Implementation of Machine Learning Systems to Enhance the Value of the CDA North Sea Data Set*

Philip Neri1

Search and Discovery Article #42132 (2017)**
Posted September 18, 2017

*Adapted from oral presentation given at 2017 AAPG Geoscience Technology Workshop, Big Data & Deep Learning in the Oil Industry: Basic Applications, Houston, Texas, May 22, 2017
**Datapages © 2017 Serial rights given by author. For all other rights contact author directly.

1AgileDD  (pneri@me.com)

Abstract

The CDA maintains a collection of well and seismic data submitted by the UKCS operators since the early days of the North Sea Exploration and Production in the 1960’s. The collection of CDA well data has been made available to operators and authorities as a database of 11,500 well headers and as a set of 450,000 documents under various formats such as .pdf, .xls, .doc, .tiff, .jpg, .las, .dlis.

This collection of data is similar in its organization and content with legacy datasets that can be found in any industry: around 20% of the information is available in a structured form such as a relational database, and 80% in a semi-structured or unstructured form, typically grouped in folders containing various documents formatted as described above. Since most of the software and data management tools used in E&P can only access the structured information and in some cases some half-structured formats, it transpires that E&P decisions are based on a small part of the available stored information.

The low benchmark of 20% of available data is due to several factors, primarily the cost of indexing (classifying the documents per topic) and cataloguing the documents (extracting metadata from the document) which is currently a work-intensive process. But the cost is not the only limitation. The fixed nature of most of the subsurface data-models makes it almost impossible to catalog information which was not planned to be extracted in the initial stage of the data model design.
In 2016, the CDA launched a challenge to find new ways to extract value from its unstructured data assets. This article explores the application of newly developed Machine Learning Systems (MLS) to automate part of the indexing and cataloguing. MLS demonstrated a reduced time (and therefore cost) of access to information but also enriched the extracted information by qualifying its extraction confidence and source, and identifying replicates. They make it possible to perform data analysis of larger datasets in term of volume and variety.

The performance of Machine Learning Systems when applied to subsurface data management will be discussed, the limitation criteria listed, and some future possibilities to overcome the current limitations will be overviewed.
Implementation of machine learning systems to enhance the value of the CDA North Sea data set

Philip Neri
Presenter’s notes: In summer 2016 AgileDD, together with 7 other technology providers, was selected to participate in the first CDA unstructured data challenge. The idea was to illustrate how new technology could “enrich one of the world largest collections of subsurface O&G data.

This presentation will show you what we have achieved in a short period of 4 weeks in August 2016.
Presenter’s notes: In summer 2016 AgileDD, together with 7 other technology providers, was selected to participate in the first CDA unstructured data challenge. The idea was to illustrate how new technology could “enrich one of the world largest collections of subsurface O&G data.

This presentation will show you what we have achieved in a short period of 4 weeks in August 2016.
The CDA well data set

2 relational tables
- Well header
- CS8 file index

450,000 unstructured files
- End of well reports
- Wirelines, MWD
- Composites
- CCA and SCAL
- Geochemistry
- Fluid analysis
- Biostratigraphy
- Lithostratigraphy
- Cementation report
- Checkshots and VSP
- ...etc...

Presenter’s notes: CDA is a significant and vitally important component of the private/public partnership that constitutes the NDR (National Data Repository) for the UKCS

The CDA has collected information about 11,500+ offshore wells drilled over the last 60 years

This information is under the form of
450,000 unstructured files
The 80/20 ratio

**20% structured information**
- Easy to access
- Easy to query
- Easy to QC
- Easy to model and analyze

**80% unstructured information**
- Costly to index
- Costly to extract metadata
- Difficult to access and query
- Cannot feed analytical tools

Presenter’s notes: The CDA data set is not so different from many others we can find in our industry and in some other industries. According to various sources (Merryl Lynch, EMC, Oracle ...) 80 % of the information is available in an unstructured format (PDF, TIFF ...)

