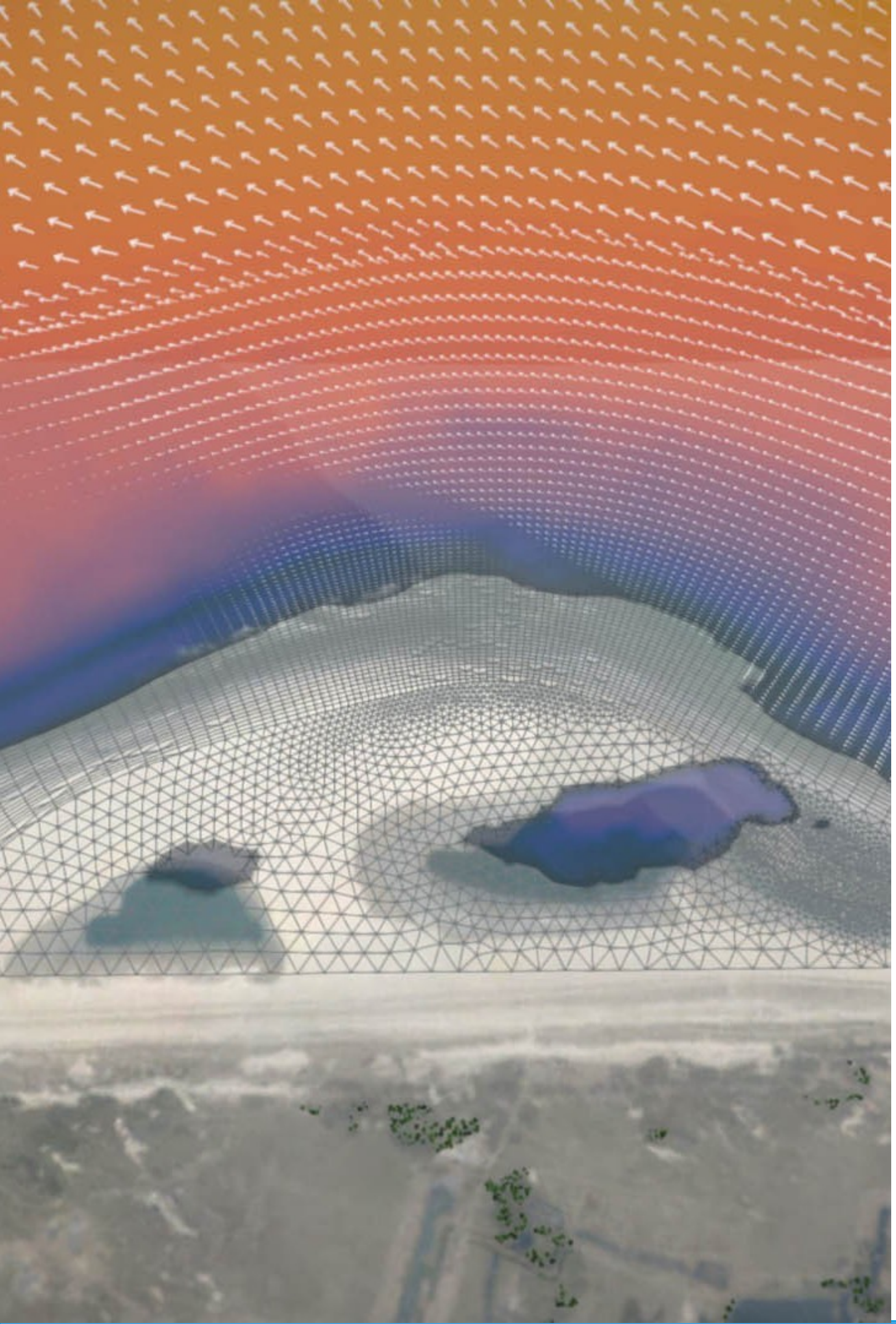


DIGISHAPE
DATASCIENCE FOR DELTAS

Automated information extraction from borehole log reports

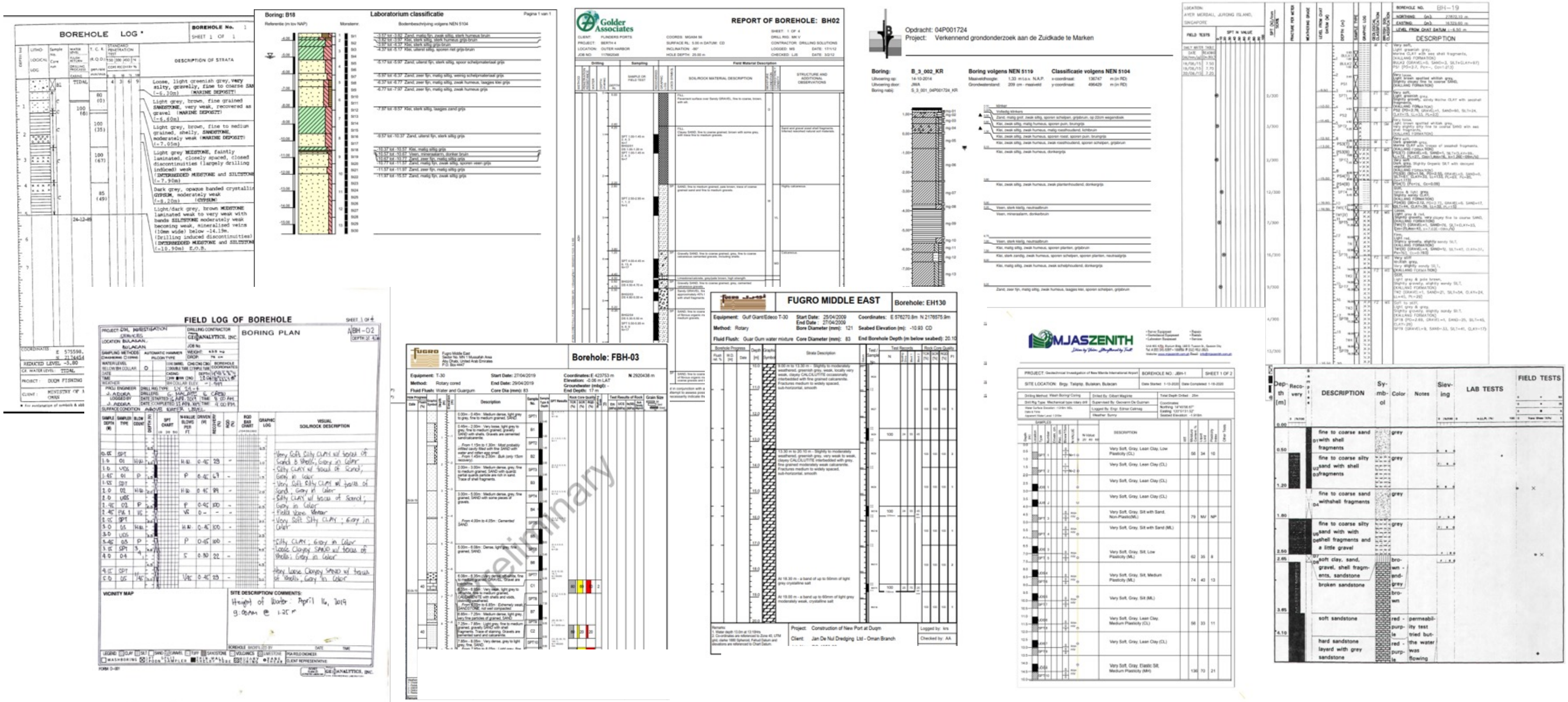
RESHMI KURIAN / WIEBKE JÄGER

Illustration of a Borehole Log Report

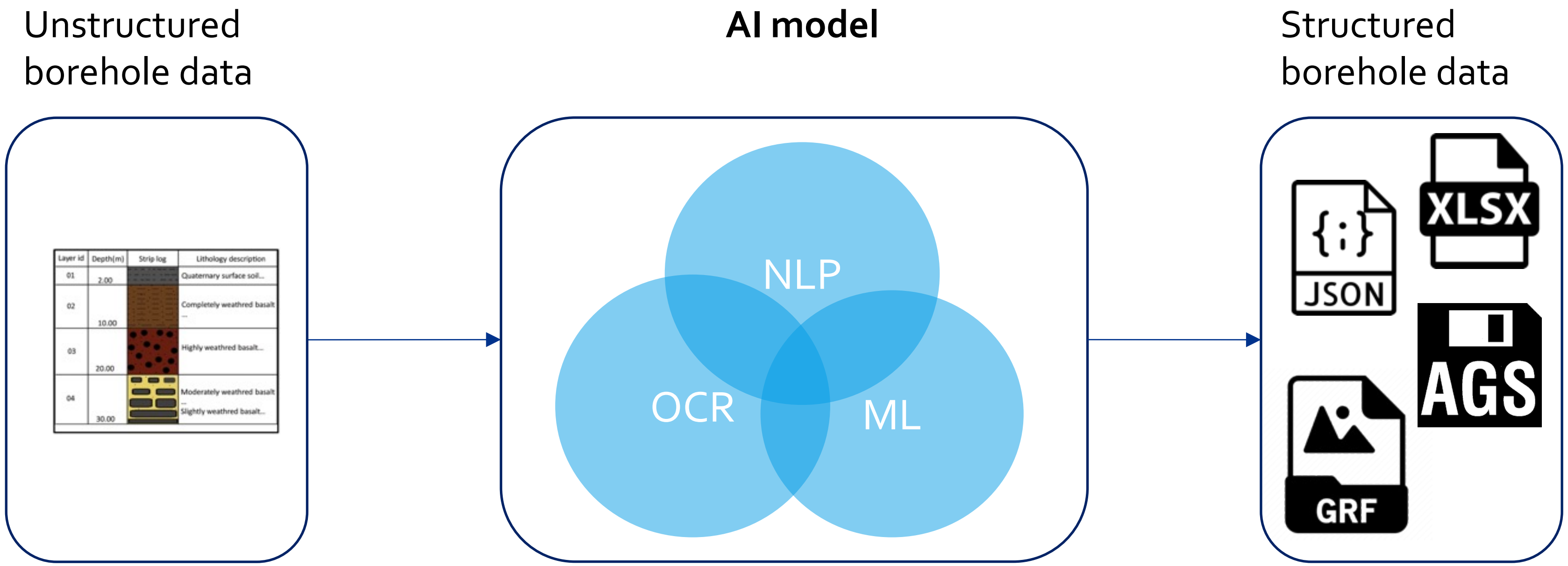
Layer id	Depth(m)	Strip log	Lithology description
01	2.00		Quaternary surface soil...
02	10.00		Completely weathred basalt ...
03	20.00		Highly weathred basalt...
04	30.00		Moderately weathred basalt ... Slightly weathred basalt...

Image from: https://www.nature.com/articles/s41597-019-0345-6

Real Borehole Log Reports



Seedmoney Project



<u>Seed money project:</u>

- Shared library with model for digitization of borehole log reports (in TUDelft CTIG Github)

<u>General:</u>

- Share knowledge and skills for processing and analysing unstructured documents with OCR, NLP & ML
- All partners can continue to improve the model with additional training using their own borehole log reports

Agenda Session

13:00 – 13:30 Seedmoney project & GitHub Code (*Reshmi Kurian & Wiebke Jäger*)

13:30 – 13:40 Experiences from IQIP (*Joost van Hooren & Jort van Wijk*)

13:40 – 13:50 Experiences from RHDHV (*Camille Habets & Jelle van der Zon*)

