AI-Enhanced Techniques for Extracting Structured Data from Unstructured Public Procurement Documents

Amina Oussaleh Taoufik
Intelligent Automation and BioMedGenomies Laboratory (IABL)
Faculty of Sciences and Techniques of Tangier, Abdelmalek Essaadi University of Tetouan
Tangier, Morocco
oussalehtaoufik.amina@etu.uae.ac.ma

Abdellah Azmani
Intelligent Automation and BioMedGenomies Laboratory (IABL)
Faculty of Sciences and Techniques of Tangier, Abdelmalek Essaadi University of Tetouan
Tangier, Morocco
a.azmani@uae.ac.ma

2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS) | 979-8-3315-4010-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISAS64331.2024.10845583

***Abstract*— This paper presents a methodology for extracting and structuring procurement data from scanned Summary Minutes documents obtained from the Moroccan Public Procurement Portal. Leveraging web scraping techniques with Scrapy-Selenium and Beautiful Soup, scanned PDFs were collected and processed using PaddleOCR for Optical Character Recognition. The treated text files were stored in a MongoDB database, and structured data extraction was performed using the RAG process with the Mistral LLM. Our methodology resulted in the extraction of 439,048 records, offering valuable insights into procurement practices. We discuss the data extraction process, including cleaning and mining techniques, and highlight limitations encountered, particularly regarding PaddleOCR's performance across different languages and scripts. Despite challenges, our methodology demonstrates the feasibility of extracting structured data from unstructured sources for informed decision-making in procurement analysis.**

Keywords—*Public Procurement, OCR, RAG*

I. INTRODUCTION

In today's data-driven landscape, the term "unstructured data" reverberates through organizational corridors, particularly within sensitive domains like public procurement, which are inherently susceptible to corruption and fraud. This unstructured data, ranging from emails and PDFs to scanned documents and handwritten notes, presents a great challenge for businesses and state entities alike [1], [2]. Analyzing such data demands significant financial and human resources, as manual efforts to sift through vast troves of information prove time-consuming and error-prone. In the realm of public procurement, unstructured data manifests in various forms: from unstandardized bid documents and contract agreements to fragmented supplier information scattered across disparate systems. Moreover, as digitalization accelerates, the volume of unstructured data burgeons exponentially, exacerbating the complexity of data management and analysis. In this milieu, the quest for an automated tool capable of structuring and contextualizing data assumes paramount importance. Such a tool could revolutionize public procurement processes by streamlining the identification of needs, optimizing supplier sourcing, reducing costs through data-driven decision-making, and enhancing audit capabilities to detect red flags of malfeasance. By harnessing the latent insights buried within myriad documents, whether published or archived, an intelligently constructed database promises to unlock untold benefits for governance, accountability, and ultimately, societal welfare.

The "Summary of Minutes" from open tendering processes represents a treasure trove of information that, if structured effectively, can be harnessed to drive insights and optimize decision-making in public procurement. By systematically organizing the data contained within these documents, a structured database can offer a comprehensive view of the procurement landscape. Firstly, the names of bidders and their respective statuses—whether admitted with or without reserves, rejected, or declared the winner—provide crucial insights into the competitive dynamics of the procurement process. This information enables stakeholders to assess the breadth of competition and evaluate the efficacy of selection criteria. Furthermore, understanding the reasons behind bidder elimination enriches the dataset with valuable contextual information. By categorizing these reasons—whether related to financial viability, technical capacity, or compliance—it becomes possible to identify recurring issues and areas for improvement in procurement procedures. Additionally, the financial offers submitted by each bidder offer a quantitative dimension to the analysis, allowing for comparative assessments of pricing strategies and cost competitiveness.

Through structured data analysis, stakeholders can derive actionable insights at every stage of the procurement lifecycle. From identifying market trends and supplier performance to optimizing sourcing strategies and detecting anomalies, the structured database serves as a powerful tool for enhancing transparency, efficiency, and accountability in public procurement. Moreover, the availability of historical data enables longitudinal analysis, facilitating the identification of patterns and the evolution of procurement practices over time. Overall, by unlocking the latent potential of "Summury of Minutes" through structured data exploitation, stakeholders can drive meaningful improvements in procurement outcomes and ensure the optimal allocation of public resources.

The present initiative can be generalized in order to create an African portal for public procurement as the European example or even integrate the global database to enable global insight into the sensitive domain of public procurement.