This type of information is difficult to use, it cannot be used directly by the modeling of BI tools. Extracting metadata to create indexes and populate DBs is extremely costly and need SMEs that today are increasingly scarce in our industry.

This translates into the fact that only 20% of the information is available within structured database.

The big problem is that since this small part is easy to access and query, it trends to be the unique source of information to base decision
Your decision is at risk

- 20% structured information
  - Easy to access
  - Easy to query
  - Easy to QC
  - Easy to analyze

- 80% unstructured information
  - Costly to index
  - Costly to extract metadata
  - Difficult to access and query
  - Cannot feed analytical tools

Presenter’s notes: And that makes your decision a very risky one.
What if we reverse the ratio?

80% structured information

- Easy to access
- Easy to query
- Easy to QC
- Easy to analyze
- Easy to source
- With a known confidence level

20% unstructured information

- Automatic cataloging and indexing

Make your decision more reliable based on more structured data

Presenter’s notes: Our ambition is to reverse the ratio. Using an automated process based on a machine learning system, we extract more information from your documents at a fraction of the cost and time, and this makes your decisions more reliable.
The advantages of automated cataloging

**SAVE MONEY**
Avoid populating databases manually

**GO FASTER**
From data to decision

**DE-RISK**
Using more verified information
Why use a machine learning system?

Presenter’s notes: Supervised Machine Learning excels at recognizing a pattern in an unstructured context
ML exceeds manual and full text indexing

- **Manual indexing**
  - Cannot process a lot of documents in a short period of time

- **OCR + Full text indexing**
  - Cannot automatically extract metadata not previously known in lists, dictionaries, taxonomies ...

- **OCR + Full text indexing + Machine learning**
  - Because the ML searches for context around the metadata, any text and numerical variable can be detected

Presenter’s notes: The capacity to detect the pattern around a target metadata item allows us to make a more efficient cataloging compared to a full text indexing. ML detects a numerical variable such as a coordinate, a depth, a temperature using the context of the value which is not possible even with the best Full text indexing.
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:

- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:

- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.

- Detect some well header metadata, text or numerical, using their context

- Associate a confidence factor to each extraction

- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
Example of indexing using the CS8 taxonomy and extraction of some well header metadata

Presenter’s notes:
These few snapshots illustrate the capacity of our tools to:
- Detect the document category using the CS8 taxonomy used in the UK to describe each subsurface document. The CS8 taxonomy defines a document according to its “container” such as a report, log, digital document ... and its “contents” such as mud-logging, petrophysics, seismic, engineering ... Does the same automatically after training.
- Detect some well header metadata, text or numerical, using their context
- Associate a confidence factor to each extraction
- Display the variability of a particular metadata value for the same well
How does it work?

Unstructured and semi-structured documents

Pre-OCR processing

OCR

Post-OCR processing

Text searchable

A uniform text layout representation

Parallelized OCR

User provided taxonomy (option)

Seeds db (option)

Seeds docs (option)

Heuristic labelling (option)

Learning models

iQC Machine Learning

Documents classification

iQC datalake

Metadata extraction

GUI

QC Data edition

Validation

Refutation

Publication

Structured information
In 2 days, iQC has been trained to use the CS8 taxonomy, using 2000 “seed” examples.
This initial training was sufficient to have an 80% match with the manual indexing done over many years on the 450,000 documents.

The ratio moved up to 90% with a seed of 5,000 documents.

