Digitization of Borehole Logs

IQIP – Sliedrecht
08-11-2022

Jort van Wijk
Joost van Hooren

Agenda

What to expect

- Introduction IQIP
- Business Challenges
- Project Digitization Borehole Logs
- Approaches
  - ❑ OCR
  - ❑ Regular expressions
- Lessons learned

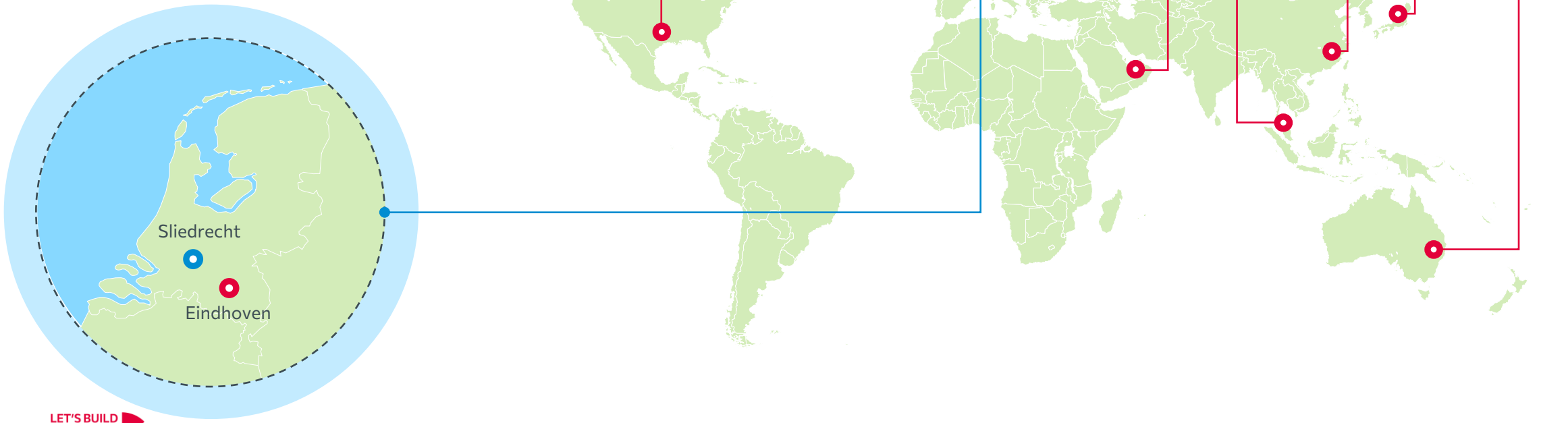
IQIP

GLOBAL LEADING EQUIPMENT SUPPLIER TO ONSHORE AND OFFSHORE CONTRACTORS

400 EMPLOYEES

OVER 80 OFFSHORE WIND PROJECTS

4.000 MONOPILES INSTALLED

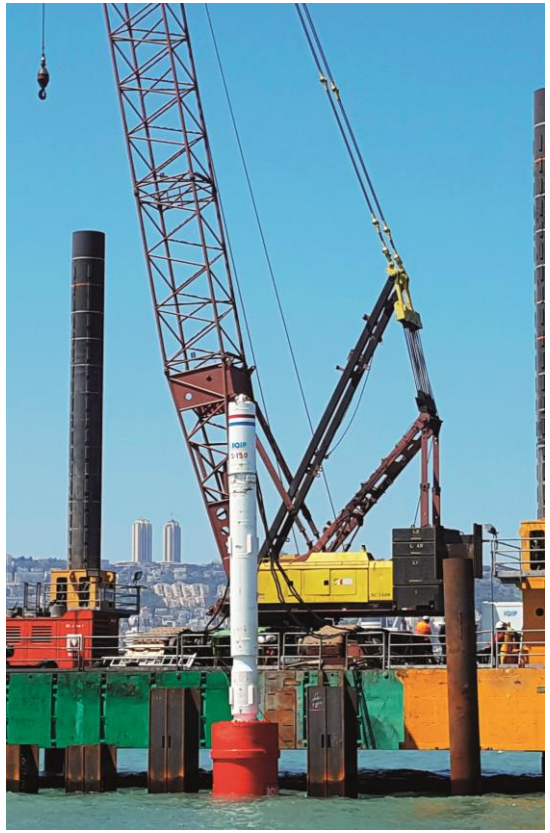


LET'S BUILD IQIP

LET'S BUILD

MARKETS

