Drilling a large corpus of document images of geological information extraction

Jean-Louis Debezia[1], Mélodie Boillet[2], Christopher Kermorvant[2], and Quentin Barral[1]

[1] Geosophy, 155 Bvd de l'Hôpital, 75013 Paris, France
[jean-louis.debezia,quentin.barral]@geosophy.io
http://www.geosophy.io

[2] TEKLIA, 30 rue Raymond Losserand, 75014 Paris, France
[boillet,kermorvant]@teklia.com
http://teklia.com

Abstract. Geo-energy is a resource widely available on Earth that consists in using the first 10-100 meters below the surface where the temperature is low and constant during the year. This 12 to 15°C is ideal for heat pumps to provide heat and cold with an excellent coefficient of performance. Despite a very high potential, this resource is not often integrated in France. One of the main reasons is the lack of knowledge of rocks capacity to deliver sufficient power on surface. More than 2 millions of scanned documents are available on the french geological survey. We propose a way to classify and analyze them in order to quantify the underground resource.

Keywords: Classification · Neural network · Geo-energy · Permeability · Green buildings.

1 Introduction

Because the building sector represents 40 % of global consumption and 36 % of greenhouse gas emissions [3], the EU is seeking sustainable solutions for heating and cooling, adapted to existing constructions. 78 % of the EU's population lives in areas where the soil thermal inertia can be useful for both cooling and heating [4], making geothermal heat pumps a highly efficient solution, able to save up to 75 % of building consumption and 90 % of its carbon emissions compared to gas systems. Today though, it covers only 2 % of total heating/cooling consumption while this abundant resource could easily be used ten times more. It is due to: 1) the lack of awareness of the available resources: the feasibility studies being risky, expensive and only performed by rare experts; 2) the uncertainty of the quality

[3] https://ec.europa.eu/energy/en/topics/energy-efficiency/heating-and-cooling

[4] https://www.powerengineeringint.com/coal-fired/advances-made-in-eu-geothermal-heating-development/

of the resource, which determines the output of the system; 3)unknown investment costs and ROI, which prevent it from being planned when it is financially relevant.

In the case of geothermal heat pumps, the operational principle does not rely on heat extraction, but on constant temperature in the subsurface, at depths typically between 10 and 400 m [5]. This level generally maintains the same temperature in winter as in summer between 15 and 20°C.

2 Underground data

The technical design offices can provide a first approximation of the geo-energy power available in the ground, given either as a water flow rate or a transmissivity/permeability. The proposed value is estimated through neighboring installations: for each new project, the design engineer studies its own data sets and the public data available in the local geological survey (BRGM in France) in order to gather information on nearby wells. From those wells, they propose an estimate for the petrophysical parameters such as porosity, permeability [3] which represents the ability of water to flow in the rocks, thermal conductivity and thermal capacity. There is no global and pre-analyzed database, only raw data are available in France.

Due to the high level on uncertainties of rocks properties (and in particular the permeability), design offices accept to confirm the first approximation only after drilling a well and running pumping tests on it. This induces for the owner a financial risk (the drilling cost is between 20 and 200 k€ depending on the accessibility of the site and the characteristics of the well namely depth, diameter and type of rocks) and a technical risk (potential modifications of the construction plans if the estimated resources are not confirmed after drilling). It is easily understandable that an important number of buildings owners don't go deeper in geo-energy potential analysis due to those risks. It is difficult to perform a market analysis to quantify the impact of the financial and technical risks, but the facts are that only 2,000 heat pumps based on geo-energy were sold in France in 2020 (including all type of buildings, from small independent houses to big office buildings) whereas 812,000 aerothermal heat pumps (classical system) were installed. From a technical point of view, it is not satisfying to impose a invasive testing well when a lot of data are not processed.

3 Automatic geological information extraction in document images

3.1 General document processing workflow

More than 800,000 underground structures (geotechnical survey, oil and gas wells, water wells, geo-energy wells,...) are listed in the BRGM database. Associated with them, 2,213,989 scanned documents are available online. They are

[5] https://ericsenergy.com/heating-cooling/learn-more-about-geothermal-heat-pumps/

drilling and tests reports, geological logs, water quality, etc. and contains many key information to model the rocks parameters. The diversity of the documents is such that it is not possible to extract geological information with a single generic procedure. It is necessary to develop several extraction pipelines depending on the typology of the documents: mainly textual, tables, forms, maps or sketches. The different pipelines are presented on fig. 1. The first stage is common for all the pipelines and consists in an automatic classification of the type of pages. This classification will determine which extraction process has to be applied to the page. We present in this section the training and evaluation of the page classification system.