II. Literature Review and Related Works

A. Risks in Public Procurement

The field of public procurement is widely acknowledged as particularly susceptible to corruption and fraud, with various forms of corrupt practices and their consequences posing significant concerns for many countries [3]. In the construction industry, corruption is pervasive, often involving public officials and professional consultants in bribery and tender manipulation, exacerbated by factors such as skills shortages and lax deterrents [4]. This corruption extends across sectors like education and healthcare, where land allocation, tender processes, and urban planning are common areas of abuse, highlighting the urgent need for improved prevention measures [5].

Transparency emerges as a primary concern in public procurement, underscoring the necessity for enhanced systems to combat corruption and ensure fairness and efficiency. Proposals for implementing frameworks promoting good procurement practices are essential to address these irregularities [6]. Moreover, fraud and corruption in public works contracts are intricately linked to critical project phases such as tender evaluation and site supervision. Measures like quality management and audits are crucial in combating these issues effectively [7].

In the public sector, procurement stands out as a significant risk area, where counter-fraud measures are essential for tackling corruption at all levels [8]. Yet, the opacity and inefficiency in controlling public procurement processes often lead to embezzlement and the misappropriation of funds, posing severe economic repercussions [9]. Notably, in countries like Ukraine and Poland, corrupt practices in public procurement have far-reaching consequences, necessitating comprehensive strategies to combat corruption effectively [10].

Efforts to detect fraud and conspiracy in public works procurements are advancing globally, with innovative methods like Bi-LSTM showing promise in improving criminal fraud investigations [11]. However, ensuring the integrity of subjects involved in public procurement remains paramount for preventing corruption, given the susceptibility of interactions between the public and private sectors to illegal activities. Addressing corruption and fraud in public procurement requires a multifaceted approach, emphasizing transparency, ethical standards, and effective oversight. Strengthening measures to prevent corruption, including improving procurement processes, enhancing quality management, and ensuring the integrity of involved subjects, are essential steps towards mitigating these risks.

B. Challenge in Information Extraction from Unstructured Documents

The extraction of information from unstructured documents presents several challenges due to the complexity and variability of data formats and content:

- *Complex Layouts and Formats*: Information extraction from unstructured documents is complicated by the diverse layouts and formats, which require models to efficiently utilize textual, visual, and spatial information [12].
- *Domain-Specific Challenges*: The efficiency of extraction methods often depends on the linguistic structure and keyword taxonomy, which may not be suitable for domain-specific applications that require semantic and contextual understanding [13].
- *Dataset Quality and Diversity*: A significant challenge is the lack of high-quality, annotated datasets that are diverse enough to generalize key field extraction tasks across different domains [14].
- *Visual and Textual Richness*: Documents with a combination of intensive text and rich visual information, such as banking documents, necessitate advanced information extraction techniques that can handle multimodal data [15].
- *Narrative Text Data*: Extracting structured information from narrative text data, such as fire reports, is difficult due to the need for systems that can understand and process the narrative content [16].
- *Temporal Information*: The extraction of temporal information is hindered by the imprecision of temporal expressions in text, which can lead to interpretation errors [17].
- *Multilingual Documents*: Information extraction from multilingual documents adds another layer of complexity, requiring models that can handle multiple languages and their nuances [18].
- *Technical Terminology*: Scientific texts with specific technical terminology present difficulties for natural language processing tools, which are often trained on non-technical text [19].

Hence, the challenges in extracting information from unstructured documents are multifaceted. Addressing these challenges requires the development of advanced, adaptable, and robust information extraction systems capable of handling the intricacies of unstructured data.

C. Tools for Information Extraction from Unstructured Documents

Research on information extraction (IE) has evolved significantly, focusing on improving

efficiency and accuracy in various domains. Recent studies have explored different methodologies, including question-answering systems, dynamic network building, and transformers-based models, to enhance the performance of IE systems, especially when dealing with limited data.

Tesseract stands as a performant OCR tool, particularly for Latin characters and when enhanced with preprocessing techniques[20], [21]. However, its performance can be outperformed by other OCR tools like Google Cloud Vision in specific contexts, such as recognizing Thai characters [22]. Additionally, Tesseract's accuracy can be significantly improved with the right preprocessing methods and by leveraging script similarities for languages with limited training data[23]

OpenCV emerges as a highly effective tool for enhancing OCR performance across various applications [24]. By preprocessing images to improve quality and integrating with other OCR engines, OpenCV-based systems achieve high accuracy rates and versatility[25] . These systems are applicable in diverse fields, from automated vehicle guidance to historical text digitization, demonstrating their robustness and adaptability [26].