SERVING THE MARKET WITH BOTH STANDARDIZED EQUIPMENT AND CUSTOM-BUILD SOLUTIONS



COASTAL & CIVIL

OFFSHORE WIND

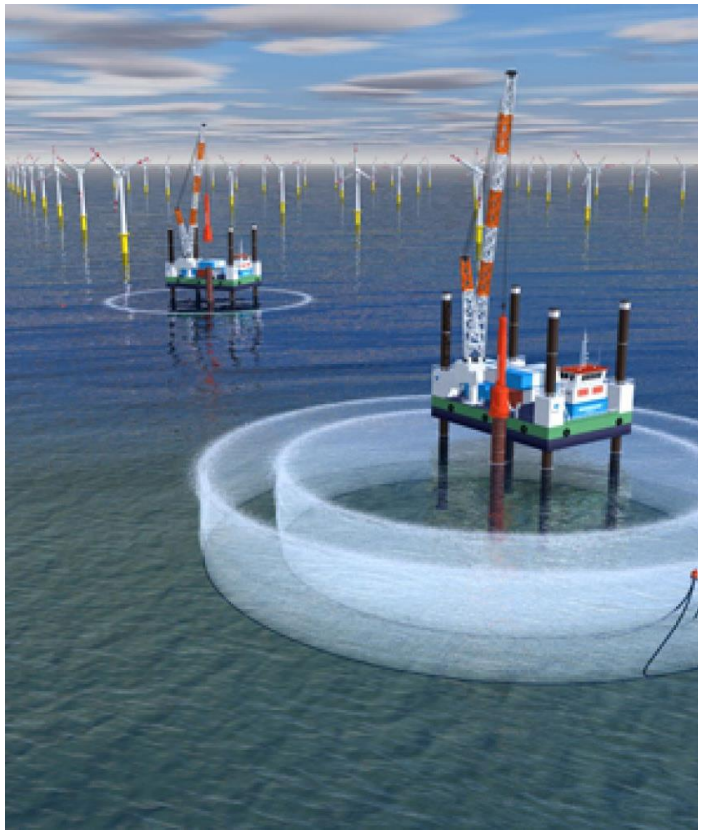
OIL & GAS

DECOMMISSIONING

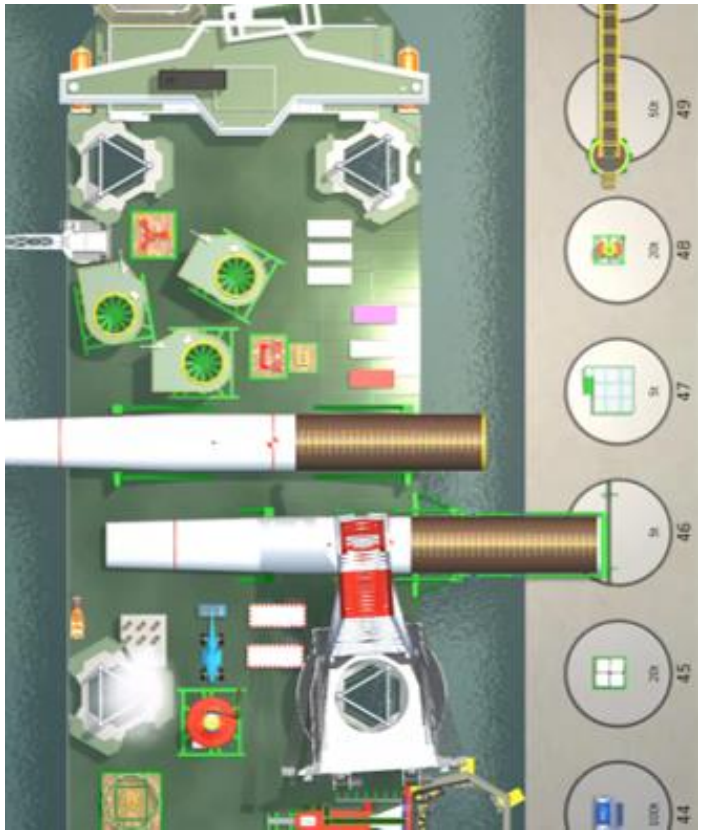
ADVISORY SERVICES

INVALUABLE SUPPORT IN THE EARLIEST PROJECT STAGE

DRIVABILITY ANALYSIS



NOISE PREDICTIONS



METHOD STATEMENTS & OPTIMAL DECK LAYOUT

LET'S BUILD IQIP

LET'S BUILD

BUSINESS CHALLANGES

Geotechnical data sources

- CPT
- Borehole logs
- Lab test

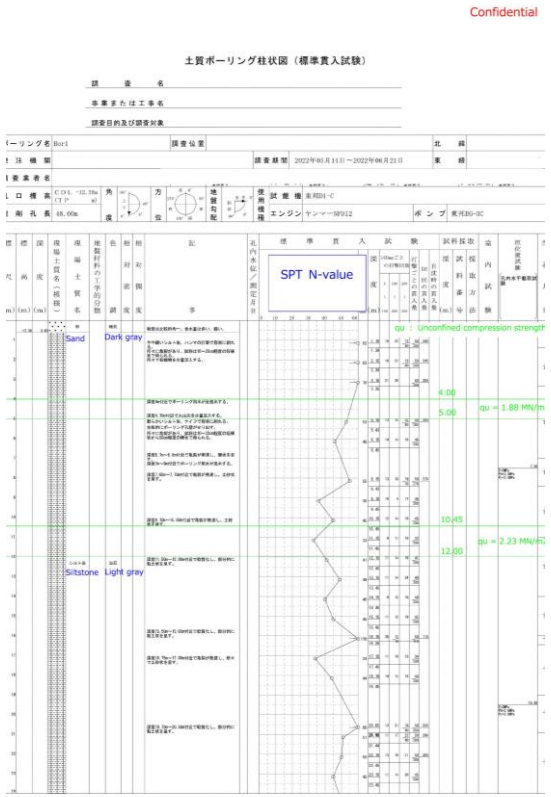
Drivability models

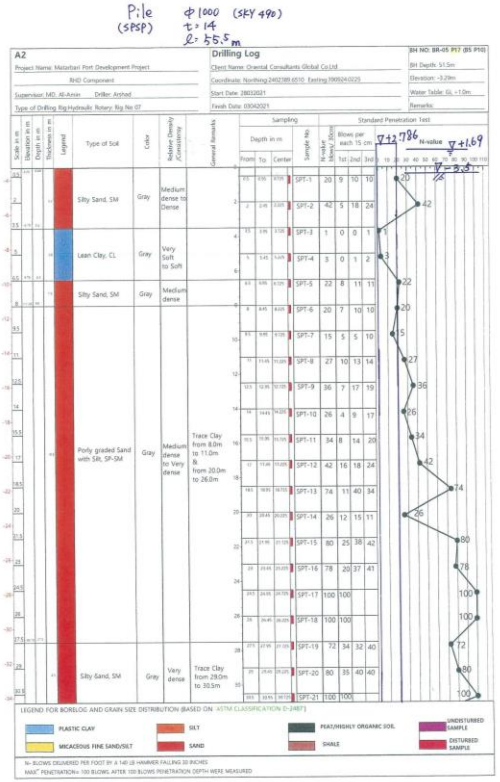
- Structured data
