AI - based digitization of legacy ground information

Jaume Llopart [1*], *Parichehr* Behjati [1], *Ignasi* Aliguer [1]

[1]*SAALG Geomechanics, Barcelona Spain*
[*]*Corresponding author: jaume.llopart@saalg.com*

ABSTRACT

Geotechnical characterization of site materials is of paramount importance in the construction and mining industry. The analysis of large volumes of geotechnical information from multiple sources leads to data-driven decisions that help to minimize uncertainty. For this purpose, a unified digital information platform becomes handy to have a global perspective and improve the analysis of available ground information data.

Access to historic ground investigation data from previous projects during the project planning stage might increase efficiency. However, accessing and processing legacy data from companies' databases is time and resources consuming. In the recent years, software tools that are capable of extracting data in a digital format from images have become popular, but still require human-supervised interpretation.

A novel tool combining Optical Character Recognition (OCR), digital data extraction technologies and AI-based data interpretation system is presented herein. The state-of-the-art OCR technology is capable of accurately recognizing and extracting text from various document types, such as scanned documents, images, and PDFs. It utilizes advanced machine learning algorithms to process text, even in challenging conditions, ensuring data is extracted accurately and reliably. Then, a data interpretation system has been trained to identify the type of site characterization data and its structure while retrieving all the content in a digital format. All components work seamlessly together to provide a comprehensive solution for automating the interpretation and extraction of site characterization data, streamlining data management and analysis processes.

The capability of gathering data from multiple sources in a unique ground information system provides valuable information for planning and design stages while decreasing costs, time and uncertainties. In addition, all these data are then available within DAARWIN platform to feed the ground model workflow.

Keywords: borehole data; digitization; image recognition; OCR.

1. Introduction

The geomechanical characterization of site materials plays a crucial role in the fields of construction and environmental engineering, with a direct impact on project planning, execution, and sustainability. Traditional methods of geotechnical data acquisition and analysis rely on the assimilation of data from diverse sources, often presenting challenges in data interpretation and integration. In recent years, the integration of digital technologies has reshaped the process of geotechnical data acquisition, analysis, and interpretation, offering a global perspective and facilitating data analysis.

Considering data from multiple sources, and even from different instants in time, contributes to a better site characterization, and the assimilation of entire geotechnical reports presents an opportunity for comprehensive analysis and data-driven decision-making. However, the analysis of large volumes of geotechnical information, including entire reports, has historically been a time-consuming and resource-intensive endeavour.

The integration of legacy data from previous field works has proven to be instrumental in reducing costs, minimizing environmental impact, and increasing the sustainability of construction activities. However, the incorporation of entire geotechnical reports into contemporary databases has posed significant time and resource-related challenges. With the objective of satisfying the need for automated data processing solutions capable of processing entire reports to solve these challenges, this study introduces a novel tool that combines Optical Character Recognition (OCR), advanced data extraction technologies, and a state-of-the-art AI-based data interpretation system to process a wide range of ground information such as: borehole logs, in-situ tests results and even entire geotechnical reports. The developed tool integrates cutting-edge OCR technology, which has demonstrated remarkable proficiency in accurately recognizing and extracting text from a spectrum of document types, including scanned documents, images, and PDFs. Leveraging advanced machine learning algorithms (Sarker 2021), the OCR technology ensures precise and reliable data extraction, even in challenging conditions such as low-resolution images, noisy backgrounds, distorted fonts, or handwritten texts. Complementing the OCR technology,

