559-FSE-000176).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chung, S.; Moon, S.; Kim, J.; Kim, J.; Lim, S.; Chi, S. Comparing natural language processing (NLP) applications in construction and computer science using preferred reporting items for systematic reviews (PRISMA). *Automation in Construction* **2023**, *154*, 105020, doi:https://doi.org/10.1016/j.autcon.2023.105020.
2. Jacques de Sousa, L.; Poças Martins, J.; Santos Baptista, J.; Sanhudo, L. Towards the Development of a Budget Categorisation Machine Learning Tool: A Review. In Proceedings of the Trends on Construction in the Digital Era, Guimarães, Portugal, 2023//, 2023; pp. 101-110.
3. Sepasgozar, S.M.E.; Davis, S. Construction Technology Adoption Cube: An Investigation on Process, Factors, Barriers, Drivers and Decision Makers Using NVivo and AHP Analysis. *Buildings* **2018**, *8*, 74.
4. Munawar, H.S.; Ullah, F.; Qayyum, S.; Shahzad, D. Big Data in Construction: Current Applications and Future Opportunities. *Big Data and Cognitive Computing* **2022**, *6*, 18.
5. Elmousalami, H.H. Data on Field Canals Improvement Projects for Cost Prediction Using Artificial Intelligence. *Data in Brief* **2020**, *31*, 105688, doi:10.1016/j.dib.2020.105688.
6. Phaneendra, Seethamraju; Reddy, E.M. Big Data - Solutions for RDBMS Problems - A Survey. 2013.
7. Jacques de Sousa, L.; Martins, J.; Baptista, J.; Sanhudo, L.; Mêda, P. Algoritmos de classificação de texto na automatização dos processos orçamentação. In Proceedings of the 4º Congresso Português de "Building Information Modelling", Braga, Portugal, 2022; pp. 81-93.
8. Xu, W.; Sun, J.; Ma, J.; Du, W. A personalised information recommendation system for R&D project opportunity finding in big data contexts. *Journal of Network and Computer Applications* **2016**, *59*, 362-369, doi:https://doi.org/10.1016/j.jnca.2015.01.003.
9. (IMPIC), I.d.M.P.d.I.e.d.C. Portal Base. Available online: https://www.base.gov.pt/ (accessed on April 2023).
10. DRE. Diário da República Electónico. Available online: https://dre.pt/dre/home (accessed on January 2023).
11. Jacques de Sousa, L.; Poças Martins, J.; Sanhudo, L. Base de dados: Contratação pública em Portugal entre 2015 e 2022. In Proceedings of the Construção 2022, Guimarães, Portugal, 2022.
12. Jacques de Sousa, L.; Poças Martins, J.; Sanhudo, L. Portuguese public procurement data for construction (2015–2022). *Data in Brief* **2023**, *48*, 109063, doi:https://doi.org/10.1016/j.dib.2023.109063.

13. Selenium. Available online: https://www.selenium.dev/ (accessed on July 2023).
14. Chrome Driver. Available online: https://chromedriver.chromium.org/downloads (accessed on July 2023).
15. Acingov. Available online: https://www.acingov.pt/acingovprod/2/index.php/ (accessed on July 2023).
16. Saphetygov. Available online: https://gov.saphety.com/bizgov/econcursos/loginAction!index.action (accessed on July 2023).
17. Vortalgov. Available online: https://www.vortal.biz/vortalgov/ (accessed on July 2023).
18. Anogov. Available online: https://anogov.com/r5/en/ (accessed on July 2023).
19. Jacques de Sousa, L.; Martins, J.; Sanhudo, L. Framework for the Automation of Construction Task Matching from Bills of Quantities using Natural Language Processing. In Proceedings of the 5th Doctoral Congress in Engineering (DCE 23'), Porto, Portugal, June 2023, 2023.