- Manual conversion

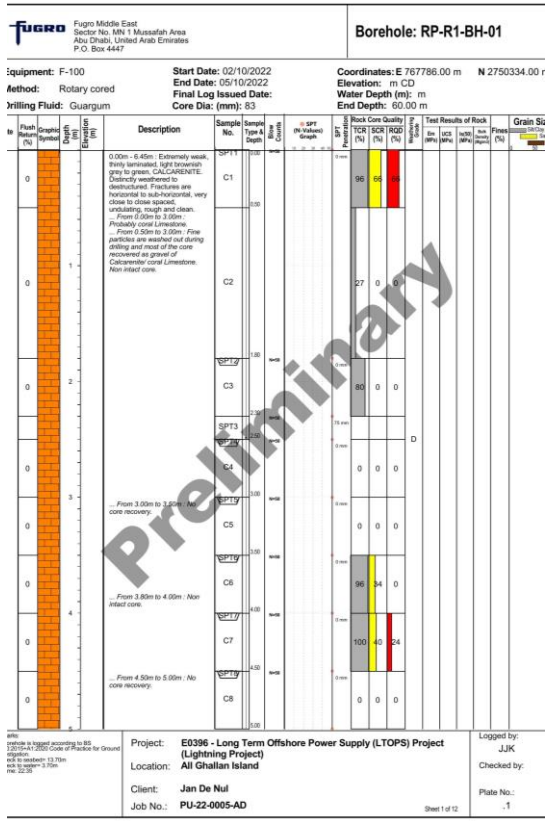
Data format

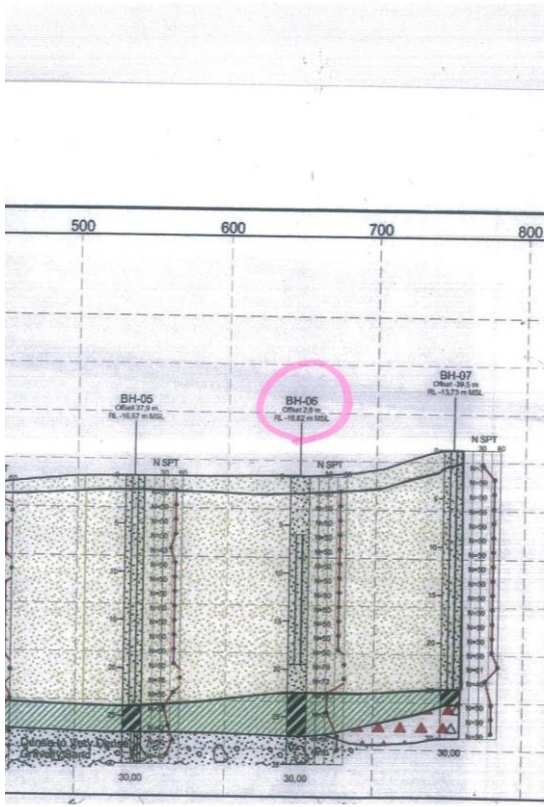
- Unstructured data (pdf/images)
- Wide range of data sources
  - Formats & layout
  - Language
  - Parameters

EXAMPLES









LET'S BUILD IQIP

LET'S BUILD

GOAL OF OUR PROJECT

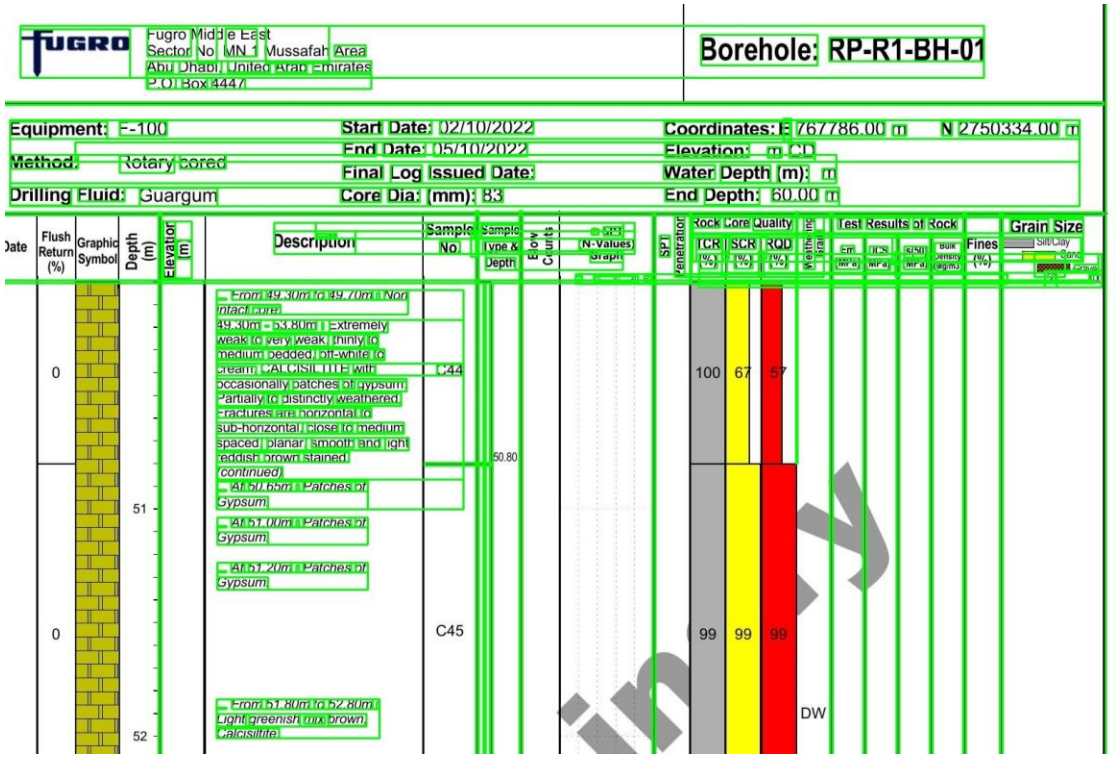
Digitizing borehole logs

"Digitize all received borehole log information so that it can be used to serve geotechnical models"

"Extract relevant parameters from PDF's and images with an **inconstant layout, quality and language**"

