TOWARDS THE DEVELOPMENT OF A HARMONIZED INVENTORY DATABASE FOR DECISION SUPPORT: AUTOMATIZED INFORMATION EXTRACTION

Alla SAPRONOVA [1], Paul J. UNTERLASS [1], Vaibhav SHRINGI [1] Thomas MARCHER[1]

[1] Institute of Rock Mechanics and Tunnelling, Graz University of Technology, Graz, Austria

Introduction

Data analysis in civil engineering is important and widely used to ensure high safety standards, optimize the construction process's time and cost, and accelerate the transition to Industry 4.0. To utilize all available information and enable efficient data management in engineering, there is a need to extract structural data from archived geotechnical documentation. Most documents in civil engineering, especially those created in the early- to mid-digitalization era, are stored as papers or files containing unstructured information (e.g., scanned documents). Such documents may include: handwritten notes, engineering sketches, printed plots, and photos. At the same time, the form and completeness of the documents vary from one project to another. Therefore, extracting, structuring, and harmonizing information from civil engineering projects via intelligent document parsing seems essential for efficient data acquisition.

Information extraction and integration

The form and content of geotechnical and tunnel documentation vary by year, region, and project. Each part of a technical document may contain various pieces of information. All available information shall be extracted and integrated with other data from the same project to form a complete dataset for one project. Figure 1 illustrates the significant steps towards such a database. Data harmonized this way can then be used for analysis or further enriched with other datasets and used for decision support.

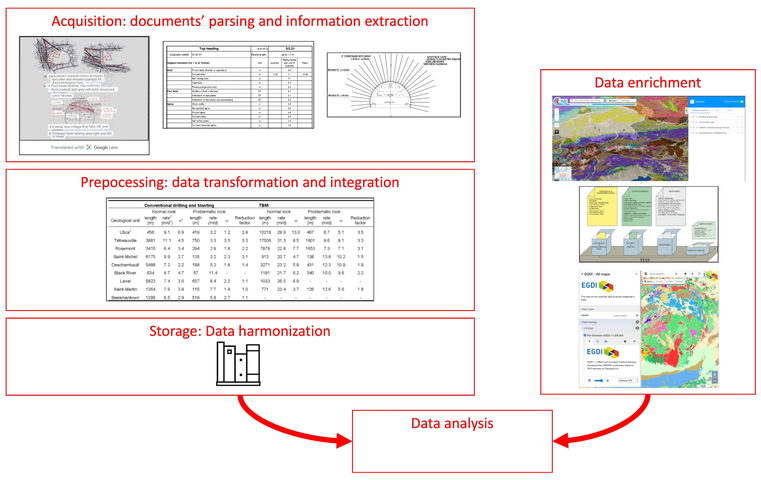

Figure 1. Shows the pathway towards a harmonized technical documentation database: from data acquisition to preprocessing and transformation. Harmonized data can be further enriched and used for decision support.

While significant efforts have already been made in collecting geological information and tunnel data from recent construction projects, there is still an issue with obtaining data from archives where the documents are only available as hard copies (in paper format).

The process of data acquisition from a scan of a document is reflected in Figure 2. It shall include several or all of the following processes: (A) translation, if the language is different from English; (B) text extraction; (C1) information extraction and integration from image and (C2) handwriting; (D) tabular information extraction and verification from images; (E) natural language processing; (F) converting images of tabular data to a table; (G1) information extraction/integration from images of tabular data and (G2) plots.

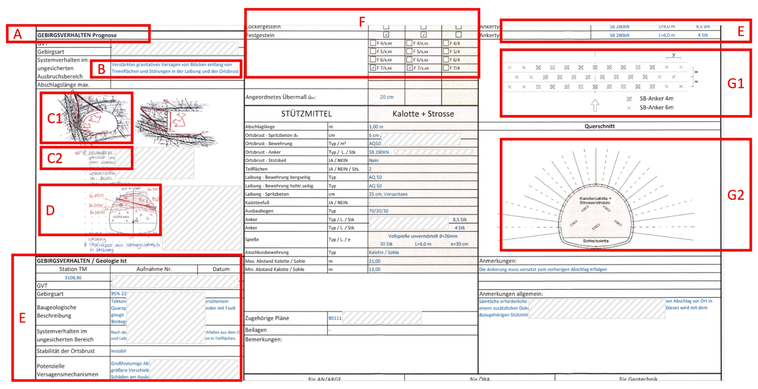

Figure 2. Shows a sample of a technical document containing information in different forms: images of text characters (A, B, E), images of tables (E, F), images of handwriting (C2, D), sketches (C1, D, G2), images of categorical data (G1), and images of plots (G2).

