PNEC Conferences

Houston, May 11th 2017

Implementation of machine learning systems to enhance the value of the CDA North Sea data set

Philip Neri, AgileDD

Henri Blondelle, AgileDD

Executive Summary

The CDA[1] 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 to 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 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 paper explores the application of Machine Learning Systems (MLS) recently developed by Agile Data Decisions to automate part of the indexing and cataloguing. iQC, the AgileDD MLS demonstrated a reduced time (and therefore cost) of access to information

[1] CDA is a wholly-owned subsidiary of Oil & Gas UK. The company was established by industry in 1995, more information at http://cdal.com

while also enriching the extracted information by qualifying the level of confidence of the extraction and its source, and identifying replicates. This makes it possible to perform data analysis on larger datasets in terms of volume and diversity.

The performance of Machine Learning Systems when applied to subsurface data management are discussed, together with the listing of some limitations and an overview of some future possibilities to overcome the current limitations.

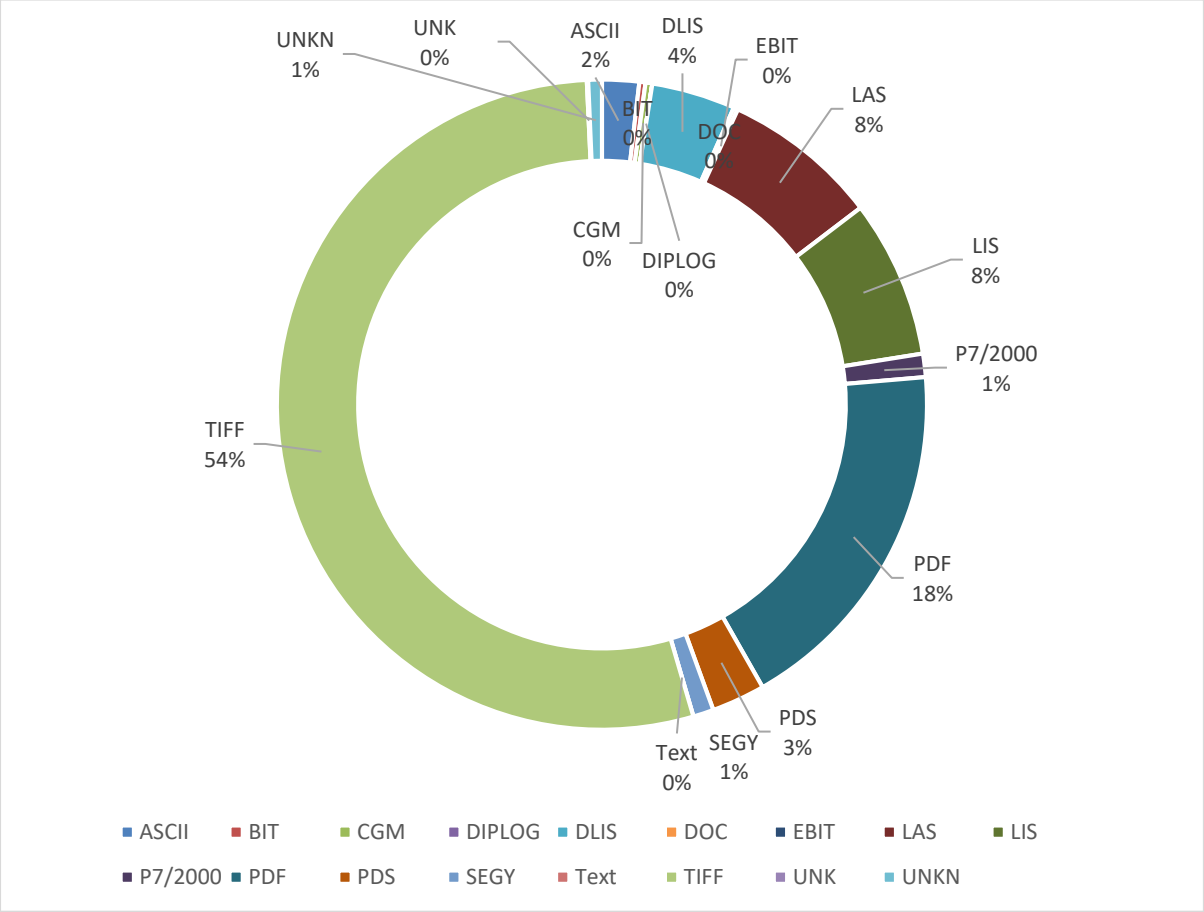
Data was always critical to E&P operations, it is now vital!