APPROACHES

Unstructured to structured data



1. OPTICAL CHARACTER RECOGNITION (OCR)

'Fugro Middle East\nSector No. MN 1 Mussafah Area\nAbu Dhabi, United Arab Emirates\nP.O. Box 4447\n\nBorehole: RP-R1-BH-01\n\nE
quipment:\n\nF-100\n\nMethod:\n\nRotary cored\n\nDrilling Fluid:\n\nGuargum\n\nDate\n\nFlush\nReturn\n(%)\n\nGraphic\nSymbol D
\n\nh\nt\np\ne\n\n)\n\nm\n\n(\n\nn\no\ni\nt\na\nv\ne\nl\nE\n\n)\n\nm\n\n(\n\nStart Date: 02/10/2022\nEnd Date: 05/10/2022\nFina
l Log Issued Date:\n Core Dia: (mm): 83\n\nCoordinates:\nElevation: m CD\nWater Depth (m): m\nEnd Depth: 60.00 m\n\nE 76778
6.00 m\n\nN 2750334.00 m\n\nSample\nNo.\n\nSample\nType &\nDepth\n\nw\no\nB\n\nl\n\ns\nt\nn\nu\no\nC\n\nSPT1\n\n0.00\n\nN=50\n
\n SPT\n(N-Values)\nGraph\n\n10\n\n20\n\n30\n\n40 50\n\nT\nP\nS\n\nn\no\ni\nt\na\nr\nt\ne\nn\ne\nP\n\n0 mm\n\nRock Core Quality
\nRQD\nTCR\nSCR\n(%)\n(%)\n(%)\n\ng\nn\ni\nr\ne\nh\nt\na\ne\nW\n\ne\nd\na\nr\nG\n\nTest Results of Rock\n\nEm\n(MPa)\n\nUCS\n(M
Pa)\n\nIs(50)\n(MPa)\n\nBulk\nDensity\n(Mg/m3)\n\nFines\n(%)\n\nGrain Size\nSilt/Clay\n\nSand\n\nGravel\n100\n\n50\n\n0\n\nDesc
ription\n\n0.00m - 6.45m : Extremely weak,\nthinly laminated, light brownish\ngrey to green, CALCARENITE.\nDistinctly weathered
to\ndestructured. Fractures are\nhorizontal to sub-horizontal, very\nclose to close spaced,\nundulating, rough and clean.\n...
From 0.00m to 3.00m :\nProbably coral Limestone.\n... From 0.50m to 3.00m : Fine\nparticles are washed out during\ndrilling and
most of the core\nrecovered as gravel of\nCalcarenite/ coral Limestone.\nNon intact core.\n\n0.50\n\nSPT2\n\n1.80\n\nN=50\n\nC3
\n\nSPT3\n\nSPT4\n\nC4\n\n2.30\n\n2.50\n\nN=50\n\nN=50\n\nSPT6\n\n3.50\n\nN=50\n\n... From 3.00m to 3.50m : No\ncore recover
y.\n\nSPT5\n\n3.00\n\nN=50\n\n... From 3.80m to 4.00m : Non\nintact core.\n\nSPT7\n\n4.00\n\nN=50\n\n... From 4.50m to 5.00m :
No\ncore recovery.\n\nSPT8\n\n4.50\n\nN=50\n\nC1\n\nC2\n\nC5\n\nC6\n\nC7\n\nC8\n\n96\n\n66\n\n66\n\n27\n\n0\n\n0\n\n0 mm\n\n75
mm\n\n0 mm\n\n0 mm\n\n0 mm\n\n0 mm\n\n0 mm\n\n80\n\n0\n\n0\n\nD\n\n0\n\n0\n\n0\n\n0\n\n0\n\n96\n\n34\n\n0\n\n100\n\n40\n\n
24\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n:\n\n \n\n/\n\n/\n\n:\n\n \n\n \n\n6\n1\n8\n3\n0\n1\n2\n2\n0\n2\n0
\n1\n6\n0\nn\no\n4\n0\n0\n0\n3\n8\nT\nN\ng\n \ny\nb\nd\ne\nc\nu\nd\no\nr\n\nI\n\n.\n\n.\n\n \n\n \n\n \n\n.\n\n \n\nP\n \nJ\nP
\nG\nN\nA\nL\nL\nA\nH\nG\nL\nA\nM\nR\nO\nF\nT\nA\nL\nP\nR\nE\nS\nR\n \n1\n \nE\nT\nU\nO\nR\n\nI\n\n \n\n \n\n-\n \n\nD\nA\n\n-\n5
\n0\n0\n0\n-\n2\n2\n-\nU\nP\n:\ne\n\nl\ni\n\nF\n\n \n,\n\nP\nR\n\n \n)\n\nN\nA\nL\nL\nA\nH\nG\n \nL\nA\n\n(\n \n1\n \nE\nT\nU\n
O\nR\nD\nA\n\n \n\n-\n5\n0\n0\n0\n-\n2\n2\n-\nU\nP\n:\ne\nt\na\np\nm\ne\nT\n\nl\n\n \nt\nr\no\np\ne\nR\n\n \n,\n\nB\nL\nG\n
\n.\n)\n\nD\nE\nT\nA\nD\nP\nU\n\n(\n \n\nY\nR\nA\nR\nB\nL\nD\nA\n\nI\n\n \n\n-\n5\n0\n0\n0\n-\n2\n2\n-\nU\nP\n:\ny\nr\na\nr\nb
\nL\n\ni\n\n1\n\n2\n\n3\n\n4\n\n5\n\nRemarks:\n1. Borehole is logged according to BS\n5930:2015+A1:2020 Code of Practice for Gr
ound\nInvestigation.\n2. Deck to seabed= 13.70m\n Deck to water= 3.70m\n Time: 22:35\n\nProject:\n\nLocation:\n\nE0396 -
Long Term Offshore Power Supply (LTOPS) Project\n(Lightning Project)\nAll Ghallan Island\n\n5.00\n\nClient:\n\nJan De Nul\n\nJo
b No.:\n\nPU-22-0005-AD\n\nLogged by:\nJJK\n\nChecked by:\n\nPlate No.:\n.1\n\nSheet 1 of 12\n\nPreliminary\x0cFugro Middle Eas
t\nSector No. MN 1 Mussafah Area\nAbu Dhabi, United Arab Emirates\nP.O. Box 4447\n\nBorehole: RP-R1-BH-01\n\nEquipment:\n\nF-10
0\n\nMethod:\n\nRotary cored\n\nDrilling Fluid:\n\nGuargum\n\nDate\n\nFlush\nReturn\n(%)\n\nGraphic\nSymbol D\n\nh\nt\np\ne\n
\n)\n\nm\n\n(\n\nn\no\ni\nt\na\nv\ne\nl\nE\n\n)\n\nm\n\n(\n\nStart Date: 02/10/2022\nEnd Date: 05/10/2022\nFinal Log Issued Dat
e:\n Core Dia: (mm): 83\n\nCoordinates:\nElevation: m CD\nWater Depth (m): m\nEnd Depth: 60.00 m\n\nE 767786.00 m\n\nN 2750
334.00 m\n\nSample\nNo.\n\nSample\nType &\nDepth\n\nw\no\nB\n\nl\n\ns\nt\nn\nu\no\nC\n\nN=50\n\n SPT\n(N-Values)\nGraph\n\n10\n
\n20\n\n30\n\n40 50\n\nT\nP\nS\n\nn\no\ni\nt\na\nr\nt\ne\nn\ne\nP\n\n0 mm\n\nRock Core Quality\nRQD\nTCR\nSCR\n(%)\n(%)\n(%)\n
\ng\nn\ni\nr\ne\nh\nt\na\ne\nW\n\ne\nd\na\nr\nG\n\nTest Results of Rock\n\nEm\n(MPa)\n\nUCS\n(MPa)\n\nIs(50)\n(MPa)\n\nBulk\nDe
nsity\n(Mg/m3)\n\nFines\n(%)\n\nGrain Size\nSilt/Clay\n\nSand\n\nGravel\n100\n\n50\n\n0\n\nDescription\n\n0.00m - 6.45m : Extre
mely weak,\nthinly laminated, light brownish\ngrey to green, CALCARENITE.\nDistinctly weathered to\ndestructured. Fractures are
\nhorizontal to sub-horizontal, very\nclose to close spaced,\nundulating, rough and clean.\n(continued)\n... From 5.28m to 6.00
m : Non\nintact core.\n\n... From 6.00m to 6.30m : No\ncore recovery.\n\n... From 6.30m to 6.45m : Non\nintact core.\n\n6.45m -
12.10m : Extremely\nweak to very weak, thinly to\nmedium bedded, light brownish\ngrey to cream, CALCARENITE\nwith voids upto 2c