13:50 – 14:15 Brainstorming in Groups: Main Challenges & Opportunities

14:15 – 14:30 Sharing Results and discussion

Development of the project with key-user group

16-05-2022 : Project kick-off meeting

- Identified that the user group is also interested in extracting the metadata from the borehole logs

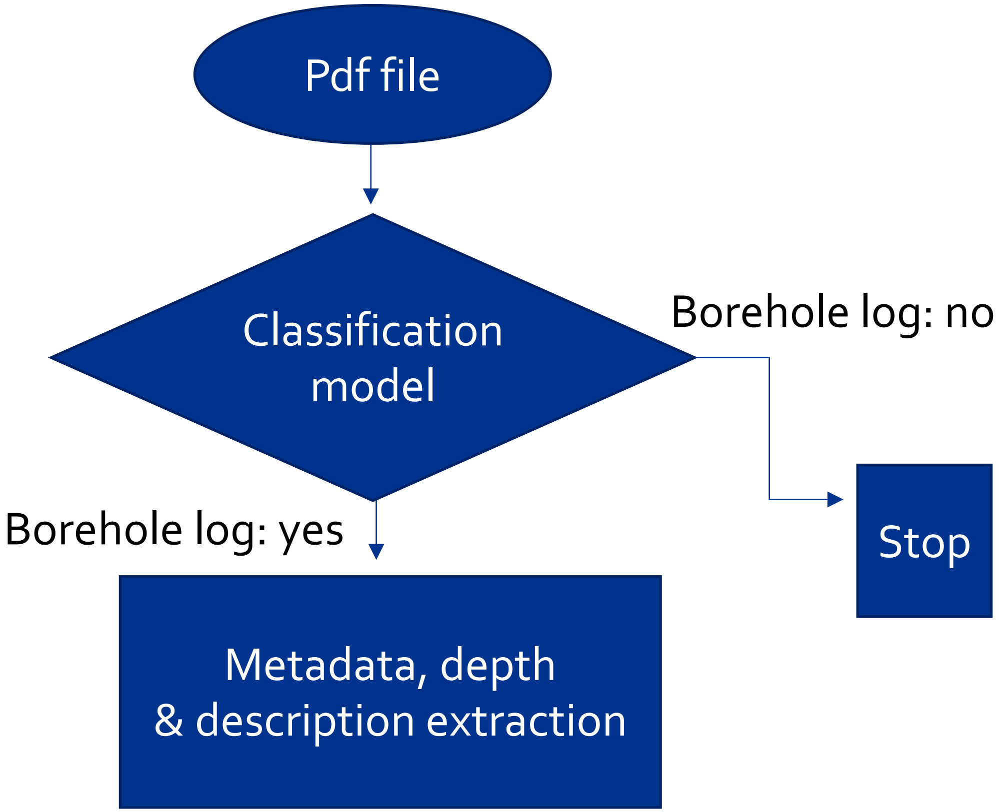
22-07-2022: Update meeting with the user group to present the intermediate results

08-11-2022: Final presentation of the seed-money project and brain-storming for the follow-up of the project.

Chosen Approach

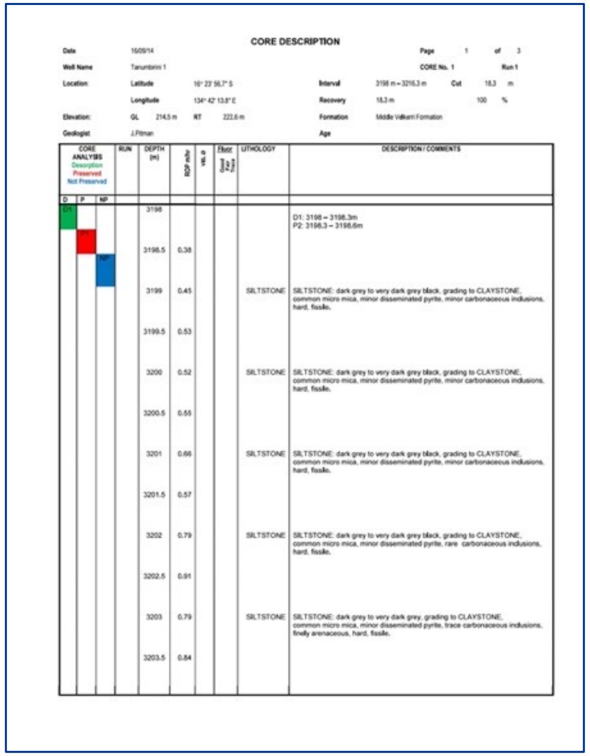
Steps:

- Identify the borehole logs from reports
- Extract metadata such as geographical coordinates, date, country etc.
- Extract depth and description fields.

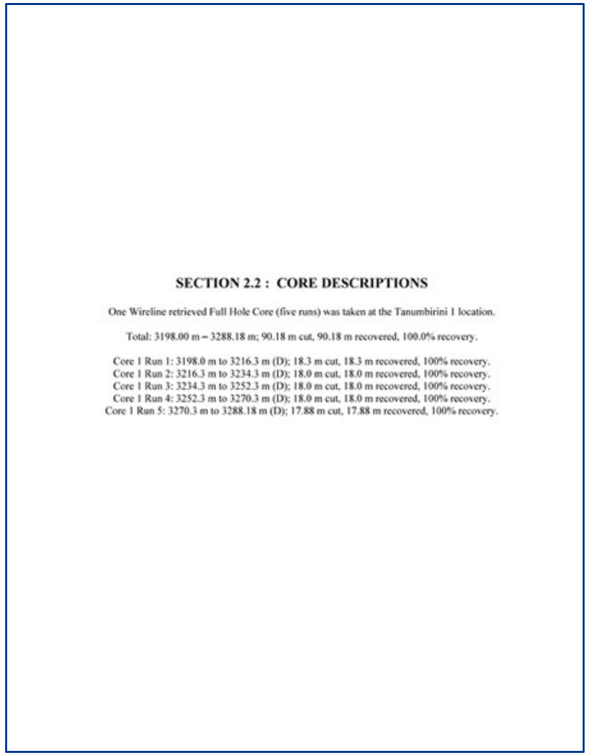


Step 1: Identify borehole logs

Borehole log



Non borehole log



SECTION 2.2 : CORE DESCRIPTIONS

One Wireline retrieved Full Hole Core (five runs) was taken at the Tanumbirini 1 location.

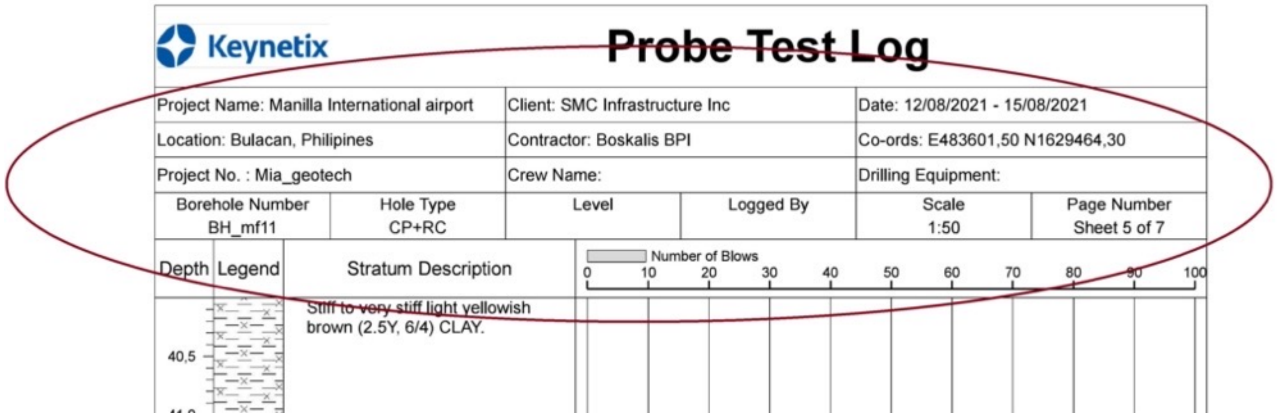
Total: 3198.00 m – 3288.18 m; 90.18 m cut, 90.18 m recovered, 100.0% recovery.