the AI-based data interpretation system has been meticulously trained

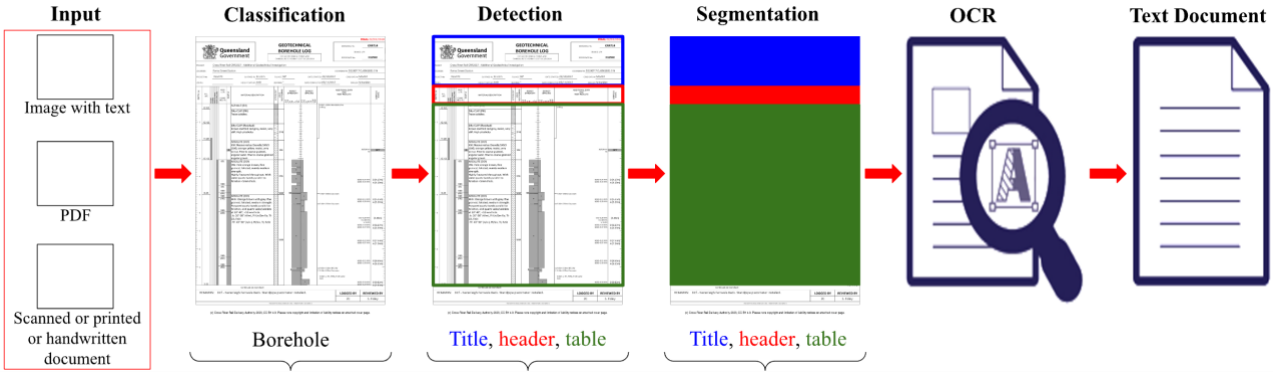

Figure 1. The overall framework of the digitalization process.

to discern the nuances of site characterization data and retrieve structured information from entire reports in digital formats.

The seamless integration of these components offers a comprehensive solution for automating the interpretation and extraction of site characterization data, including entire geotechnical reports, thus streamlining data management and analysis processes. The capability of processing entire geotechnical reports within a unified ground information system presents a transformative shift in geotechnical data management, offering valuable insights for project planning and design, and contributing to cost reduction, time efficiency, and the mitigation of uncertainties in geotechnical projects.

In this article, we present a pioneering approach to automated geotechnical data extraction and interpretation, underscoring the potential for transformative advancements in geotechnical engineering and environmental sustainability through the digitization of ground information.

2. Digitalization algorithm

In this section, a comprehensive AI-based digitization pipeline to turn analog documents into digital formats, illustrated in The overall framework of the digitalization process.Figure 1, is described. This pipeline encompasses several key stages, each leveraging cutting-edge deep learning methodologies to achieve precise and efficient document processing.

2.1. Image classification

In the initial phase of the digitization process, we employ deep learning methodologies for image classification, a pivotal step in discerning and categorizing various components within documents. Leveraging state-of-the-art convolutional neural networks (CNNs) (He et al. 2015; Simonyan 2014) alongside other advanced architectures (Reis et al. 2023), our system undergoes rigorous training to achieve precise recognition and classification of distinct document sections. Our neural network is trained on an extensive dataset comprising over 20,000 images for each class, covering a diverse range of materials such as borehole logs, test reports, and other relevant documents. Employing diverse evaluation metrics including accuracy, precision, and recall, we meticulously assessed the model's performance. Through extensive exposure to varied datasets, our deep learning models develop a nuanced understanding of visual features that distinguish different types of information. This comprehensive classification framework not only establishes a foundation for subsequent processing steps but also enables accurate identification and segregation of document elements for further analysis.

2.2. Object detection

Following the initial classification phase, the digitization process progresses with deep-learning techniques for object detection. This stage involves identifying various elements within the documents, such as tables, text blocks, plots, and other relevant entities. We apply transfer learning by leveraging pretrained models like region-based convolutional neural networks (R-CNNs) (Girshick 2014) and You Only Look Once (YOLO) (Reis 2023). These models are adapted using our proprietary dataset, which comprises over 70,000 images. This transfer learning process allows us to tailor the models to our specific task, thereby improving their accuracy and performance in detecting diverse document elements. The system accurately pinpoints and localizes specific entities, with the output being the bounding boxes of the detected objects. Through ongoing refinement and optimization, the object detection module ensures comprehensive coverage and precise detection of various document elements, ultimately enhancing the overall efficiency and reliability of the digitization process.

2.3. Image segmentation

Segmentation is a crucial stage in the digitization pipeline, where sophisticated algorithms such as semantic segmentation and instance segmentation are utilized to accurately delineate and extract specific regions of interest from the identified document components. Employing the same training process and dataset as the object detection section, the system divides the document into cohesive segments corresponding to tables, text passages, and other relevant entities, utilizing cutting-

edge techniques such as DeepLab (Liang-Chieh, 2018). Through examination of visual cues and contextual information, these segmentation algorithms enable precise isolation of target regions, ensuring that only pertinent information is extracted for further processing. By seamlessly integrating segmentation into the digitization workflow, the system achieves a high level of accuracy and fidelity in capturing the desired content.