The first step is to translate the document. There are a number of existing solutions for optical character recognition, including solutions that can extract and translate text from images simultaneously (e.g., Google Cloud ML Tools). These tools can also be applied to handwriting. But in this case, the translated text would require post-processing, where natural languages or expressions are used. Therefore, natural language processing (NLP) models may be better suited for this task. The field of NLP and document parsing is well-developed, and a number of available solutions (paid and open access) can be found.

After the text and tables are extracted and structured, the image processing step can take place.

Image Processing

The images occurring in archived documentation can belong to one of three main classes: plots, sketches, or photos.

Restoring plots from images to numerical values has been solved using the Hough transform method (Duda & Hart, 1972) and density based spatial clustering (DBSCAN) (Ester et al., 1996) to detect lines and points. With the Hough method the idea is to use automatic image thresholding to filter large non-connecting objects. These methods are realized in, e.g., the Open Source Computer Vision library (OpenCV) (Bradski, 2000). Figure 3 shows the digitization procces for scans of geodetic monitoring results (3a.). In Fig 3b.) the DBSCAN method has been used to restore the cyan coloured measurements and in 3c.) the Hough transform method was used for the green measurements.

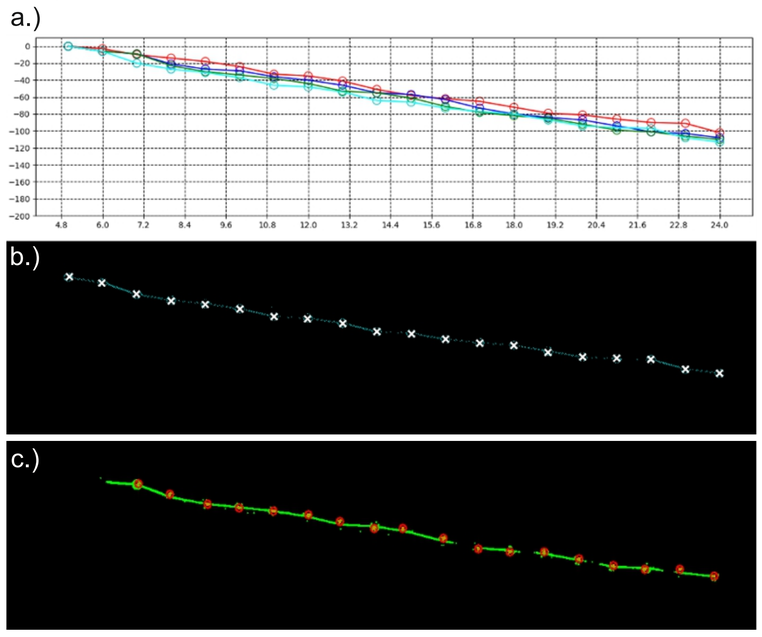

Figure 3. Shows the restoration of scanned geodetic monitoring results. a.) Scanned image of the monitoring results. b.) DBSCAN method to restore the cyan coloured measurements and c.) Hough transform method to restore the green coloured measurements.

The accuracy of the data extraction depends on image resolution and separation of pixels to be detected.

Conclusion

The core idea behind this work was to provide users regardless of their expertise level with a flexible and scalable tool that quickly integrates complex information stored in technical documentation as hard copy format. Such a tool would offer the possibility to enhance decision support during construction. Furthermore, the harmonization of data stored this way, would support its exploration (down-drilling) and enable the provision of tailored information for specific use.

By retrieving, structuring, and building links between pieces of data, as proposed in this work, the increasingly important yet hard-to-reach transition from information to knowledge can be addressed. The authors see the proposed approach as a major step towards improving cross-site knowledge transfer and information flow. One use case could be the comparative analysis of tunnel data from different sites, which could help in the future selection of effective strategies during project planning, execution, and maintenance.

DaVinci Project

This work is part of the Data Advance via INtelligent Content Integration (DaVinci) project. Initiated by RMT, under the leadership of Dr. Sapronova and Prof. Marcher, the project aims to develop a toolkit for precise data management in civil engineering and the advanced processing of technical documentation. In 2023, the DaVinci project earned recognition by being listed in the top 100 projects addressing issues related to the 17 United Nations Sustainable Development Goals, as categorized by IRCAI as an 'Early Stage Project.'

Currently, the project is pivoting towards leveraging cutting-edge technologies like transformers and Large Language Models (LLMs) to enhance image description capabilities. This new focus aims to further refine data interpretation and contribute to more effective and sustainable solutions in civil engineering.

References

Bradski, G., 2000. The OpenCV Library. Dr. Dobb's J. Softw. Tools.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, pp. 226–231).

Duda, R. O., & Hart, P. E. (1972). Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM, 15(1), 11–15.

Contact presenting author: unterlass@tugraz.at

ISRM · ÖGG · AUSTRIAN SOCIETY FOR GEOMECHANICS · RMT · TU Graz · MLGT · NAWI Graz GEOZENTRUM