2. REGULAR EXPRESSIONS

1. OCR

Overview

Methodology

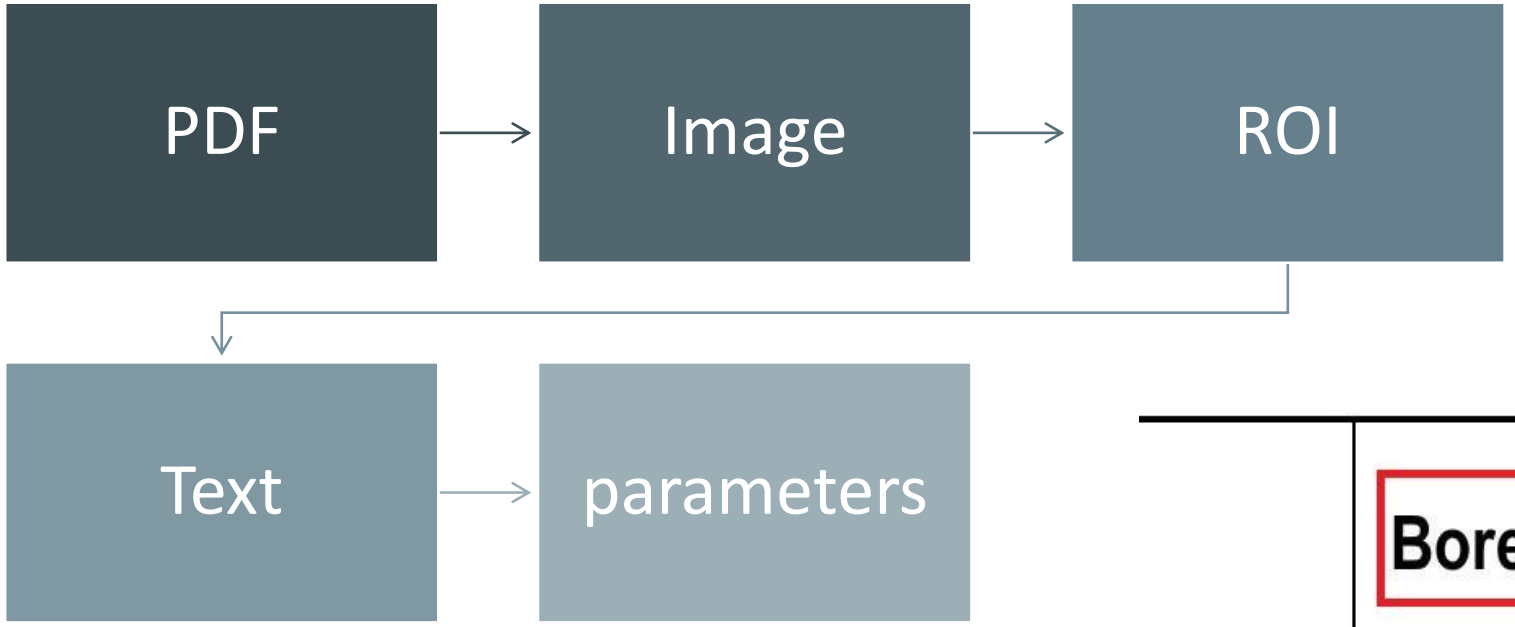
- OCR stands for "Optical Character Recognition”
- Recognizes text within a digital image
- Commonly used to recognize text in scanned documents and images

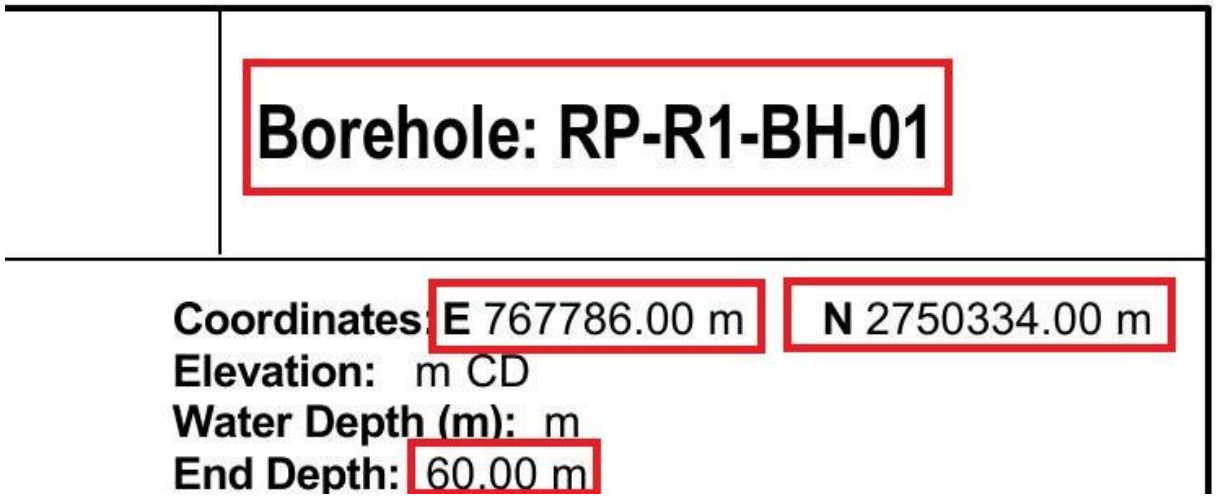
Tools

- Python & Tesseract (Open source, Google)