2.4. Optical character recognition

Following the segmentation phase, the digitization process integrates Optical Character Recognition (OCR), a cornerstone in document digitization. OCR technology enables the conversion of handwritten, typed, scanned text, or text within images into machine-readable text, applicable to various file formats. Leveraging advanced OCR algorithms and deep learning models, as documented by Ray (2007) and Du et al. (2020), the system adeptly transforms scanned or photographed documents into machine-readable text. This relies on neural network architectures like recurrent neural networks (RNNs) and transformers. In our work, we trained our OCR model with more than 50,000 text images. The significance of this transformation extends beyond conversion; it ensures the preservation of the original document's integrity and fidelity while facilitating downstream natural language processing tasks. Consequently, OCR bridges the gap between physical and digital realms.

2.5. Natural language processing

After the Optical Character Recognition (OCR) stage, the digitization process proceeds to the Mapper stage, where the output from OCR is processed further. In the Mapper stage, we leverage Natural Language Processing (NLP) techniques for text understanding and classification. Using deep learning architectures such as transformer-based models like BERT (Devlin et al. 2018) and GPT (OpenAI 2024), the system comprehensively analyzes and interprets the digitized borehole logs and other textual data. Through advanced semantic parsing, entity recognition, and classification methodologies, the NLP module extracts valuable insights, discerns patterns, and categorizes information embedded within the digitized text. This enables actionable intelligence, facilitates informed decisions, and derives meaningful insights from digitized documents. Our NLP model has been trained on more than 5000 classified texts, covering diverse domains and topics, ensuring robustness and versatility in text classification tasks. This extensive training corpus enhances the model's ability to generalize across different datasets and effectively classify text documents, regardless of their specific domain or subject matter.

3. Data mapping

The variety of possible layout formats for presenting the geotechnical and geological information in a report is enormous. To address this challenge, an unsupervised data mapping approach is required, wherein the software dynamically adapts to recognize structures and patterns, extracting the right information from the right place. The implementation outlined below employs a decision tree methodology. It takes machine-readable digital text extracted from the report (Section 2) as input and generates a digital representation of the enclosed elements in the report as illustrated in Figure 2. This methodology relies on a predefined structure to direct the processing and analysis of data. It allows the software to adapt and identify patterns for data extraction, leveraging the hierarchical organization of the acquired data.

The data mapping module comprises specialized pipelines, each designed to perform distinct operations tailored to the specifications of the guiding structure. The output generated by these pipelines is stored in a structured JSON file, facilitating efficient organization and retrieval of the extracted data.

3.1. Tailored pipelines

Upon classification of pages by the automatic identification algorithm, the data mapping algorithm, henceforth referred as the ‘mapper’, directs the data to specific sub-processes accordingly. As such, two primary pipelines have been developed: a) borehole data, b) laboratory and in situ test data.

Figure 2. Automatic detection of each column type in a borehole log.

3.1.1. Borehole interpretation

The borehole interpretation pipeline receives data extracted by the OCR algorithm. Given that geotechnical reports commonly feature multiple boreholes or Ground Information Points (GIPs), the initial step involves grouping all detected information pertaining to each borehole or GIP. To accomplish that, the information from each borehole is cross-checked such as the borehole name, but also other properties such its location, drilling date, total depth, etc. (Figure 3). Once the grouped information referring to a single borehole is gathered, the multiple pages of the same borehole are merged to obtain a continuous object. Before extracting the borehole data *per se*, the different groups of information present in the borehole log are identified. Thus, the mapper extracts the following data: depths and description of the identified

layers, samples and types of samples retrieved and whether laboratory tests have been performed on these samples or whether Standard Penetration Tests (SPT) have been recorded. These data can be presented in multiple formats, and this is the main strength of the mapper, to recognize the type of the data present in the borehole log and classify it before its extraction. The most complex interpretation for the mapper is the layer geometry and description pairing, since sometimes there is no record of the layers' depth boundaries in the borehole log (4a). Thus, the mapper can

