Towards the development of a harmonized inventory database for decision support: automatized information extraction

Alla Sapronova
Institute of Rock Mechanics and Tunnelling, Graz University of Technology, Graz, Austria

Paul Johannes Unterlass
Institute of Rock Mechanics and Tunnelling, Graz University of Technology, Graz, Austria

Vaibhav Shringi
Institute of Rock Mechanics and Tunnelling, Graz University of Technology, Graz, Austria

Thomas Marcher
Institute for Rock Mechanics and Tunnelling, Graz University of Technology, Graz, Austria

ABSTRACT: Decisions made during tunnel construction are ultimately always based on the opinion of safety-oriented engineers and therefore utilize the knowledge of the employed human experts in the best possible way. Because every tunnel is unique to some extent, it can be assumed that experts' decisions are often "reinvented" on-site at short notice. Given the number of completed, ongoing, and planned tunnel projects in Europe and worldwide, it is possible to identify projects which could be used as an extra reference to assist the decision-making processes for new constructions. For human experts, though, it is difficult and time-consuming to identify all similar reference projects. Aiming at developing a harmonized inventory database for decision support (DS) during the tunnel's planning and construction phases, this work discusses a pathway for retrieving and processing data from archived projects. The major steps for information extraction are described, and the process of developing a harmonized database based on archived documentation is discussed. This work specifically addresses the process of extracting tabular information from images using machine learning methods.

Keywords: data analysis, information extraction, technical documentation, decision support.

1 INTRODUCTION

Data analysis in civil engineering is important and widely used to ensure high safety standards, optimize the time and cost of the construction process, leverage the principles of sustainable logistics, and overall accelerate the transition to Industry 4.0. In order 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 e.g., handwritten notes, engineering sketches, printed plots, and photos. Whereas 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 to be essential for efficient data acquisition. Extracted data can be used, e.g., for comparative analysis: to identify and further explore projects with similar conditions worldwide.

2 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, even when the documentation is from the same project. All available information shall be extracted and integrated with other data from the same project to form a complete *dataset* for one particular project. The harmonization of the data from *multiple* datasets helps to form a *database* that can be used for comparative analysis and data-driven modeling. Figure 1 illustrates the major steps that shall be taken prior to site comparison: data acquisition, integration, and harmonization. Data harmonized this way can then be utilized for analysis or can be further enriched with other GIS datasets and used for decision support.