Core 1 Run 1: 3198.0 m to 3216.3 m (D); 18.3 m cut, 18.3 m recovered, 100% recovery.
Core 1 Run 2: 3216.3 m to 3234.3 m (D); 18.0 m cut, 18.0 m recovered, 100% recovery.
Core 1 Run 3: 3234.3 m to 3252.3 m (D); 18.0 m cut, 18.0 m recovered, 100% recovery.
Core 1 Run 4: 3252.3 m to 3270.3 m (D); 18.0 m cut, 18.0 m recovered, 100% recovery.
Core 1 Run 5: 3270.3 m to 3288.18 m (D); 17.88 m cut, 17.88 m recovered, 100% recovery.

- Image classification using convolutional neural network with TensorFlow.
- 800 images (70% for training, 30% for testing the model)
- Accuracy on the test set: 92%

Step 2: Extract metadata

Extract geographical co-ordinates, city, country, dates and other information from the borehole logs.


Keynetix Probe Test Log

Project Name: Manilla International airport	Client: SMC Infrastructure Inc	Date: 12/08/2021 - 15/08/2021
Location: Bulacan, Philipines	Contractor: Boskalis BPI	Co-ords: E483601,50 N1629464,30
Project No. : Mia_geotech	Crew Name:	Drilling Equipment:

Borehole Number	Hole Type	Level	Logged By	Scale	Page Number
BH_mf11	CP+RC			1:50	Sheet 5 of 7

Depth | Legend | Stratum Description | Number of Blows 0 10 20 30 40 50 60 70 80 90 100

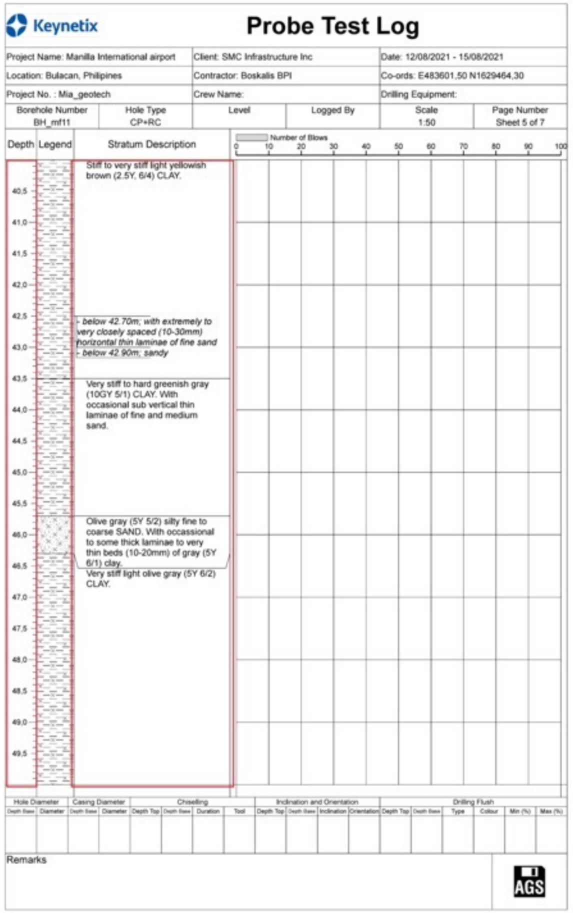
Stiff to very stiff light yellowish brown (2.5Y, 6/4) CLAY.