Transformer models, with their ability to model long dependencies, support parallel processing, and require minimal inductive biases [27], [28]. Excelling in multi-modal processing, achieve high performance in vision tasks, and have been adapted for specialized applications with notable success[28], [29]. Furthermore, efficiency improvements and specialized variants (3X-formers" models such as Reformer, Linformer, Performer) further enhance their applicability, making Transformers a powerful tool in modern deep learning [27].

RAG (Retrieval-Augmented Generation) can be efficient in information extraction by leveraging metadata through a filtering mechanism and using a fuse-and-oversample approach for transfer learning [30]. These approaches enhance the precision and performance of IE systems, particularly in domain-specific applications with limited data [31].

Additionally, it's worth noting that the landscape of models for OCR, language modeling, and RAG is vast and diverse, with numerous techniques achieving outstanding results. Table 1 provides a summary of some of these techniques.

I. METHODOLOGY

A. Objectif

The objective of this study is to extract structured data from unstructured scanned PDFs obtained from the Moroccan Public Procurement Portal "https://www.marchespublics.gov.ma". The data within these PDFs, referred to as Summary Minutes, contains essential information including the names of bidders, their status (whether they have been admissible with reserve, without any reserve, rejected, or retained as a winner), their financial offers, and occasionally details on the cancellation of the open tendering process as stipulated by the Moroccan Decree of public procurement (see Figure *2*).

B. Data Collection

Data collection is initiated by scraping the Moroccan Public Procurement Portal using the Scrapy framework. The portal provides structured and unstructured data. Specifically, we focus on scraping the links to the Summary Minutes documents. Once the links are obtained, Beautiful Soup, a Python library, is used to download the scanned PDF documents. The source code is available at https:// github.com/aminaot/ PublicProcurementDatabase.git..

C. Data Preprocessing

The scanned PDFs, being unstructured, required preprocessing to extract meaningful information. PaddleOCR, a comprehensive Optical Character Recognition (OCR) toolkit developed by Baidu's PaddlePaddle platform, is employed to convert scanned PDFs into editable text files. PaddleOCR's deep learning models are instrumental in accurately detecting and recognizing text within scanned documents, overcoming the challenges posed by varying fonts and languages.

D. Data Extraction and Structuring

To extract structured data from the text files, we use Groq, a powerful text extraction tool, in conjunction with the Mistral model. Groq's capabilities facilitated the extraction of relevant information such as bidder names, status, financial offers, and details on tendering process cancellations. The extracted data is then structured into a data frame, allowing easy analysis and manipulation. The source code is available at https:// github.com/aminaot/ PublicProcurementDatabase. git.

II. Results and Discussion

A. Data Extraction Process

A total of 75,788 links to Summary Minutes documents were extracted from the Moroccan Public Procurement Portal. However, only 54,881 scanned PDFs were included in the OCR process due to the presence of unsupported file formats such as zip files in the remaining links. The use of Selenium for web scraping was essential in navigating the dynamic nature of the website, which required input filtering and pagination to access all relevant links. Selenium's ability to interact with web elements dynamically justified its selection over other scraping tools.

B. Document Downloading Process

The scraped links were processed using Beautiful Soup to download the corresponding scanned PDF documents. Beautiful Soup's simplicity and ease of use made it the preferred choice for this task, as it efficiently retrieved the documents based on the scraped links. Additionally, Beautiful Soup's robust parsing capabilities ensured reliable extraction of the required data, further validating its suitability for the downloading process.

C. OCR and Data Extraction

The downloaded PDF documents were first converted into images using pdf2image, an essential step for preparing the documents for Optical Character Recognition (OCR) processing. This conversion enabled the extraction of both textual and tabular data from the scanned documents. PaddleOCR was employed as the primary OCR tool due to its strong performance in handling complex data structures. Specifically, PaddleOCR excelled in accurately extracting tabular data, outperforming other OCR tools, including Tesseract and EasyOCR, in both accuracy and reliability.

A comparative analysis was performed on a set of scanned procurement documents containing intricate tables. Tesseract frequently encountered issues with table formatting and misalignment, leading to inaccuracies in data extraction. Similarly, EasyOCR exhibited limitations in handling multi-column tables, resulting in data misalignment. Google Cloud Vision displayed the highest accuracy overall in text recognition, including complex tabular data. However, despite its superior accuracy, Google Cloud Vision is a more costly solution. In contrast, PaddleOCR demonstrated competitive accuracy, maintaining precise row and column alignment while offering a more cost-effective alternative.