It could be said that subsurface data has always been critical for E&P operations," from the drill floor to the top floor" but the sharp fall in oil prices in 2014 makes it even more vital for the industry to fully leverage existing information to maximize the recovery of producing fields and improve the performance of exploration drilling activities.

In a study entitled "The business value case for data management – a study", (Hawtin and Lecore, 2011) commissioned by Common Data Access, one conclusion was that between 25% and 33% of oil company value creation can be reasonably estimated to be attributed to data. In mature exploration and production provinces, such as the North Sea, as well as in frontier areas, such as the Atlantic Margins or the Irish Sea, the data is the foundation of value creation for O&G companies, and this involves existing as well as new acquired data. There are hundreds of thousands of files and documents which have been collected over the past decades.

To maximize this value creation from knowledge currently stored in unstructured formats, new tools are needed that can cost-effectively identify, extract and make the required data accessible.

Accessing the unstructured documents may not be the most difficult task in a country encouraging competition in the development of their subsurface acreage. On the UK continental shelf, the CDAL releases well header and associated documents to its members. While data in LAS, LIS and DLIS file formats easily transfer into the various structured database systems and application repositories used by operators, vast quantities (76% of the CDA's 450,000 document collection) of unstructured files and documents consisting of scans of paper documents in TIFF, JPG or PDF formats, and associated reports and tabulations in Microsoft Excel and Word formats (XLS and DOC formats respectively) do not fit into the databases and are typically stored in a tree of hierarchical folders that make it possible to locate them, but not directly use their contents.



CDA subsurface document collection per format

To reduce the risk associated with decisions all along the E&P value chain, the rich content of these very unstructured files cannot be neglected. They are the only source of information about basin and reservoir properties for geoscientists in charge of exploration, development or production.

Therefore, until recently, the extraction of this information, done as a cataloging of metadata, was performed by data-managers and Subject Matter Experts (SMEs) using some manual processes and on some occasions OCR (Optical Character Recognition) systems and full text indexing. But the cost and the slowness of this type of process results in a ratio between structured information and structured information rarely surpassing 20/80.

In the case of the CDA document collection, it could be estimated that a basic cataloging of the more important data could cost between $3M to $4.5M considering a productivity of 120 to 180 documents per day per person.