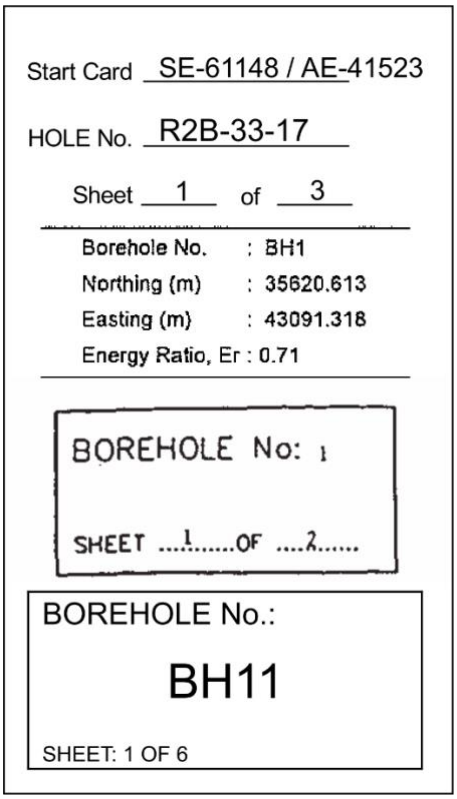

Figure 3. Four examples of borehole log header information that mapper needs to be able to detect and extract the borehole name from them.

transform the lines between layers description to depth values. Furthermore, given the stacking of descriptions, especially in thin layers, the lines division between layers can be uneven. However, the mapper is adept at determining the actual depth to which the bottom of each layer corresponds in the log (4b). To match the layer description to a material type and show it within DAARWIN platform, the NLP algorithm described previously is used to extract from each description the matching material type.

Regarding the possible SPTs, sample records, and laboratory tests in the borehole log, their depth is calculated if no reference is found in the log. Depending on the structure of the SPT data, the mapper algorithm can detect what standard is used for the SPT test procedure. Also, the mapper detects automatically if the measurements are in the IS or imperial metric system and transform accordingly to the metric system used to be visualized within DAARWIN platform.

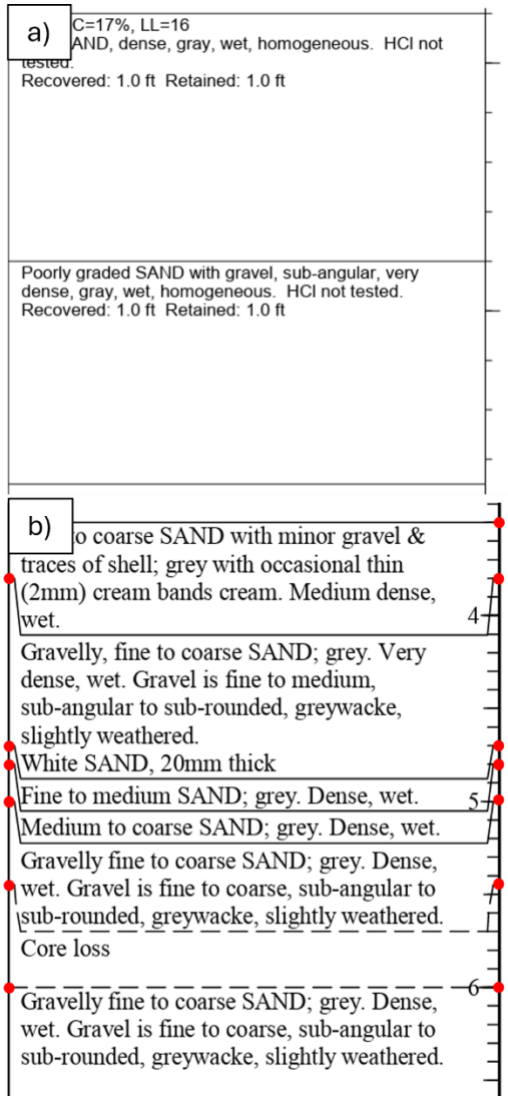

Figure 4. a) Layer description boundaries without depth data. b) Layer description boundaries with uneven bottom lines. Red dots mark the verge of the uneven lines where the mapper detects the real layer bottom.

3.1.2. Laboratory and in situ test interpretation

Typically, factual reports contain pages dedicated to laboratory and in situ tests, which often feature plots, images, and tables in various formats. Occasionally, they also present the primary outcomes of these tests. In this context, the mapper initially categorizes each page labeled as tests by utilizing an image classification algorithm. Nowadays, most test reports are already in digital format. However, older reports are likely to consist of scanned pages, rendering the information on these pages as images. An algorithm discerns the page's nature, directing digital pages to a digital data extraction pipeline and image pages to an OCR algorithm for text retrieval. In both scenarios, the images and figures from each page are extracted and subsequently linked to the corresponding Geotechnical Information Point (GIP) in DAARWIN.