PaddleOCR follows a structured process for both Layout Information Extraction and Key Information Extraction as illustrated in Figure *1*. Initially, an image undergoes Image Direction Correction to ensure the proper orientation for analysis. For Layout Information Extraction, the corrected image is processed through Layout Analysis to recognize tables and restore their format through Layout Recovery, allowing for the extraction of accurately structured tabular information. Simultaneously, Key Information Extraction involves the use of OCR for extracting text-based content, followed by Semantic Entity Recognition to identify important entities like names, dates, or titles. These recognized entities are then processed through Relation Extraction to form Structured Information, combining the layout and key details to provide restored and organized data.

To improve the accuracy and performance of OCR, we applied several preprocessing techniques, including:

- *Skew Correction:* This helped to align text horizontally, which improved the OCR performance by reducing the chance of misinterpretation of slanted text.
- *Noise Removal:* By eliminating background noise, the clarity of the text increased, resulting in more accurate character recognition.
- *Color to B&W Conversion:* Converting colored text to a binary image simplified the OCR task by focusing on the contrast between the text and background.

These steps helped reduce errors caused by image distortions and noise. Notably, thinning and skeletonization were not applied as the documents being processed (Summury of Minutes) consist of printed text, and these techniques are typically used for handwritten text.

In a broader evaluation using a sample dataset of 10,000 documents, PaddleOCR achieved an accuracy rate of 95.2%, as shown in Table 2. While Google Cloud Vision attained the highest accuracy in the evaluation, PaddleOCR remains an efficient and cost-effective solution, especially when combined with preprocessing techniques such as skew correction, noise removal, and color conversion.

D. RAG Process and Database Storage

The treated text files, obtained through PaddleOCR, were stored in a MongoDB database along with other structured and unstructured data extracted from the portal. To extract structured data from the text files, the RAG process was employed,

utilizing the Mistral LLM (Large Language Model) from the Groq Hub. Figure 1 illustrates an example of the pipeline result. The choice of RAG with Mistral LLM was motivated by its superior performance compared to other language models, such as LLAMA3, in summarizing diverse Summary Minutes documents published by various public entities. The RAG process involved several steps, including creating chunks of text from the treated files, building a vector database, and generating context and prompts for the Mistral LLM. This facilitated the generation of a structured dataframe containing the extracted data, allowing for systematic analysis and comparison across different public procurement documents.

During our evaluations, we initially tested the **Meta-Llama-3-8B** model from Meta-Llama. However, its size (16GB) exceeded the automatic loading capacity (10GB), even when running on Colab Pro with high CPU, making it impractical for our needs. We then tested the **Mistral-8x7B-Instruct-v0.1** model from Mistralai, which delivered excellent results with metrics such as Faithfulness/Groundedness = 1, Answer Relevance = 0.99989, Context Relevance = 1, Context Recall = 1, and Context Precision = 1. Despite Mistral 8's outstanding performance with French texts, its processing time was too long to be viable for the entire dataset of 95,942 documents. As a result, we opted for the **Mistral-7B-Instruct-v0.2** model, which offered similarly high-quality results but with significantly faster processing times. This model's efficiency allowed us to process the full dataset while maintaining a balance between language generation quality and the practical requirements of handling large-scale documents, meeting the demands of our project effectively.

E. Data Cleaning and Mining

The extracted data from the treated text files resulted in a total of 439,048 records stored in the MongoDB database. To ensure the accuracy and consistency of the dataset, a comprehensive data cleaning process was conducted. This involved removing duplicate records, correcting formatting inconsistencies, and standardizing data fields to adhere to predefined schemas. A major challenge we addressed was the inconsistency in bidder names, as different public buyers referred to the same bidder with variations in spelling and acronyms. For example, one buyer might list a bidder as "HALLAOUI," while another refers to the same bidder as "HALAOUI."

To resolve this, we applied several techniques for bidder name normalization:

- *Data Cleaning*: Special characters were removed, and names were standardized to lowercase, ensuring consistent formatting across the dataset.
- *Acronym Mapping:* We used a pre-defined mapping dictionary to replace acronyms with their full names, such as converting "LPEE" to "Laboratoire Public des Essais et Etudes."
- *Fuzzy Matching:* Using fuzzy string matching (Levenshtein distance), we identified and merged names with minor spelling differences, resulting in a 92% match rate for names like "HALLAOUI" and "HALAOUI."
- *Clustering:* Using hierarchical clustering based on fuzzy matching similarity scores, we grouped similar bidder names, such as "HALAOUI" and "HALLAOUI," under a common standardized version. This approach improved the consistency of bidder names by an additional 8%.