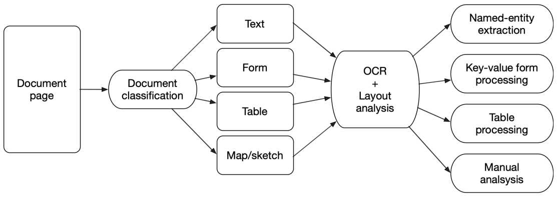


Fig. 1: The different workflows needed to extract the geological information from the scanned documents, depending on their types

3.2 BRGM Dataset annotated for classification

The BRGM document data set is composed of 2,213,989 PDF documents, containing multiple pages. Each page can be composed of different elements such as text zones, tables, forms, graphs, maps, photos, logs. Since it is not possible to know in advance which of this structure contains the geological information we are interested in, each page should be labeled with the different types of structure it contains. The classification problem is therefore inherently multi-class and multi-label (cf. fig.2).

In order to create a training sample for the automatic document classifier, we sampled 2,000 documents from the complete data set, which represent 2,323 images after splitting the documents into pages. The sample was manually labeled using the document processing platform Arkindex[6]. Seven labels were defined to represent the different elements located in the documents: `log - form - graph - map - photo - table - text`. Examples of three pages showing the different elements are presented on fig. 3.

The labeling of the pages was not straightforward. First identifying the different types of structure is sometimes not obvious. For example, a table with only 2 columns could be considered as a form or a simple form containing a lot of text could be considered a plain text zone. Moreover, some structural elements cover only a small part of the page and may not be relevant for the information extraction. However, since we don't know in advance where lies the information, we decided to focus on labeling the different elements present on the page whether they contain an information of interest or not.

[6] https://demo.arkindex.org

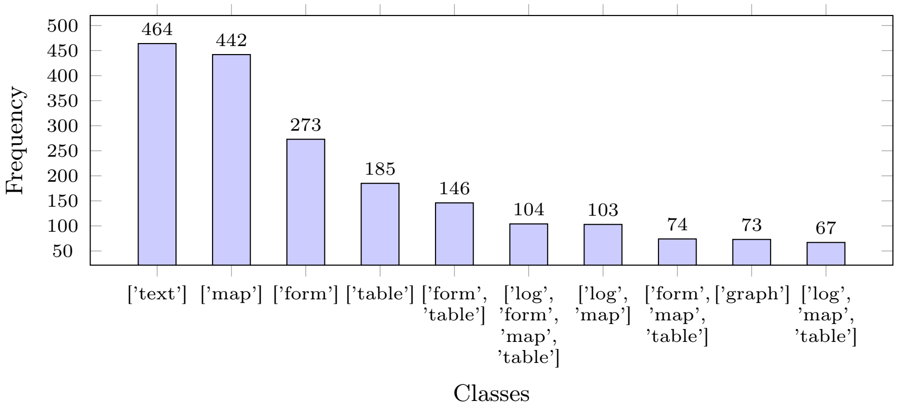
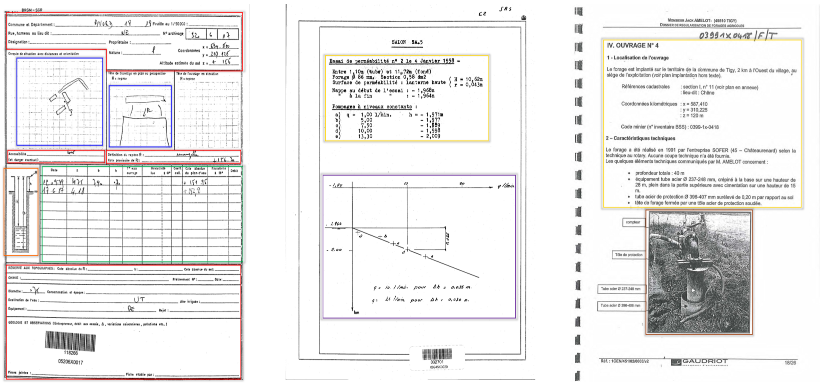


Fig. 2: The 10 most frequent classes combination in the BRGM dataset of scanned documents.