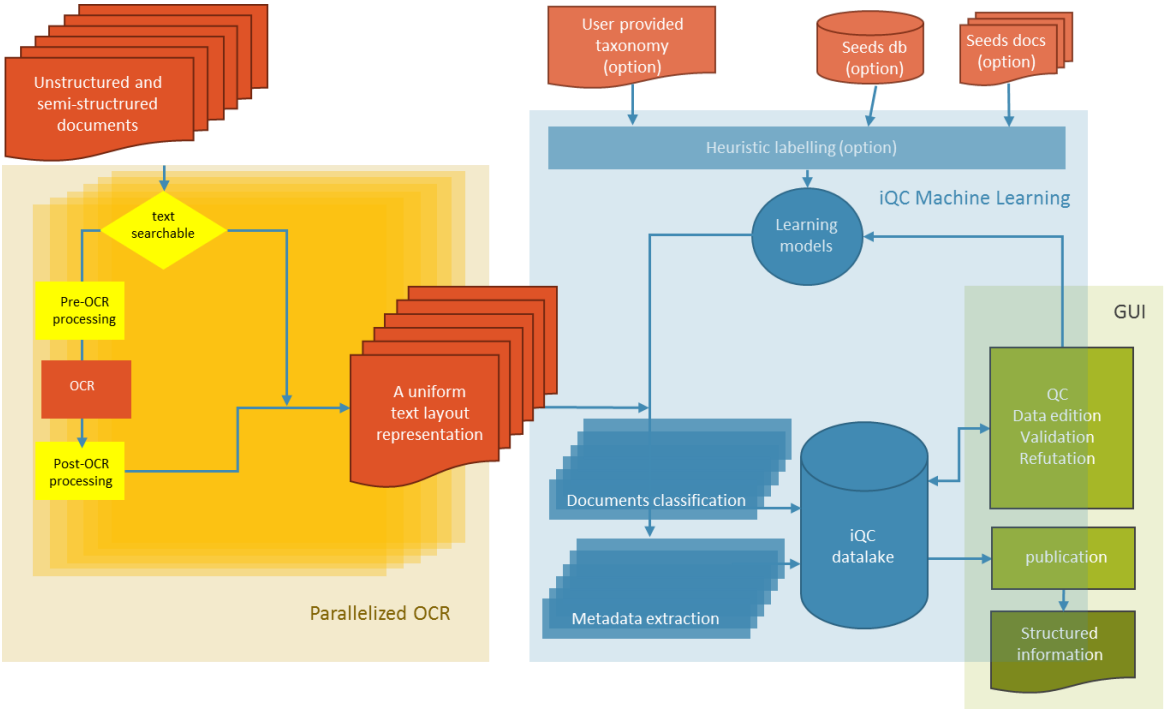
In addition to the current acute cost controls prevalent in the industry, we have also to consider today the dwindling availability of resource such as SMEs able to execute this task with a minimal amount of errors.

Automation seems to be the only cost-effective and time-efficient solution to this problem, but the variability of the documents, of the quality of scanning and the organization of material within reports or spreadsheets have been a formidable challenge to achieving a computerized solution.

An emerging technology: Machine Learning

The variability of documents is significant, but still subject to patterns and similarities that make them decipherable to the human eye. Therefore, a pattern recognition technology that aims at replicating human behaviour is well-suited to solving the problem. The process used by humans to approach the task of identifying and verifying data for extraction is empirical. A computer system that teaches itself to emulate the decision process of the human mind will be much quicker at converging on a usable solution than the traditional process of explicitly programming an extraction system.

Machine Learning is one of the fastest-growing domains of computer technology in recent years. Recognizing the potential for Machine Learning technologies to revolutionize data indexing and cataloguing for E&P files and documents, AgileDD has developed a Machine Learning System (MLS), named iQC, that could be taught and operated by data managers and SMEs to handle the cataloguing of data and indexing using metadata. As the name implies, an MLS learns without being explicitly programmed. The learning process itself is supervised by one or more data management experts who initially perform the task to be automated, in our case, the extraction of information from scanned documents and Microsoft Office files. The MLS observes their actions and decisions and builds a reference system that models these actions. During subsequent automated operations by the MLS, experts modify or invalidate any wrong decisions made by the MLS. These corrections are registered by the MLS and update the model, which over time increases the reliability and success ratio of the system. Our primary goals were to both significantly lower costs and ensure execution within a time span consistent with a project aiming at a review of a field or exploration area.



Machine learning surpasses the results of alternative solutions

Machine learning is not the only solution to extract metadata from text documents, even deeply unstructured. For decades now a full text indexing after an OCR makes it easy to search for key words in the text on the condition that you know the key words to search for. In addition to full indexing, we have also seen some very elegant characterisations of documents (or subparts of the documents' ontologies) which make it possible to query the documents using geological operators (formation names, chronostratigraphic unit names …) and even some queries using natural languages.
But all these techniques face a strong limitation: they can only propose pages, or locations within a page of text known by the user. As an example, you can search wells per operator in the North Sea under the condition that you know the list of North Sea operators over the last 60 years. In addition, these search methods are non-operative if somebody wants to search for numerical variables. It is impossible to query a full text index for wells deeper than 3000m.

The ability to use the context of the metadata allows us to overcome these two limitations. A well total depth is not detected by a MLS according to its value but according to the value's context. This capacity to recognize text or graphic patterns or a combination of both makes it possible to detect more metadata, identify unexpected metadata values and evaluate a confidence coefficient associated to the detection.