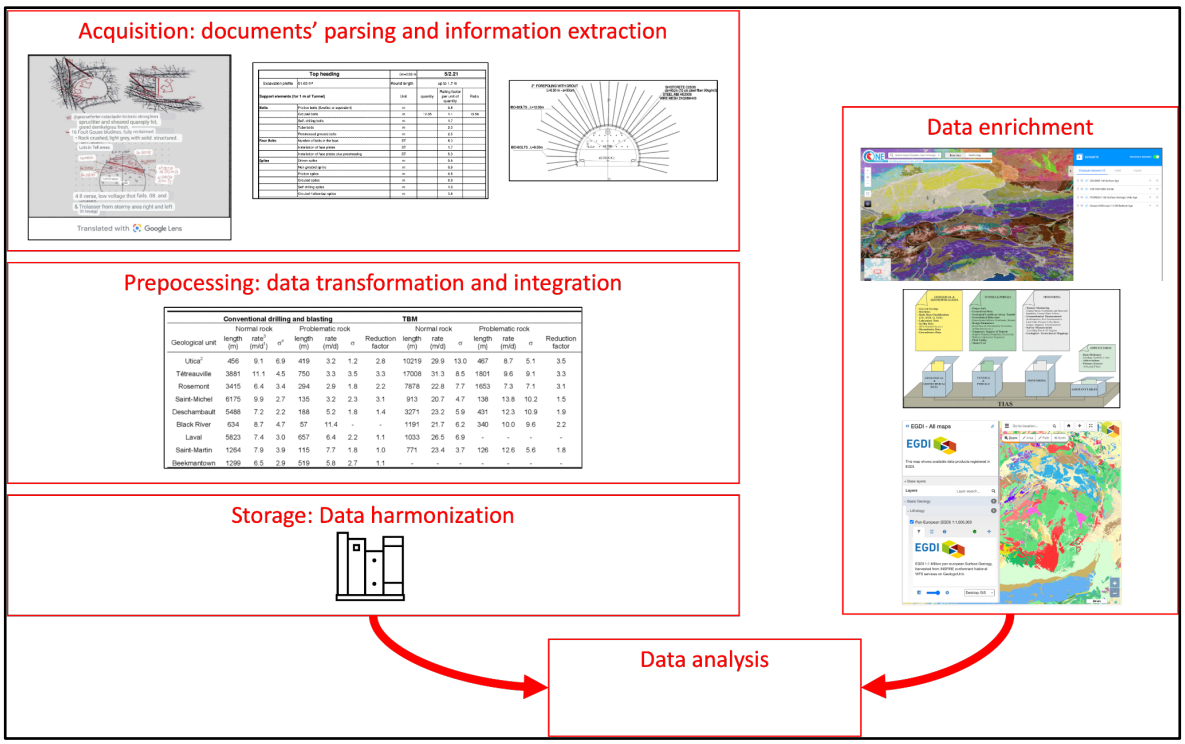

Figure 1. Shows the pathway towards a harmonized database of technical documentation: from data acquisition to preprocessing and transformation. Harmonized data can be further enriched and used for decision support. (For illustration, here used images from Aygar, 2021, Leroux, 2018, Austrian standards, OneGeology, Europe Geology, and Marinos, 2013).

While great efforts have already been made in collecting geological information and tunnel data from recent construction projects (e.g., the Tunnel Information and Analysis System (TIAS) project proposed by Marinos et al., 2013), 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 scanned copy of a document is reflected in Figure 2 and shall include several or all of the following processes: translation if the language is different from English (A), text extraction (B), information extraction and integration from image (C1) and handwriting (C2), tabular information extraction and verification from images (D), natural language

processing (E), converting images of tabular data to a table (F), information extraction and integration from images of tabular data (G1) and plots (G2).

To ensure the correct interpretation of data types and accurate categorization of the technical documentation (e.g., documentation on the selected support, face mapping, and description), 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 at the same time (e.g., Google Cloud ML Tools); hence this part of preprocessing seems relatively trivial.

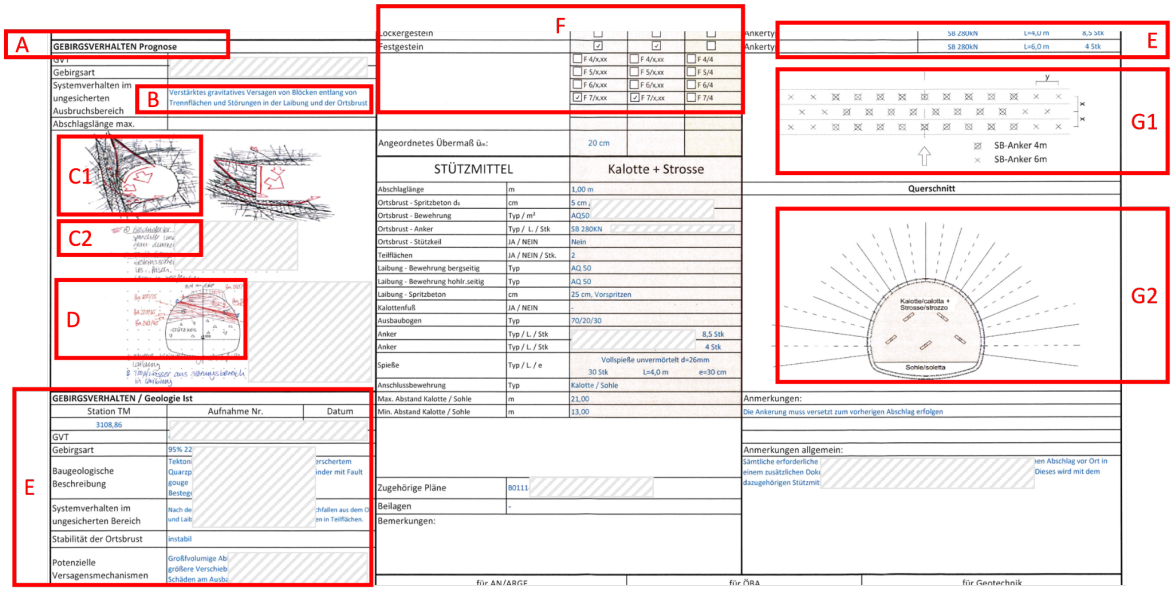

Figure 2. Shows a sample of a technical document containing information in different forms: images of text characters (A, B, E), images of tables (E, F), images of handwriting (C2, D), sketches (C1, D, G2), images of categorical data (G1), and images of plots (G2). Due to an existing NDA on the data, the text here is partially covered in order to hide the content, but the type of information represented by a text (strings of natural language, numerical values, images, etc.) is left uncovered.