The 10% of remaining discrepancies relate to “manual errors” or rare and ambiguous taxonomy classes.
This initial training was sufficient to have an 80% match with the manual indexing done over many years on

relate to "manual errors" or rare and ambiguous taxonomy classes.
QC of the CDA tables according to the source documents

<table>
<thead>
<tr>
<th>Well</th>
<th>Attribute</th>
<th>Value</th>
<th>Confid CDA-wellare</th>
<th>Original operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>209/12-1</td>
<td>Operator Marathon Oil U.K. Ltd.</td>
<td>100  209/12-1</td>
<td>MARATHON OIL (UK) LTD</td>
<td>MARATHON OIL (UK) LTD</td>
</tr>
<tr>
<td>209/3-1a</td>
<td>Operator MOBIL NORTH SEA LTD.</td>
<td>100  209/03-1a</td>
<td>MOBIL NORTH SEA LIMITED</td>
<td>MOBIL NORTH SEA LIMITED</td>
</tr>
<tr>
<td>209/4-1a</td>
<td>Operator North Sea San Oil Company Ltd</td>
<td>100  209/04-1a</td>
<td>AGIP (U.K.) LIMITED</td>
<td>AGIP (U.K.) LIMITED</td>
</tr>
<tr>
<td>209/6-1</td>
<td>Operator Gulf U.K.</td>
<td>100  209/06-1</td>
<td>CHEVRON U.K. LIMITED</td>
<td>CHEVRON U.K. LIMITED</td>
</tr>
<tr>
<td>213/23-1</td>
<td>Operator MOBIL NORTH SEA LTD</td>
<td>100  213/23-1</td>
<td>MOBIL NORTH SEA LIMITED</td>
<td>MOBIL NORTH SEA LIMITED</td>
</tr>
<tr>
<td>213/26-1a</td>
<td>Operator chevron</td>
<td>76  213/26-1</td>
<td>CHEVRON</td>
<td>CHEVRON</td>
</tr>
<tr>
<td>217/25-1x</td>
<td>Operator Chevron North Sea Ltd.</td>
<td>100  217/25-1x</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
</tr>
<tr>
<td>220/26-1</td>
<td>Operator B.P. PETROLEUM DEVELOPMENT</td>
<td>100  220/26-1</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
</tr>
<tr>
<td>220/26-1st</td>
<td>Operator B.P. PETROLEUM DEVELOPMENT</td>
<td>100  220/26-1st</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
</tr>
<tr>
<td>220/26-2</td>
<td>Operator BP Petroleum Development Ltd</td>
<td>100  220/26-2</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
</tr>
<tr>
<td>747-1</td>
<td>Operator Gas Council (Exploration) Ltd</td>
<td>100  747-1</td>
<td>BG INTERNATIONAL LTD</td>
<td>BG INTERNATIONAL LTD</td>
</tr>
<tr>
<td>83/24-1</td>
<td>Operator DEPARTMENT OF ENERGY</td>
<td>100  83/24-1</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
</tr>
<tr>
<td>86/17-1</td>
<td>Operator Murphy Petroleum Limited</td>
<td>100  86/17-1</td>
<td>MURPHY PETROLEUM LIMITED</td>
<td>MURPHY PETROLEUM LIMITED</td>
</tr>
<tr>
<td>86/18-1</td>
<td>Operator The British National Oil Corporation</td>
<td>100  86/18-1</td>
<td>MURPHY PETROLEUM LIMITED</td>
<td>MURPHY PETROLEUM LIMITED</td>
</tr>
<tr>
<td>87/72-1a</td>
<td>Operator BP Petroleum Development Ltd</td>
<td>100  87/72-1a</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
<td>BP EXPLORATION OPERATING COMPANY LIMITED</td>
</tr>
<tr>
<td>87/16-1</td>
<td>Operator Department of Energy, UK</td>
<td>100  87/16-1</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
</tr>
<tr>
<td>87/16-1</td>
<td>Operator Department of Energy, U.K.</td>
<td>100  87/16-1</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
</tr>
<tr>
<td>88/2-1</td>
<td>Operator Department of Energy, U.K.</td>
<td>100  88/2-1</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
<td>HMG DEPARTMENT OF TRADE AND INDUSTRY</td>
</tr>
<tr>
<td>89/15-1</td>
<td>Operator UNICAL NORTH SEA LTD</td>
<td>100  89/15-1</td>
<td>UNICAL NORTH SEA EXPLORATION LIMITED</td>
<td>UNICAL NORTH SEA EXPLORATION LIMITED</td>
</tr>
<tr>
<td>89/15-2a</td>
<td>Operator Total Exploration and Production UK Ltd</td>
<td>100  89/15-2a</td>
<td>TOTAL</td>
<td>TOTAL</td>
</tr>
<tr>
<td>89/20-1</td>
<td>Operator App (UK) Ltd</td>
<td>100  89/20-1</td>
<td>ASIP (U.K.) LIMITED</td>
<td>ASIP (U.K.) LIMITED</td>
</tr>
<tr>
<td>89/25e-1</td>
<td>Operator nautical petroleum</td>
<td>100  89/25e-1</td>
<td>NAUTICAL</td>
<td>NAUTICAL</td>
</tr>
<tr>
<td>89/30-1</td>
<td>Operator Sovereign Oil &amp; Gas Plc</td>
<td>100  89/30-1</td>
<td>SANDERSON'S EXPLORATION LIMITED</td>
<td>SANDERSON'S EXPLORATION LIMITED</td>
</tr>
<tr>
<td>89/4-1</td>
<td>Operator Chevron UK Ltd</td>
<td>100  89/4-1</td>
<td>CHEVRON U.K. LIMITED</td>
<td>CHEVRON U.K. LIMITED</td>
</tr>
</tbody>
</table>