1. OCR

Workflow





1. OCR

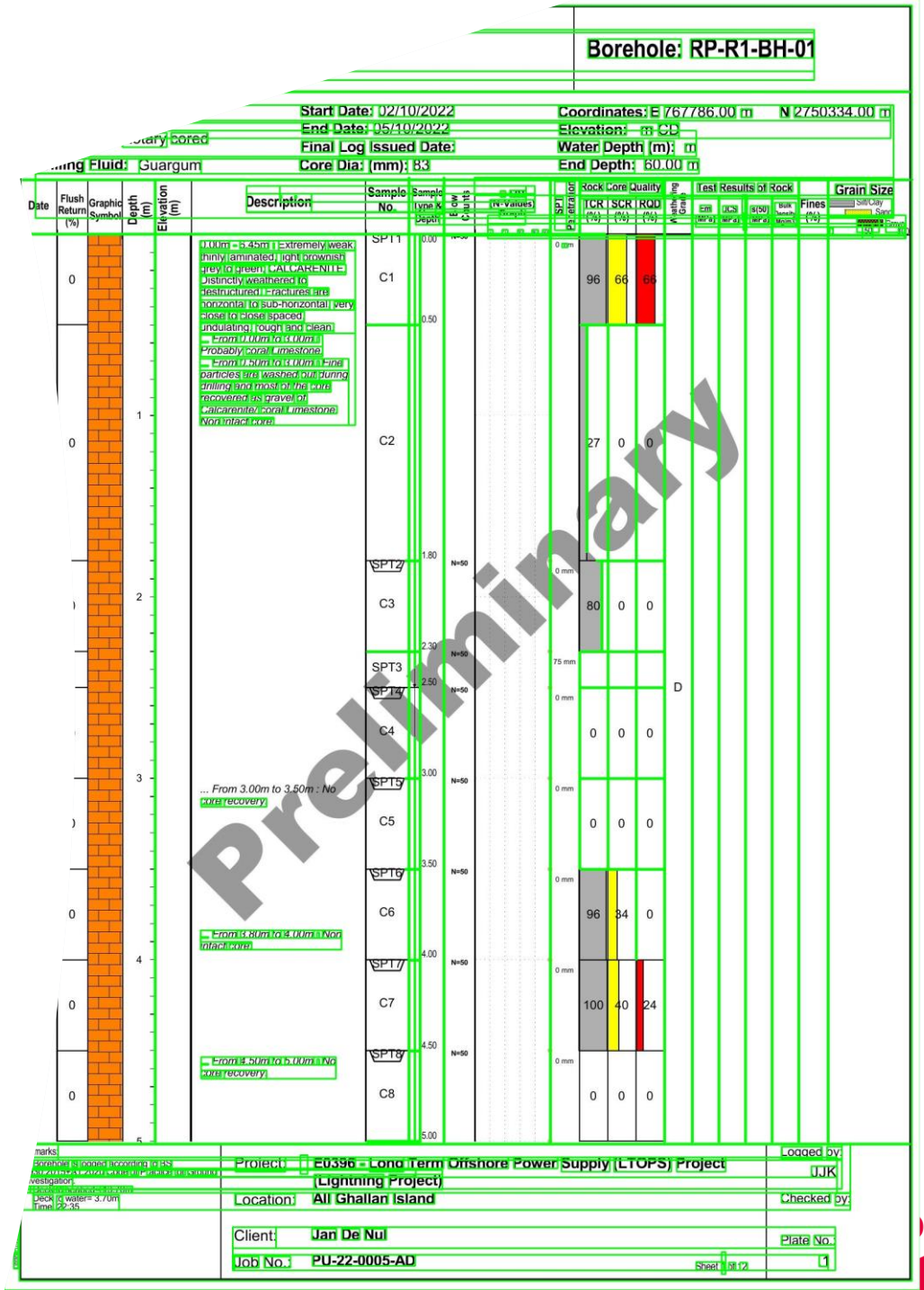
LESSONS LEARNED

Structuring

- ROI approach
- Every format needs custom ROI configuration

Digitization

- OCR is not 100% accurate
  - PDF → Image → image processing
  - '5 '→ 'S' / 'S' → '5'
  - I' → '1' / '1 → 'I'
- Language specification required
- Relative high computing power
  - PDF → Image → image processing



2. REGULAR EXPRESSIONS

Overview

Methodology

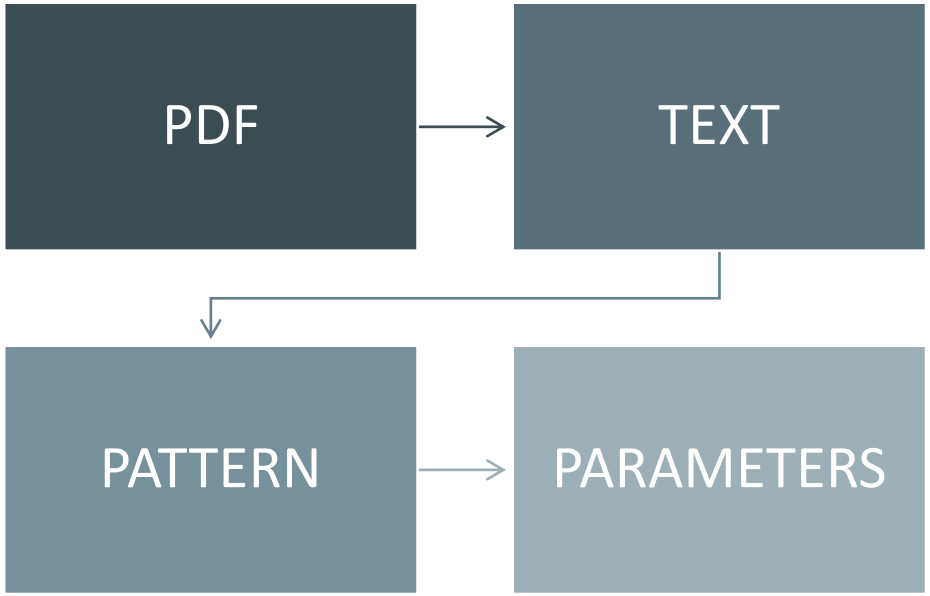
- Text extraction from pdf
- Regular Expression
- Break down document text into interpretable structure
- Data cleaning