Cannot process a lot of documents in a short period of time

Can automatically extract metadata already known in lists, dictionaries, taxonomies ...

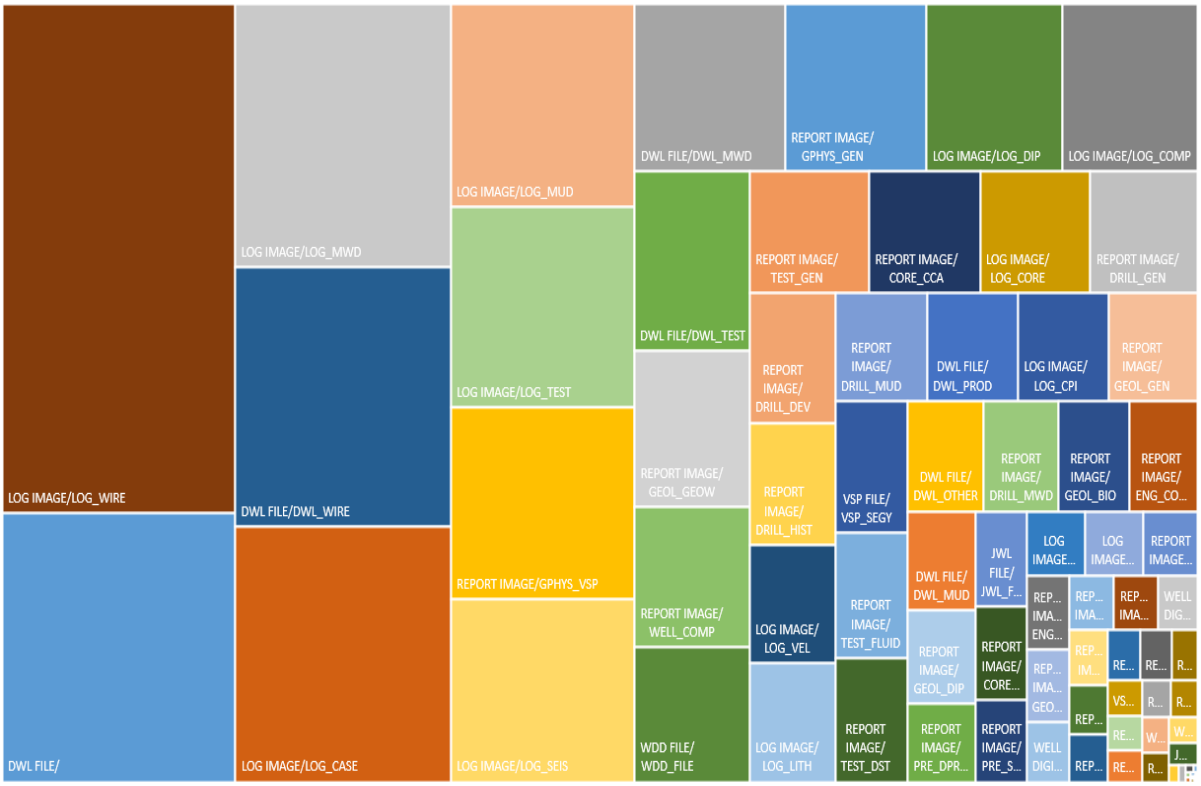
Because the ML searches for context around the metadata, text and numerical variables can be detected

CDA data challenge results

Our objectives for the CDA challenge were several:

- Benchmarking the automatic document indexing using the CS8 taxonomy done by iQC versus the manual indexing done over the last 60 years by operators and CDA data-managers
- QCing the well header metadata extracted automatically by iQC versus the cataloguing (metadata extraction) done over the last 60 years by operators and CDA data-managers.
- Evaluate the possibility to enrich rapidly the CDA structured database using the unstructured information.

The first objective (automatic indexing) was reached very quickly. We started with only 2,000 seeds (4.4% of the full document set), in other terms 2,000 documents representing all the taxonomy classes. Using these 2,000 seeds a heuristic learning model was created in less than 2 days. A first run of iQC using this first version of the learning model was assigned to detect the document category on several UKCS blocks. It showed that 80% of the document categories detected automatically matched the document indexing done manually over the past decades.