Table 1:Exploring the Landscape of Information Extraction Tools and Models

Model/Technique	Description	Examples
Optical Character Recognition (OCR)	Converts scanned documents, PDFs, or images into editable and searchable data.	Tesseract, OpenCV, Fast Fourier Transform (FFT), Scale-Invariant Feature Transform (SIFT), Support Vector Machines (SVMs), Template Matching, Gaussian Mixture Models (GMMs).
Convolutional Neural Networks (CNN)	Deep learning model commonly used for image recognition tasks, including OCR.	LeNet, AlexNet, VGG, ResNet
Long Short-Term Memory (LSTM)	A type of recurrent neural network (RNN) capable of processing sequences of data.	LSTM, GRU [31]
Bidirectional LSTM (Bi-LSTM)	An extension of LSTM that processes input sequences in both forward and backward directions.	Bi-LSTM[31]
Transformer Models	State-of-the-art models for sequence-to-sequence tasks, such as language translation.	GPT, BERT, T5, RoBERTa, ALBERT [31]
Retrieval-Augmented Generation (RAG)	Integrates retrieval-based methods with generative models for text generation tasks.	RAG, DPR
Reinforcement Learning (RL)	Machine learning paradigm where an agent learns to interact with an environment through trial and error.	Q-learning, Deep Q-Network (DQN)
Language Model (LLM)	Models trained on large text corpora to understand and generate human-like text.	GPT ; BERT; RoBERTa ; T5 ; BART ; DistilBERT , LLAMA, Mistral

Table 2: Steps and Techniques for Bidder Name Normalization

Step	Description	Example	Method/Tool
1.Data Cleaning	Remove special characters, normalize case, and trim whitespace.	"HALAOUI." -> "HALAOUI"	Python (string manipulation)
2.Acronym Mapping	Replace acronyms with full names using a pre-defined mapping table.	"LPEE" -> "Laboratoire Public des Essais et Etudes"	Dictionary-based mapping
3.Fuzzy Matching	Use fuzzy string matching to identify close name variations.	"HALAOUI" ~ "HALLAOUI"	Levenshtein Distance, fuzzywuzzy, rapidfuzz
4. Clustering	Cluster similar names together based on similarity score or fuzzy matching.	"HALLAOUI", "HALAOUI", "HALAOUI & Sons" -> "HALLAOUI"	K-means, Hierarchical Clustering

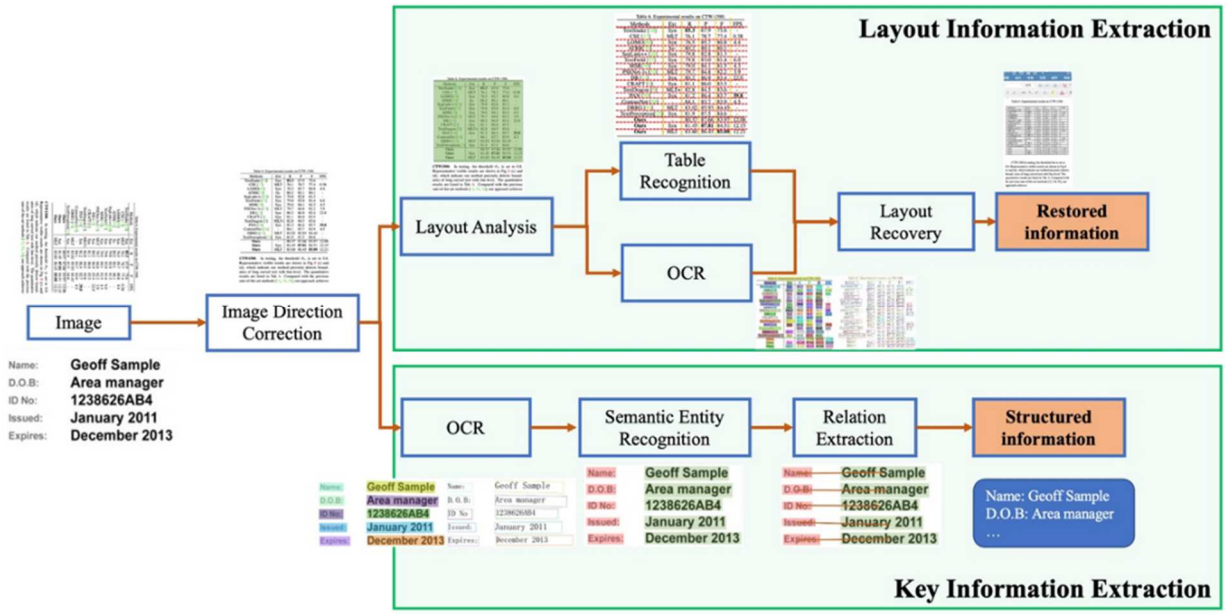

Figure 1: Paddle OCR-Based Pipeline. *Source* [32]