Tools

- Python (pdfminer/PyPDF2)
- Regex

2. TEXT PATTERN EXTRACTION

Workflow



'Fugro Middle East\nSector No. MN 1 Mussafah Area\nAbu Dhabi, United Arab Emirates\nP.O. Box 4447\n\nBorehole: RP-R1-BH-01\n\nE
quipment:\n\nF-100\n\nMethod:\n\nRotary cored\n\nDrilling Fluid:\n\nGuargum\n\nDate\n\nFlush\nReturn\n(%)\n\nGraphic\nSymbol D
\n\nh\nt\np\ne\n\n)\n\nm\n\n(\n\nn\no\ni\nt\na\nv\ne\nl\nE\n\n)\n\nm\n\n(\n\nStart Date: 02/10/2022\nEnd Date: 05/10/2022\nFina
l Log Issued Date:\n Core Dia: (mm): 83\n\nCoordinates:\nElevation: m CD\nWater Depth (m): m\nEnd Depth: 60.00 m\n\nE 76778
6.00 m\n\nN 2750334.00 m\n\nSample\nNo.\n\nSample\nType &\nDepth\n\nw\no\nB\n\nl\n\ns\nt\nn\nu\no\nC\n\nSPT1\n\n0.00\n\nN=50\n
\n SPT\n(N-Values)\nGraph\n\n10\n\n20\n\n30\n\n40 50\n\nT\nP\nS\n\nn\no\ni\nt\na\nr\nt\ne\nn\ne\nP\n\n0 mm\n\nRock Core Quality
\nRQD\nTCR\nSCR\n(%)\n(%)\n(%)\n\ng\nn\ni\nr\ne\nh\nt\na\ne\nW\n\ne\nd\na\nr\nG\n\nTest Results of Rock\n\nEm\n(MPa)\n\nUCS\n(M
Pa)\n\nIs(50)\n(MPa)\n\nBulk\nDensity\n(Mg/m3)\n\nFines\n(%)\n\nGrain Size\nSilt/Clay\n\nSand\n\nGravel\n100\n\n50\n\n0\n\nDesc
ription\n\n0.00m - 6.45m : Extremely weak,\nthinly laminated, light brownish\ngrey to green, CALCARENITE.\nDistinctly weathered
to\ndestructured. Fractures are\nhorizontal to sub-horizontal, very\nclose to close spaced,\nundulating, rough and clean.\n...
From 0.00m to 3.00m :\nProbably coral Limestone.\n... From 0.50m to 3.00m : Fine\nparticles are washed out during\ndrilling and
most of the core\nrecovered as gravel of\nCalcarenite/ coral Limestone.\nNon intact core.\n\n0.50\n\nSPT2\n\n1.80\n\nN=50\n\nC3
\n\nSPT3\n\nSPT4\n\nC4\n\n2.30\n\n2.50\n\nN=50\n\nN=50\n\nSPT6\n\n3.50\n\nN=50\n\n... From 3.00m to 3.50m : No\ncore recover
y.\n\nSPT5\n\n3.00\n\nN=50\n\n... From 3.80m to 4.00m : Non\nintact core.\n\nSPT7\n\n4.00\n\nN=50\n\n... From 4.50m to 5.00m :
No\ncore recovery.\n\nSPT8\n\n4.50\n\nN=50\n\nC1\n\nC2\n\nC5\n\nC6\n\nC7\n\nC8\n\n96\n\n66\n\n66\n\n27\n\n0\n\n0\n\n0 mm\n\n75
mm\n\n0 mm\n\n0 mm\n\n0 mm\n\n0 mm\n\n0 mm\n\n80\n\n0\n\n0\n\n0\n\nD\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n96\n\n34\n\n0\n\n100\n\n40\n\n
24\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n0\n\n:\n\n \n\n/\n\n/\n\n:\n\n \n\n \n\n6\n1\n8\n3\n0\n1\n2\n2\n0\n2\n0
\n1\n6\n0\n\nn\no\n4\n0\n0\n0\n3\n8\nT\nN\ng\n \ny\nb\nd\ne\nc\nu\nd\no\nr\n\nI\n\n.\n\n.\n\n \n\n \n\n \n\n.\n\n \n\nP\n \nJ\nP
\nG\nN\nA\nL\nL\nA\nH\nG\nL\nA\nM\nR\nO\nF\nT\nA\nL\nP\nR\nE\nS\nR\n \n1\n \nE\nT\nU\nO\nR\n\nI\n\n \n\n \n-\n \n\nD\nA\n\n-\n5
\n0\n0\n0\n-\n2\n2\n-\nU\nP\n:\ne\n\nl\ni\n\nF\n\n \n,\n\nP\nR\n\n \n)\n\nN\nA\nL\nL\nA\nH\nG\n \nL\nA\n\n(\n \n1\n \nE\nT\nU\n
O\nR\nD\nA\n\n \n\n-\n5\n0\n0\n0\n-\n2\n2\n-\nU\nP\n:\ne\nt\na\np\nm\ne\nT\n\nl\n\n \nt\nr\no\np\ne\nR\n\n \n,\n\nB\nL\nG\n
\n.\n)\n\nD\nE\nT\nA\nD\nP\nU\n\n(\n \n\nY\nR\nA\nR\nB\nL\nD\nA\n\nI\n\n \n\n-\n5\n0\n0\n0\n-\n2\n2\n-\nU\nP\n:\ny\nr\na\nr\nb
\nL\n\ni\n\n1\n\n2\n\n3\n\n4\n\n5\n\nRemarks:\n1. Borehole is logged according to BS\n5930:2015+A1:2020 Code of Practice for Gr
ound\nInvestigation.\n2. Deck to seabed= 13.70m\n Deck to water= 3.70m\n Time: 22:35\n\nProject:\n\nLocation:\n\nE0396 -
Long Term Offshore Power Supply (LTOPS) Project\n(Lightning Project)\nAll Ghallan Island\n\n5.00\n\nClient:\n\nJan De Nul\n\nJo
b No.:\n\nPU-22-0005-AD\n\nLogged by:\nJJK\n\nChecked by:\n\nPlate No.:\n.1\n\nSheet 1 of 12\n\nPreliminary\x0cFugro Middle Eas
t\nSector No. MN 1 Mussafah Area\nAbu Dhabi, United Arab Emirates\nP.O. Box 4447\n\nBorehole: RP-R1-BH-01\n\nEquipment:\n\nF-10
0\n\nMethod:\n\nRotary cored\n\nDrilling Fluid:\n\nGuargum\n\nDate\n\nFlush\nReturn\n(%)\n\nGraphic\nSymbol D\n\nh\nt\np\ne\n
\n)\n\nm\n\n(\n\nn\no\ni\nt\na\nv\ne\nl\nE\n\n)\n\nm\n\n(\n\nStart Date: 02/10/2022\nEnd Date: 05/10/2022\nFinal Log Issued Dat
e:\n Core Dia: (mm): 83\n\nCoordinates:\nElevation: m CD\nWater Depth (m): m\nEnd Depth: 60.00 m\n\nE 767786.00 m\n\nN 2750
334.00 m\n\nSample\nNo.\n\nSample\nType &\nDepth\n\nw\no\nB\n\nl\n\ns\nt\nn\nu\no\nC\n\nN=50\n\n SPT\n(N-Values)\nGraph\n\n10\n
\n20\n\n30\n\n40 50\n\nT\nP\nS\n\nn\no\ni\nt\na\nr\nt\ne\nn\ne\nP\n\n0 mm\n\nRock Core Quality\nRQD\nTCR\nSCR\n(%)\n(%)\n(%)\n
\ng\nn\ni\nr\ne\nh\nt\na\ne\nW\n\ne\nd\na\nr\nG\n\nTest Results of Rock\n\nEm\n(MPa)\n\nUCS\n(MPa)\n\nIs(50)\n(MPa)\n\nBulk\nDe
nsity\n(Mg/m3)\n\nFines\n(%)\n\nGrain Size\nSilt/Clay\n\nSand\n\nGravel\n100\n\n50\n\n0\n\nDescription\n\n0.00m - 6.45m : Extre
mely weak,\nthinly laminated, light brownish\ngrey to green, CALCARENITE.\nDistinctly weathered to\ndestructured. Fractures are
\nhorizontal to sub-horizontal, very\nclose to close spaced,\nundulating, rough and clean.\n(continued)\n... From 5.28m to 6.00
m : Non\nintact core.\n\n... From 6.00m to 6.30m : No\ncore recovery.\n\n... From 6.30m to 6.45m : Non\nintact core.\n\n6.45m -
12.10m : Extremely\nweak to very weak, thinly to\nmedium bedded, light brownish\ngrey to cream, CALCARENITE\nwith voids upto 2c