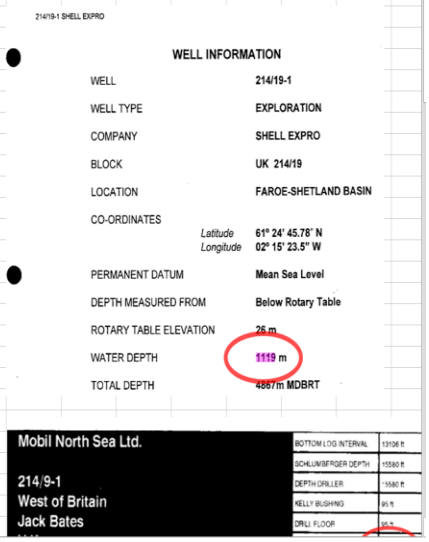


A reason for the discrepancy between the two indexing methods was found to be the difficulty to correctly represent the under-used classes in the seed dataset (bottom right corner of the figure above. By increasing the seed data set to 5,000 documents we reached a matching ration exceeding 90%. At this level of similarity, most of the discrepancies were explained by manual indexing errors and ambiguous indexing (some documents could be assigned to several categories)

The metadata detection and extraction was benchmarked on two types of metadata: the text metadata such as the well operator at the time of drilling and numerical metadata such as the wellbore total depth. Despite text metadata such as the well operator being searched for using a query on the full text index after the document OCR, the iQC version used for the challenge used only metadata detection by a machine learning model trained on the metadata surrounding context characterisation. With a heuristic training on around 1,000 seeds (documents for which the metadata value was known), we observed a result of around 80% of matching between the automatized and manual cataloguing. But this time we observed that the number of human errors in the CDA tables was greater than for the document category.

The short duration of the challenge didn't allow us to improve the cataloguing learning model and immediately measure the improvement. Nevertheless, we noticed that the tuning of our learning model done during the challenge has participated in improving the overall quality of the learning models.

iQC wellbore	title	iQC val	unit	confidenc	CDA wellname	meters	CDA Water Depth	Delta	
214/17-1	Water Depth	3832	ft	80	214/17- 1	1167.99	1167.99	0.00	
214/19-1	Water Depth	1119	m	100	214/19- 1	1119.00	1000.05	118.95	Water Depth'!M3
214/4-1	Water Depth	5317	ft	60	214/04- 1	1620.62	1620.62	0.00	
214/9-1	Water Depth	5089	ft	60	214/09- 1	1551.13	1556	4.87	Water Depth'!M23
1/4-1	Water Depth	506	ft	100	1/04- 1	154.23	154.23	0.00	
1/4-2	Water Depth	479	ft	100	1/04- 2	146.00	146	0.00	
18/3-1	Water Depth	263	ft	100	18/03- 1	80.16	80.1624	0.00	
18/5-2	Water Depth	92.6	ft	100	18/05- 2	28.22	119	90.78	Water Depth'!M32
26/12-1	Water Depth	144	ft	100	26/12- 1	43.89	43.89	0.00	
26/14-1	Water Depth	225	ft	100	26/14- 1	68.58	68.58	0.00	
26/4-1	Water Depth	219	ft	100	26/04- 1	66.75	67.06	0.31	
26/7-1	Water Depth	64.9	m	100	26/07- 1	64.90	64.9	0.00	
26/8-1	Water Depth	186	ft	100	26/08- 1	56.69	56.69	0.00	
27/10-1	Water Depth	9/18/1900	ft	100	27/10- 1	79.86	79.86	0.00	
27/3-1	Water Depth	246	ft	100	27/03- 1	74.98	74.98	0.00	
4/26-1A	Water Depth	351	ft	100	4/26- 1A	106.98	132.89	25.91	Water Depth'!M51
4/26-2	Water Depth	12/7/1900	ft	100	4/26- 2	104.24	104.24	0.00	
7/16-1	Water Depth	417	ft	100	7/16- 1	127.10	127.1	0.00	
102/28-2	Water Depth	315	\N	100	102/28- 2	96.01	96.01	0.00	
103/1-1	Water Depth	316	ft	100	103/01- 1	96.32	96.32	0.00	
103/1-2	Water Depth	314	ft	100	103/01- 2	95.71	#N/A	#N/A	
103/18-1	Water Depth	232	\N	100	103/18- 1	70.71	70.71	0.00	
103/1a-2	Water Depth	314	ft	26	103/01a- 2	95.71	98.45	2.74	
103/2-1	Water Depth	375	ft	100	103/02- 1	114.30	114.3	0.00	
103/21-1	Water Depth	276	ft	100	103/21- 1	84.12	84.12	0.00	
106/18-1	Water Depth	299	ft	100	106/18- 1	91.14	91.14	0.00	