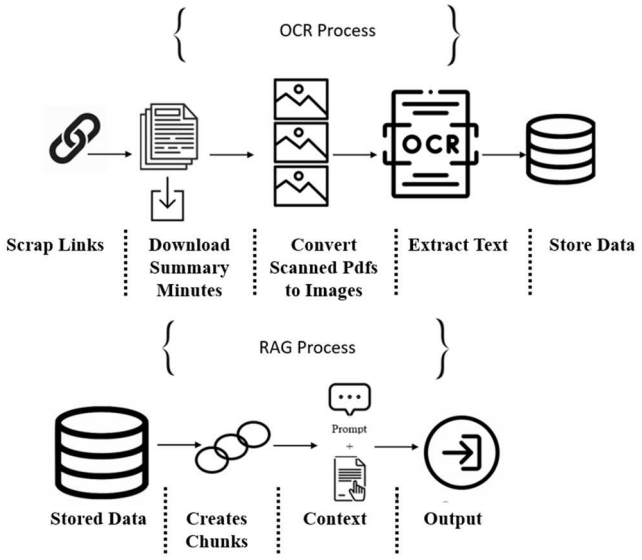

Figure 2: "Summary Minutes" Information Extraction

Table 3: Comparative Accuracy of OCR Tools with Applied Preprocessing Techniques

OCR Tool	Accuracy before pre-processing (%)	Accuracy after pre-processing (%)	Remarks
PaddleOCR	88.3	95.2	Consistent with printed text
Tesseract	82.7	90.5	Struggles with noise in documents
EasyOCR	85.1	92.3	Better performance with low-noise
Google Cloud Vision	93.4	97.1	High accuracy but costly solution

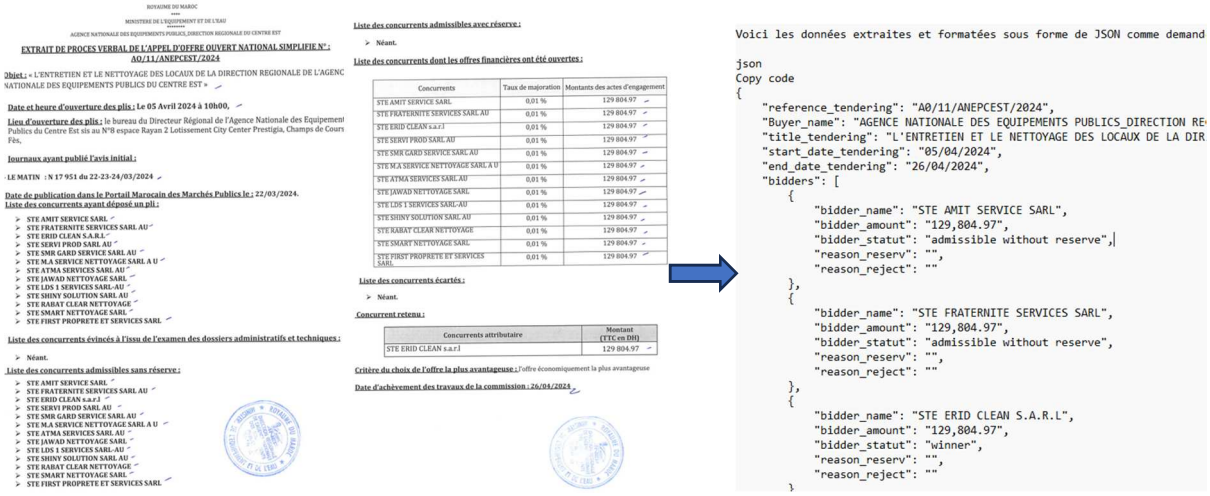

Figure 3: Example of a Summary Minutes and the Pipeline Output Related to the Example

F. Limitations Encountered

While implementing the data extraction process using PaddleOCR, we encountered several limitations inherent to the tool, which affected our ability to process approximately 15,000 documents. One significant challenge was the sensitivity of PaddleOCR to image quality and resolution. Documents with low resolution or poor image quality resulted in inaccuracies during text recognition, leading to errors in data extraction. Despite attempts to enhance image quality through preprocessing techniques, a subset of documents remained challenging to process, resulting in their exclusion from the analysis. Additionally, we observed variations in PaddleOCR's performance across different languages and scripts. While it demonstrated high accuracy in recognizing text in documents written in French, we encountered challenges with documents written in Arabic.