LET'S BUILD IQIP

LET'S BUILD

2. TEXT PATTERN EXTRACTION

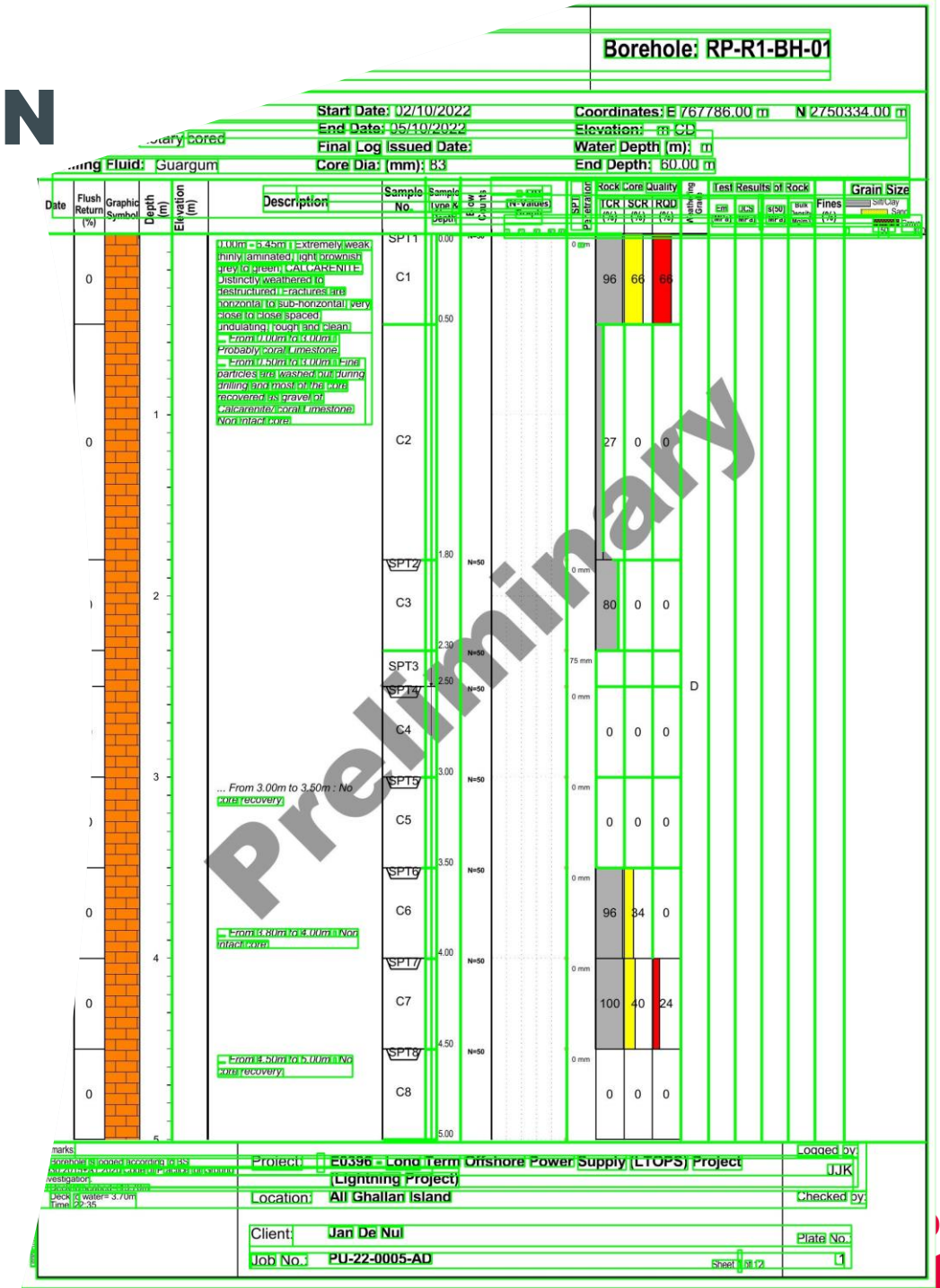
LESSONS LEARNED

Text extraction

- Not all pdf's have a text layer that can be extracted
- Not all text is in the text layer
- Only applicable to PDF

Regular expression

- 100% accurate
- Highly depended on text structure (patterns)
- Patterns unique for every layout



Conclusion

Project summary

Key findings:

OCR vs Regular Expression

- OCR not 100% accurate
- ROI selector is essential
- OCR widely applicable (scans, images, pdf)

- Regular Expression only applicable with ‘readable’ PDF’s
- Regular Expressions highly depend on text structure

High variety of source documents limits scalability

The process has become party automated due to manual selection of ROI and Regular Expressions