(a) map log form table (b) text graph (c) text photo

Fig. 3: Sample images from the BRGM database along with their labels

3.3 Document classification model

Training a model from scratch requires a very large amount of data to achieve an accurate classification. Since we had only 2,323 images, we chose to use transfer learning in order to fine-tune a pre-trained model on our images and get accurate classification. We first compared two Deep Neural Network architectures (VGG [4] and ResNet [2]) pre-trained on ImageNet [1] as feature extractor. This latter is followed by layers randomly initialized [Linear-ReLU-Dropout-Linear] to run the classification and a final sigmoid layer to obtain probabilities for each class. Since we have a small number of training sample, the first Linear layer reduces the number of features to 500, which also limits the number of parameters to be computed. To run the experiments, we randomly split the dataset into training, validation and testing sets which respectively corresponds to 1858, 232 and 233 images (80-10-10%). The images are resized to fit the input size of the pre-trained model (224x224 RGB for VGG and ResNet), normalized and randomly flipped for data augmentation. The parameters of the VGG and ResNet layers are kept constant during training, only the added classification layers are fully trained on our images. Since each image can belong to multiple classes, the probabilities output by the sigmoid layer are thresholded with $t = 0.5$ to assign the final classes. First, we chose to compare various feature extractors to select the best one before improving the results. These first experiments showed that, under the same conditions, VGG16 yielded better results on our dataset (based on BCELoss and the F1-Score) than the other ResNet and VGG models. For the next experiments, VGG16 pre-trained model is then used as feature extractor.

The second set of experiments consisted in optimizing the model's hyper-parameters: batch size, number of epochs, learning rate and optimizer. We also tested to change the loss to the BCEWithLogitsLoss, however the results were similar to the standard BCELoss. Table 1 shows the ranges of values tested.

Hyperparameter	**Range**
Learning rate	[0.01, 0.005, 0.001, 0.0005, 0.0001]
Batch size	[16, 32, 64]
Optimizer	[Adam]

Table 1: Ranges of the tested values for VGG16's hyparameters optimization

Batch size of 32, 30 epochs and a 0.005 learning rate gave the accurest model whose results are presented in the next section.

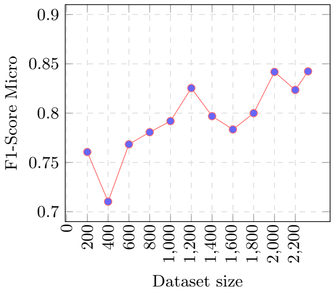
3.4 Results and discussion

The results of the best model are presented on Table 4a. The average F1-Score is 0.84 which is a good score considering our data set size and the complexity of the task. When we focus on `form - table - text` classes, 78% of form elements are detected by the model (86% for tables and 87% for text) and only 17% are confused with another element type (8% for tables and 11% for text). The training curve presented on fig. 4b shows a positive correlation between the F1-Score Micro and the data set size suggesting a potential further increase of the model performance with more labeled data.

One limitation of our approach is that the definition of the different classes is not very precise and subjective so that the manual classification is sometimes difficult to achieve. Moreover, this approach does not take into account the proportion of the document for the classification: a small table of 2 lines and 2 columns has the same impact on the manual classification as a full page table. Finally the approach does not provide either the location of the different elements, which will need to be predicted by the following processing steps.

Class	Precision	Recall	F1-Score
Coupe	0.88	0.79	0.83
Form	0.78	0.83	0.80
Graph	0.81	0.52	0.63
Map	0.92	0.83	0.87
Photo	1.0	1.0	1.0
Table	0.86	0.92	0.89
Text	0.87	0.89	0.88
Micro Avg	**0.86**	**0.83**	**0.84**

(a) Performance metrics by labels



(b) F1-Score Micro and dataset size correlation

Fig. 4: Results on test set (233 images)

4 Conclusion

We presented the first step of a fully automatic workflow to extract geological information in scanned drilling reports. On a small sample of 200 documents, selected randomly on the entire database, we observed that 3% of the documents contain information on well productivity and rocks capacity. If this ratio is also observed in the entire dataset, it means that more than 60,000 water rates can be found in the documents. As a reference, our geologists team went manually through documents: after three man.months of hard work, 5,000 flow rates have been collected. With more than 60,000 wells identified as water productors and quantified with a water flow rate, the hydro-geological model of the first hundred meters of France will be dense and robust, thus improving the prediction of geo-energy feasibility and selecting properly the right buildings to connect to their underground.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
3. Lamé, A.: Modélisation hydrogéologique des aquifères de Paris et impacts des aménagements du sous-sol sur les écoulements souterrains. Theses, Ecole Nationale Supérieure des Mines de Paris (Dec 2013), https://pastel.archives-ouvertes.fr/pastel-00973861
4. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.1556