III. Conclusion

In this study, we have presented a methodology for extracting and structuring procurement data from scanned Summary Minutes documents obtained from the Moroccan Public Procurement Portal. Through the utilization of web scraping techniques, Optical Character Recognition (OCR) with PaddleOCR, and the RAG (Retrieval Augmented Generation) process with the Mistral LLM (Large Language Model), we successfully extracted 439,048 records, providing valuable insights into procurement practices. Overall, the methodology presented in this study lays the foundation for future research in procurement analysis and data extraction from unstructured sources. By continuing to refine and improve data extraction techniques and conducting in-depth analysis of extracted datasets, researchers can contribute to the advancement of knowledge in procurement practices and support evidence-based decision-making in public procurement.

References

[1] P. Kumar, A. Tveritnev, S. A. Jan, and R. Iqbal, "Challenges to Opportunity: Getting Value Out of Unstructured Data Management," in *Day 2 Tue, March 14, 2023*, SPE, Mar. 2023. doi: 10.2118/214251-MS.

[2] E. I. Voevodina, D. A. Prytyka, Y. M. Gulyaeva, A. E. Emelyanova, and D. E. Varakhtin, "PROBLEMS OF PROCESSING AND USE OF SEMI-STRUCTURED AND UNSTRUCTURED DATA IN THE MANAGEMENT OF PUBLIC ORGANIZATIONS," *EKONOMIKA I UPRAVLENIE: PROBLEMY, RESHENIYA*, vol. 1/2, no. 133, pp. 163–167, 2023, doi: 10.36871/ek.up.p.r.2023.01.02.022.

[3] P. A. Bowen, P. J. Edwards, and K. Cattell, "Corruption in the South African construction industry: a thematic analysis of verbatim comments from survey participants," *Construction Management and Economics*, vol. 30, no. 10, pp. 885–901, Oct. 2012, doi: 10.1080/01446193.2012.711909.

[4] M. Urazaliev, "ISSUES OF PREVENTION OF CORRUPTION CRIMES IN THE FIELD OF PUBLIC PROCUREMENT," *European International Journal of Multidisciplinary Research and Management Studies*, vol. 02, no. 08, pp. 1–6, Aug. 2022, doi: 10.55640/eijmrms-02-08-01.

[5] S. Z. S. Tabish and K. N. Jha, "Analyses and evaluation of irregularities in public procurement in India," *Construction Management and Economics*, vol. 29, no. 3, pp. 261–274, Mar. 2011, doi: 10.1080/01446193.2010.549138.

[6] J. Lester, "How to Minimise Corruption in Public Works Construction Contracts," *J Financ Crime*, vol. 7, no. 2, pp. 161–169, Apr. 1999, doi: 10.1108/eb025934.

[7] T. Peter, *Fraud and Corruption in Public Services*. Routledge, 2017. doi: 10.4324/9781315093949.

[8] A. M. Cheredarchuk, "Socially Dangerous Consequences Of Criminal Offenses In The Field Of Public Procurement," *Actual problems of improving of current legislation of Ukraine*, no. 55, pp. 120–129, Jan. 2021, doi: 10.15330/apiclu.55.120-129.

[9] V. Sukhonos, L. Pavlenko, O. Krukhmal, A. Ivanovska, and D. Maletov, "Forms of committing corrupt abuses of public finances and ways to counteract them in Ukraine," *Revista Amazonia Investiga*, vol. 10, no. 39, pp. 149–158, May 2021, doi: 10.34069/AI/2021.39.03.14.

[10] A. Borowiec, "Corrupt Practices in Public Procurement: Evidence from Poland," 2019, pp. 71–82. doi: 10.1007/978-3-030-18565-7_6.

[11] M. Lima, R. Silva, F. Lopes de Souza Mendes, L. R. de Carvalho, A. Araujo, and F. de Barros Vidal, "Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1580–1588. doi: 10.18653/v1/2020.findings-emnlp.143.

[12] A. Yeghiazaryan, K. Khechoyan, G. Nalbandyan, and S. Muradyan, "Tokengrid: Toward More Efficient Data Extraction From Unstructured Documents," *IEEE Access*, vol. 10, pp. 39261–39268, 2022, doi: 10.1109/ACCESS.2022.3164674.

[13] R. K., H. Srinivas, and S. S., "Industrial information extraction through multi-phase classification using ontology for unstructured documents," *Comput Ind*, vol. 100, pp. 137–147, Sep. 2018, doi: 10.1016/j.compind.2018.04.007.