The character recognition process can also be applied to handwriting: the results of our experiment in handwriting recognition and translation are illustrated in Figure 3.

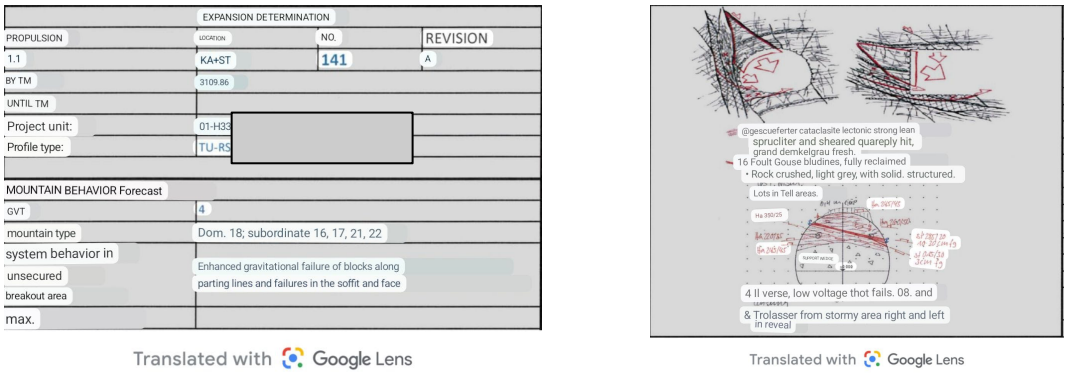

Figure 3. Two documents, in which images of text- (left) and handwritten characters (right) were processed with the Google Lens tool and translated from German to English.

With this method, the translated text would require post-processing: where natural languages or expressions are used, therefore, natural language parsing (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 can be found, examples of commercial solutions include (Diatoz, Semantha). On the other hand, there are many useful free Python libraries available: PyPDF2 (Fenniak et al., 2022) for text extraction from images, Camelot (Mehta, 2023) or Tabula-Py (Ariga, 2023) for extraction of tables, PyMuPDF (Artifex et al., 2022) for unstructured information extraction, PyTesseract (Hoffstaetter, 2022) for data extraction from images, or LayoutParser (Shen et al., 2021) for multiple image analysis tasks.

Generated output from the NLP models also requires certain validation: in Figure 3, for example, the word "demkelgrau" from german-language handwriting was not correctly recognized ("dunkelgrau"); thus, it shall be flagged as "not in a dictionary" and returned to an expert for review. A possible solution to enhance the validation of NLP outputs would be to develop a routine that searches for extracted words in a catalog and flags the unrecognized terms. Subsequently, the catalog shall be automatically updated when a new term is approved by an expert.

After the text and tables are extracted and structured, the image processing step can take place. Because a significant volume of information can be extracted from these images, the following section of this work is dedicated to an overview of methods useful for extracting structured information from images.

3 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 can be solved using the Hough transform method (Duda & Hart, 1972) and morphological operations (Sreedhar & Panlal, 2012) to detect lines and points. The idea is to use automatic image thresholding (Otsu, 1979) to filter large non-connecting objects. These methods are realized in, e.g., the Open Source Computer Vision library (OpenCV) (Bradski, 2000). The accuracy of the data extraction greatly depends on image resolution. Some preprocessing might be required to improve the resolution of the original document with methods like filtering, histogram equalization, contrast adjustment, or machine learning.

For structured information extraction from the engineering sketches, the classification one hot encoding method can be applied where a routine can be developed to assign a binary feature for each possible category of each sample. For example, in Fig.2D, the categories can be a position of a possible mark with respect to a tunnel face (top/bottom/middle, left/right/center), so the corresponding table (Table 1) can be created.

Table 1. Sample table of structured information (part) extracted from image D, Figure 2.

Position	Top	Middle
left	220/35	245/45
right	260/50	285/20
center	350/25	-

The accuracy of the suggested solution can be validated with the data extracted from the text of the same document.