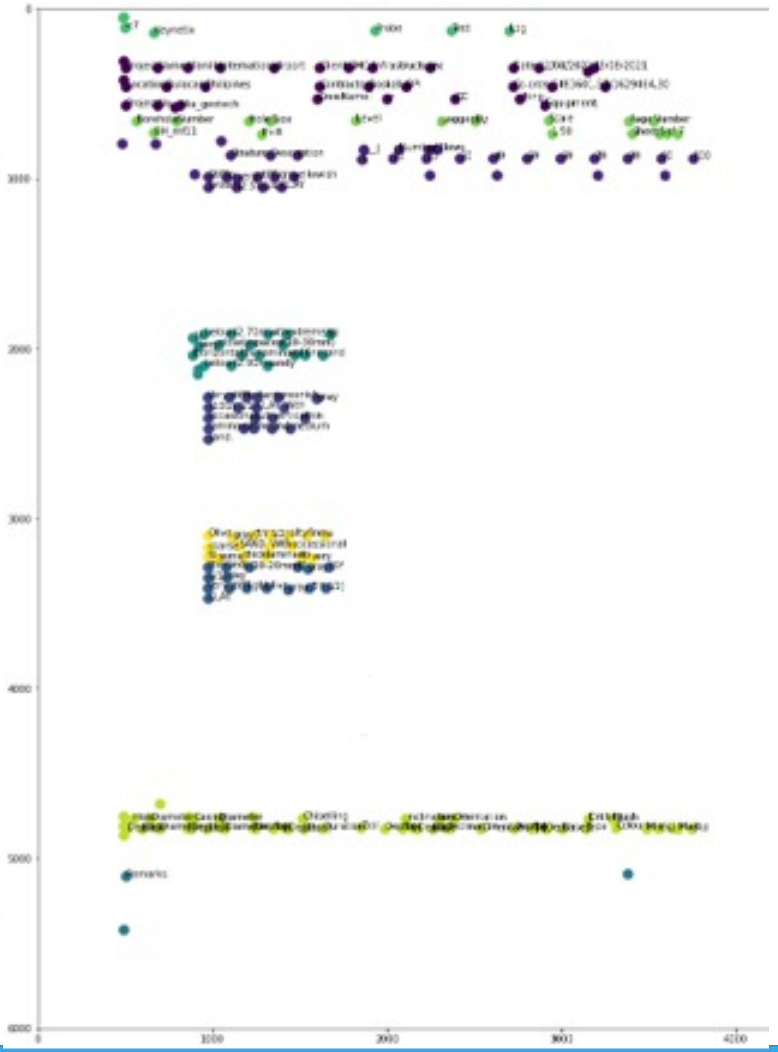
40,5

- Use Tesseract to extract all text from pdf files
- Named Entity Recognition (Spacy NER) to identify country, city
- Patterns to identify geographical co-ordinates, dates, borehole number (regular expressions)

Step 3: Extract depth and lithology description

- Use Tesseract to localize text in the table and extract the bounding box *(x, y)*-coordinates of the text in the table.
- Hierarchical agglomerative clustering (HAC) to associate rows and columns. Clusters are created based on the y-coordinate of the table with a maximum distance threshold cutoff.
- Identify the 'depth' values based on the minimum and maximum y-coordinates of each of the clusters.