Presenter’s notes: The automatic well header metadata extraction authorize to QC the CDA DB according to the source documents. This example shows the discrepancies on a text metadata (the well operator)
Presenter’s notes: Our capacity to extract numerical variables allowed us to QC the CDA DB according to the sources docs. It showed surprising “human errors” on some very easy to detect metadata. The main value is not only to alert on discrepancies but also to show immediately the documented source of information we have used to alert us, as well as our confidence in the automatic detection.
Presenter’s notes: Our capacity to extract numerical variables allowed us to QC the CDA DB according to the sources docs. It showed surprising “human errors” on some very easy to detect metadata. The main value is not only to alert on discrepancies but also to show immediately the documented source of information we have used to alert us, as well as our confidence in the automatic detection.
Lessons learnt from the CDA challenge

• Machine learning detects metadata in unstructured documents that other methods cannot detect
• It supports the QC of a structured database using unstructured sources
• It makes it possible to easily extend the contents of the database “on demand”
• For us, the CDA challenge was also an opportunity to enrich our learning model and make it more stable

Presenter’s notes: Our capacity to extract numerical variables allowed us to QC the CDA DB according the sources documents. It showed surprising “human errors” on some very easy to detect metadata. The main value is not only to alert on discrepancies but also to show immediately the documented source of information we have used to alert as well as our confidence in the automatic detection.
The (machine) learns with us!

Overview

Tom M. Mitchell provided a widely quoted, more formal definition: "A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$ if its performance at tasks in $T$, as measured by $P$, improves with experience $E$.\textsuperscript{[9]} This definition is notable for its defining machine learning in fundamentally operational rather than cognitive terms, thus following Alan Turing’s proposal in his paper "Computing Machinery and Intelligence" that the question "Can machines think?" be replaced with the question "Can machines do what we (as thinking entities) can do?\textsuperscript{[10]}"
The (machine) learns with fresh data!

Looking ahead
• We are looking for documents to crunch
• We are engaging with early adopters
  • To perform pilot projects and feasibility studies system relative to their needs and objectives
  • To evaluate the performance of our machine learning
• We work in the cloud (Microsoft Azure®)
  • Our Learning models improve continuously
  • New users benefit from all accrued learning
  • An alternative configuration is to install our system locally
Most promising IT & web startup 2017

philip.neri@agileddd.com
www.agileddd.com