We noticed also that the efficiency of the learning models built to detect a particular metadata was not the same and didn't improve at the same speed with experience. By the way, the subsurface data managers know that some metadata are more difficult to "catch" than others. As an example, a casing shoe may be found in the middle of a table with a high level of variability of the surrounding context.

Creating heuristic learning models using seeds "on demand" for some "requested" metadata has a strong advantage over the attempt to extract all the "enterprise metadata" at the time of the manual cataloguing. It makes it possible to delay some extractions to the time when it will be necessary and avoids the investment of scarce human resource to catalogue information which will not be immediately used by geoscientists and engineers. This approach is not economy viable using manual processes since it forces the operators to repeat the labour-intensive document reading process. In iQC this task is done only once for the first metadata detection and not redone for the next one. iQC remember the structure of the documents read in the past and therefore does not need to read the same docuemnts again. This is also a way, in addition to the learning model tuning, to capitalise on the past user experience and to make the tool more efficient and faster with time.

Limitations

The MLS system has been successfully used in other parts of the world where English is the working language (e.g. Australia), because terminology is mostly identical and many of the documents have similar layouts and flows. The situation is more complex for countries or regions with dominant languages other than English (e.g. Latin America – Spanish/Portuguese) or where both the language and the alphabet are a challenge (e.g. Russian and Cyrillic in Russia). For the same alphabet, it is not necessary to start a new

Learning Model, and we have made good progress with French and Spanish data sets. Tests on Cyrillic are due to start later this year.

A limitation lies in the quality of the scans, which affects the performance of the OCR system, and also the use of handwritten data fields that are much more error-prone and, in many instances, impossible to be reliably identified and read. Our experience to date shows that, using the MLS, we have a reliable outcome for 70-80% of the data. The remaining 20-30% is completed by our data management experts using established and robust traditional manual methods.

Some current limitations such as the difficulty to extract metadata from large and complex tables (eg: poroperm values from CCA or SCAL core measurement tables) can be overcome by an improved implementation of the MLS library that we use today (Python) while some other items such as identifying an oil or gas show in a composite log and defining its depth could be solved using some graphic pattern recognition and a deep learning library we plan to implement in the near future.

Conclusion

The iQC Machine Learning System developed by Agile DD brings together an essential combination of data management expertise and technology, and effectively automates the very labour-intensive and therefore time-consuming and expensive process of manually cataloguing and extracting data and metadata from many files and scanned documents. After the initial investment in the technology and training model, the time taken to extract an initial metadata-set from a document is reduced from minutes to seconds. Additionally, QC is enriched by enabling the comparison of values across related documents. If further metadata values are required they can extracted without the need to rerun the whole process. Early learning models already deliver positive outcomes for 70-80% of the overall number of documents relating to wells, as demonstrated in CDAL's *2016 Unstructured Data Challenge*. This success rate is expected to improve as the learning models grow in maturity. Automation leads to more automation, and it is quite conceivable that an MLS solution will progress from the current ability to include more legacy data in routine technical workflows to the ability to create new workflows in which the MLS plays an active role in recognizing data patterns typical to improving E&P performance, such as finding missed pay, for example.

References

Hawtin, S. and Lecore, D., 2011, The business value case for data management – a study: A publication by Common Data Access Limited. http://cdal.com/wp-content/uploads/2015/09/Data-Management-Value-Study-Final-Report.pdf