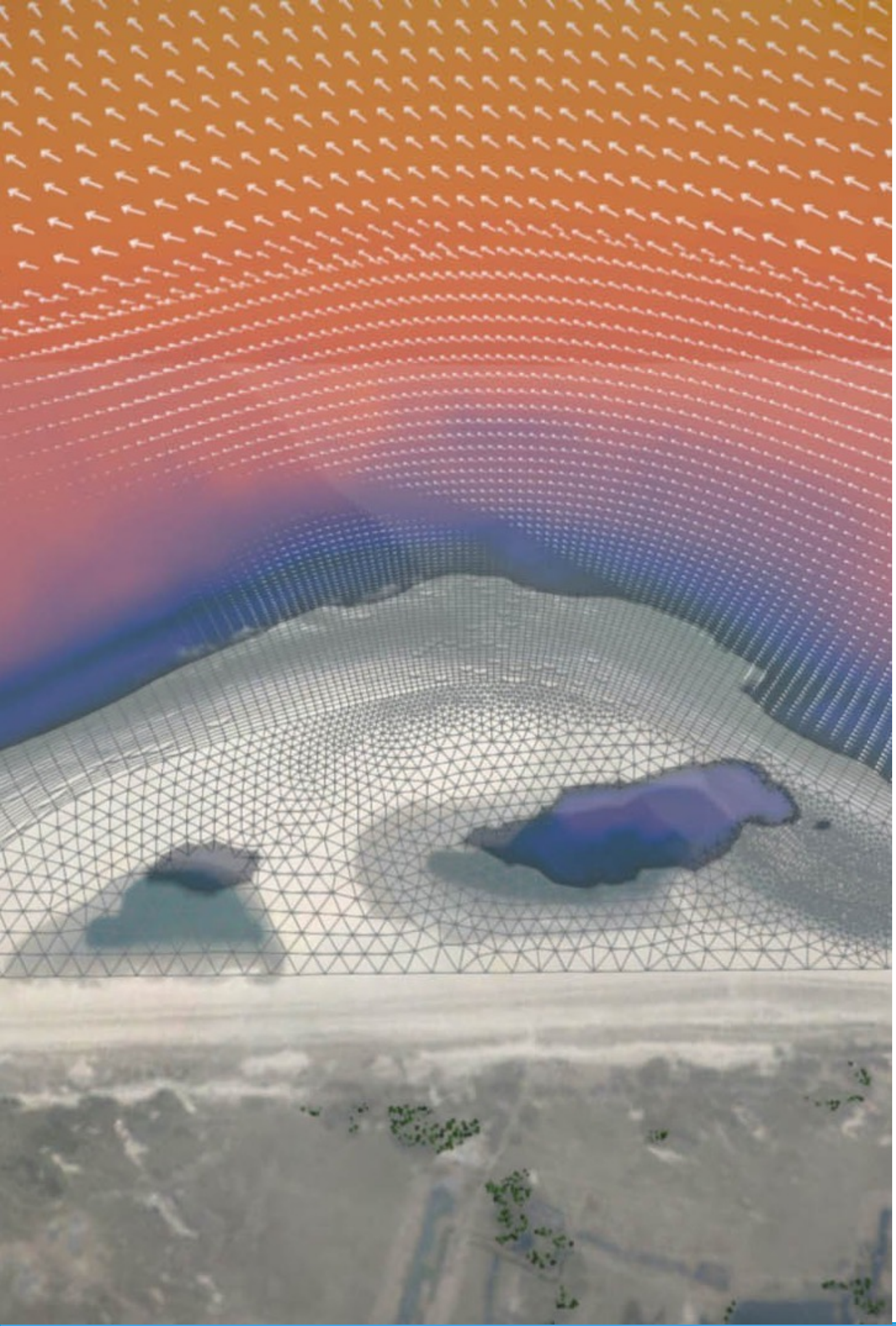


Code Demo

- Code and examples available at TUDelft CTIG Github (access needs to be requested)

Conclusions

- Step 1: Identify borehole logs
  - Achieved 92% accuracy in recognizing borehole log pages from other pages
  - Limitations:
    - Training data set of only 800 pages
- Step 2: Extract meta-data
  - Extraction of meta-data possible for several borehole logs (see examples in GitHub)
  - Limitations:
    - Not all formats of, e.g., coordinates, borehole number recognized yet
    - Validation strategy?
- Step 3: Extract depth and lithology description
  - Extraction of depth and lithology description possible for several borehole logs (see examples in GitHub)
  - Limitations:
    - Manual fine-tuning of parameters can be required for different templates
    - Validation strategy?



Questions?

Wiebke.Jager@boskalis.com

Reshmi.Kurian@boskalis.com

www.digishape.nl