For processing photos, two routines can be suggested: image quality enhancement and texture image analysis. The review of the state-of-the-art in texture recognition (Yang, Li & Ma, 2022) shows the power of the YOLO (You Only Look Once) model based on compound-scaled object detection models that are trained on the COCO (Common Objects in Context) (Lin et al., 2014) dataset. Another approach is to perform texture analysis in the photos with the GLCM (Gray-Level Co-Occurrence Matrix) approach: contrary to deep learning YOLO methods, this is a statistical method that characterizes the texture of an image by calculating patterns. GLCM relates the frequency of appearance of pixels with specified spatial patterns occurring in an image and then extracts statistical measures from this relationship.

4 CONCLUSIONS

The core idea behind this work was to provide users: workers (engineers, site supervisors), and businesses (constructors, consultants, policymakers), regardless of their expertise level (students, researchers, experts), with a flexible and scalable tool that quickly integrates complex information stored in technical documentation from different times and locations as hard copy format or images (i.e., scans) of those. Such a tool would offer the possibility to enhance decision support during construction, having effects not only on the overall safety at the construction site but also beneficially influencing construction costs and responsible use of resources. 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 of structural information retrieval from archived technical documentation as a major step towards improving cross-site knowledge transfer and information flow, ultimately enabling the integration of proactive management. One use case of the latter could be the comparative analysis of tunnel data from different sites, which will help in the selection of effective strategies during project planning, execution, and maintenance.

REFERENCES

Ariga, A., 2023. tabula-py [WWW Document]. URL https://pypi.org/project/tabula-py/

Artifex, McKie, J.X., Liu, R., 2022. PyMuPDF [WWW Document]. URL https://pypi.org/project/PyMuPDF/

Austrian Practice of NATM Tunnelling Contracts. Austrian Society for Geomechanics, web site https://www.austrian-standards.at/

Aygar, E.B., Gokceoglu, C. An assessment on the inner lining need for a large-span tunnel (a case from Turkey, Akyazı Tunnel, Trabzon). SN Appl. Sci. 3, 457 (2021). https://doi.org/10.1007/s42452-021-04366-1

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

Diatoz, online https://diatoz.com/

Dickmann, T., Hecht-Méndez, J., Krüger, D., Sapronova, A., Unterlaß, P.J. and Marcher, T. 2021. Towards the integration of smart techniques for tunnel seismic applications. *Geomechanics and Tunnelling*, Vol. 14, No. 5, pp. 609—615. DOI: https://doi.org/10.1002/geot.202100046

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.

EuroGeoSurveys, Europe Geology data, online https://www.europe-geology.eu/

Fenniak, M., Stamy, M., pubpub-zz, Thoma, M., Peveler, M., exiledkingcc, pypdf Contributors, 2022. The {pypdf} library.

Google Cloud ML Tools for Optical Character Recognition, online https://cloud.google.com/vision/docs/ocr

Hoffstaetter, S., 2022. pytesseract [WWW Document]. URL https://pypi.org/project/pytesseract/

Leroux, Virginie and Audrey Campeau. "Tunnel Database: An Information System Useful for Underground Construction in Montreal." (2018).

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8693 LNCS, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Marinos, V., Prountzopoulos, G., Fortsakis, P. et al. "Tunnel Information and Analysis System": A Geotechnical Database for Tunnels. Geotech Geol Eng 31, 891–910 (2013). https://doi.org/10.1007/s10706-012-9570-x

Mehta, V., 2023. Camelot: PDF Table Extraction for Humans [WWW Document]. URL https://pypi.org/project/camelot-py/

Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys. Man. Cyber. 9 (1): 62–66

Semantha, online https://www.semantha.de

Shen, Z., Zhang, R., Dell, M., Lee, B.C.G., Carlson, J., Li, W., 2021. LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. arXiv Prepr. arXiv2103.15348.

Simons, Bruce & Raymond, Oliver & Jackson, Ian & Lee, Katy. (2012). OneGeology—Improving global access to geoscience. Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia. 265. 10.1201/b12728-53.

Sreedhar, K., & Panlal, B. (2012). Enhancement of images using morphological transformation. arXiv preprint arXiv:1203.2514.

Yang, Nan & Li, Yongshang & Ma, Ronggui. (2022). An Efficient Method for Detecting Asphalt Pavement Cracks and Sealed Cracks Based on a Deep Data-Driven Model. Applied Sciences. 12. 10089. 10.3390/app121910089.