After classifying each page and detecting the type of tests, the pages are processed by a dedicated driver. In general terms, these drivers look for key words in the

tables depending on the test type and language. For the table detection within the page, the drivers automatically expand the search area until the desired data is obtained.

All data retrieved by each driver is then stored in a purpose-built data structure and transmitted back to the mapper. Upon receiving data from various pipelines and drivers, the mapper organizes it into a comprehensive data structure, which is stored for future review by the end user.

4. Accuracy and Processing performance

Addressing the reliability of results obtained through AI techniques is a significant concern. To tackle this challenge, an automatic test pipeline has been developed to verify the accuracy of processed data. This pipeline also ensures compatibility between new implementations and previous datasets. While AI algorithms inherently provide accuracy scores, it is worth to mention that the mapper currently lacks this capability.

To assess the accuracy of the mapper's output, a pool containing various borehole log formats has been established as templates. The accuracy is evaluated by comparing the mapper's output to the templates of each borehole log, key by key within the JSON data structure. This meticulous comparison ensures a comprehensive examination of the mapper performance across different log formats.

Subsequently, we calculate accuracy (A) using the following equation:

$$A = \frac{\kappa_t - \sum \kappa_i}{\kappa_t} \quad (1)$$

Where κ_t is the total number of keys in the JSON data structure and $\Sigma\kappa_i$ is the count of key values which differ from the reference template.

The processing performance stands out as another crucial aspect of such implementations. While manually inputting a new borehole log into the system might take minutes, the pipeline described in this paper can process each log page in under thirty seconds. Consequently, the system can handle a one hundred pages geotechnical report in less than an hour.

5. Conclusions

The integration of digital technologies has significantly reshaped the process of geotechnical data acquisition, analysis and interpretation facilitating data analysis. The algorithmic pipeline presented in this manuscript introduces a novel solution that combines optical character recognition, advanced data extraction technologies, and a state-of-the-art AI-based data interpretation system to process entire geotechnical reports. This transformative shift in geotechnical data management offers valuable insights for project planning and design, contributing to cost and material reduction and time efficiency. The comprehensive AI-based digitization pipeline described above demonstrates a pioneering approach to automate geotechnical data extraction and interpretation, underscoring the potential of incorporating legacy data into current digital workflows.

This AI-based software digitization pipeline includes image classification, object detection, image segmentation, OCR, natural language processing, and data mapping. These stages and algorithms have been rigorously developed and trained to achieve precise and efficient document processing, with the system demonstrating the ability to process a geotechnical report of a hundred pages in less than an hour, significantly improving the overall efficiency and reliability of the digitization process.

Acknowledgements

The authors are grateful for the financial support provided by the Torres Quevedo grant from the Agencia Estatal de Investigación (AEI) under Grant Agreement No. PTQ2022-012630 and the European Innovation Council (EIC) through the Project GEORGIA - 190151860.

References

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. 801-818. https://doi.org/10.1007/978-3-030-01234-2_49

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., et al & Wang, H. 2020. Pp-ocr: A practical ultra lightweight ocr system. https://doi.org/10.48550/arXiv.2009.09941

Girshick, R., Donahue, J., Darrell, T., & Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. 580-587. 10.1109/CVPR.2014.81

He, K., Zhang, X., Ren, S., & Sun, J. 2015. Deep residual learning for image recognition. 770-778.10.1109/CVPR.2016.90

OpenAI. (2024). ChatGPT (3.5) https://chat.openai.com

Reis, D., Kupec, J., Hong, J., & Daoudi, A. 2023. Real-Time Flying Object Detection with YOLOv8. https://doi.org/10.48550/arXiv.2305.09972

Sarker, I.H. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 https://doi.org/10.1007/s42979-021-00592-x.

Simonyan, K., & Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. 1-14. https://doi.org/10.48550/arXiv.1409.1556

Smith, R. An overview of the Tesseract OCR engine. 2007. vol. 2, 629-633. 10.1109/ICDAR.2007.4376991