[14] D. Baviskar, S. Ahirrao, and K. Kotecha, "Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches," *IEEE Access*, vol. 9, pp. 101494–101512, 2021, doi: 10.1109/ACCESS.2021.3096739.

[15] B. Oral, E. Emekligil, S. Arslan, and G. Eryiğit, "Information Extraction from Text Intensive and Visually Rich Banking Documents," *Inf Process Manag*, vol. 57, no. 6, p. 102361, Nov. 2020, doi: 10.1016/j.ipm.2020.102361.

[16] M. M. Mirończuk, "Information Extraction System for Transforming Unstructured Text Data in Fire Reports into Structured Forms: A Polish Case Study," *Fire Technol*, vol. 56, no. 2, pp. 545–581, Mar. 2020, doi: 10.1007/s10694-019-00891-z.

[17] H. Tissot, M. D. Del Fabro, L. Derczynski, and A. Roberts, "Normalisation of imprecise temporal expressions extracted from text," *Knowl Inf Syst*, vol. 61, no. 3, pp. 1361–1394, Dec. 2019, doi: 10.1007/s10115-019-01338-1.

[18] D. Vukadin, A. S. Kurdija, G. Delac, and M. Silic, "Information Extraction From Free-Form CV Documents in Multiple Languages," *IEEE Access*, vol. 9, pp. 84559–84575, 2021, doi: 10.1109/ACCESS.2021.3087913.

[19] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, and G. Ceder, "Opportunities and challenges of text mining in materials research," *iScience*, vol. 24, no. 3, p. 102155, Mar. 2021, doi: 10.1016/j.isci.2021.102155.

[20] G.-S. Lin, S.-K. Chai, H.-M. Li, and J.-Y. Lin, "Vision-based patient identification recognition based on image content analysis and support vector machine for medical information system," *EURASIP J Adv Signal Process*, vol. 2020, no. 1, p. 27, Dec. 2020, doi: 10.1186/s13634-020-00686-3.

[21] S. Idrees and H. Hassani, "Exploiting Script Similarities to Compensate for the Large Amount of Data in Training Tesseract LSTM: Towards Kurdish OCR," *Applied Sciences*, vol. 11, no. 20, p. 9752, Oct. 2021, doi: 10.3390/app11209752.

[22] S. Rakshit, A. Kundu, M. Maity, S. Mandal, S. Sarkar, and S. Basu, "Recognition of handwritten Roman Numerals using Tesseract open source OCR engine ," in *Proc. Int. Conf. on Advances in Computer Vision and Information Technology* , 2009, pp. 572–577.

[23] D. Sporici, E. Cușnir, and C.-A. Boiangiu, "Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing," *Symmetry (Basel)*, vol. 12, no. 5, p. 715, May 2020, doi: 10.3390/sym12050715.

[24] J. Cai, E. Sun, and Z. Chen, "OCR Service Platform Based on OpenCV," *J Phys Conf Ser*, vol. 1883, no. 1, p. 012043, Apr. 2021, doi: 10.1088/1742-6596/1883/1/012043.

[25] V. Kumar, P. Kaware, P. Singh, R. Sonkusare, and S. Kumar, "Extraction of information from bill receipts using optical character recognition," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, IEEE, Sep. 2020, pp. 72–77. doi: 10.1109/ICOSEC49089.2020.9215246.

[26] Florentinus Budi Setiawan, F. Adriantama, Leonardus Heru Pratomo, and Slamet Riyadi, "Improving AI Text Recognition Accuracy with Enhanced OCR For Automated Guided Vehicle," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 728–734, Oct. 2022, doi: 10.29207/resti.v6i5.4279.

[27] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey," *ACM Comput Surv*, vol. 55, no. 6, pp. 1–28, Jun. 2023, doi: 10.1145/3530811.

[28] K. Han *et al.*, "A Survey on Vision Transformer," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[29] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Comput Surv*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[30] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis Support Syst*, vol. 115, pp. 24–35, Nov. 2018, doi: 10.1016/j.dss.2018.09.002.

[31] M.-T. Nguyen, D. T. Le, and L. Le, "Transformers-based information extraction with limited data for domain-specific business documents," *Eng Appl Artif Intell*, vol. 97, p. 104100, Jan. 2021, doi: 10.1016/j.engappai.2020.104100.

[32] C. Li *et al.*, "PP-StructureV2: A Stronger Document Analysis System," Oct. 2022.