Introduction
Exploring a frontier basin or a more mature block, developing a new static or a dynamic model frequently starts by collecting documents from various internal and external sources and by extracting valuable data.
This is made possible thanks to the support of data management teams which develop methodology and mobilize resources to accurately index a growing number of documents and to extract the information needed by the geoscientists.
Unfortunately, this effort frequently reaches limits due to the fact that:
- The number of documents to be managed by data-managers is growing,
- The received information on the organization data-gates remains unstructured at 80%,
- The appetite for data from geoscientists and data analysts is growing thanks to the development of data analysis platforms and dashboards.
- The recruitment of data managers cannot be aligned with the data growth
As a result, geoscientists establishing their interpretation work on a large dataset frequently recognize that:
- They have to complement the data collection done by the data manager
- They have little information about the quality of the data they receive and have to spend time evaluating information probably already checked by somebody else in the past
- They spend more time searching for data and QCing the data than interpreting
And data manager recognizes that the file systems of their organization contain millions of orphan documents, that is, documents with not enough indexing to have a chance to be found by a query.

Method
Since the principles of the Artificial Intelligence (AI) are now adopted in the exploration domain, a cloud document platform was developed to automate the indexing of the unstructured documents. This platform includes
- A document pre-processor performing an improved Optical Character Recognition for difficult multi-orientation documents
- A training module to generate machine learning (ML) models able to detect text and graphical information
- A detection module to apply the models on large volumes of documents in batch on a computing cluster
- A user interface to check the results and interact with the models
- A restAPI (or any other API compatible with the organization data architecture) to export the detections to the organization’s databases or applications

Figure 1: General workflow to train and to detect text and graphic information from a geoscience document
Using different ML algorithms, it is possible to perform different document indexing tasks:

**Documents classification**
A document can be seen as a bag of letters, words or better couples of words. The repetitions of these elements create features which can define models using algorithms as deep learning, Support Vector Machines, Random Forest, Decision Trees to name the most used. With only 20 documents per class of the taxonomy the system starts to learn and a good classification accuracy (F1 score close to 90%) is obtained with around 100 training documents per class.
Once trained, the classification model can label thousands of documents in parallel and eventually associate several labels per documents in case the document covers topics of several classes. Over the last few months, several classifiers have been established to classify the CDA well related documents, the public Australian geochemical documents, the cased hole documents of a mature North Sea field, the well logs of the Netherlands continental shelf. The classifiers are accurate enough to able to distinguish between a seismic contractor report and QAQC report, between a Rock-Eval evaluation and a Vitrinite measurement or between a wireline resistivity log and a MWD resistivity log.
The classifiers have frequently been compared to black boxes very dependent of their training data. But in fact, the accuracy of the classifiers can be measured using some statistical methods but it is also possible to identified the words or set of words influencing the classifiers using algorithms as LIME or ELI5, hence making the classifiers more trustable.

**Extracting metadata from the text content**
Most of the geoscience documents describe entities such as wellbore, seismic survey, basin, reservoir, etc. Extracting trustable and sourced data from a set of unstructured documents makes it possible to populate or QC subsurface database, hence build a solid background for any further analysis.
To automate this task, we build one model per metadata to be searched in the documents and extracted. Our models combine some classical Natural Language Processing rules which described the values to be searched (e.g. values may be a float or an integer in a range, a date, a coordinate, a money value, a word belonging to a specific dictionary of synonyms, a person name, etc.) and ML text patterns which described the text environment in the neighbourhood of the value to be searched.

As the document classifiers, the metadata extractors start to learn with 25 training examples of targeted metadata values and obtain a F1 score of around 85% (similar that the score of a human indexing documents) with 100 to 200 training examples.

**Detecting graphical patterns in unstructured documents**
In the geoscience documents, the subsurface is described with numerical values, with text and also with graphical symbols like in the case of well composite logs. The improvement of the computer vision algorithms, and particularly the convolutional neural networks (CNNs) designed to search for local pattern in pictures, has made possible to train some graphical metadata detectors to automatically detect and segment a lithology column or to detect oil and gas show symbols, as shown in Figure 4 below:

Figure 4 (on right): Detection of lithological intervals on of composite logs using YOLO as a computer vision tool

Figure 5 (on left): Detection and classification of geological description using an hybrid graphical and text model.

Detecting and segmenting tabular data
The targeted data may be textual or graphical as described above but also printed as table of values in the unstructured documents. The data from the tables are very useful for any further analysis but at the same time very costly to be extracted manually into csv formats. Few data management tools propose a solution to support this task (the opensource Tabula being one of them) and are not able to detect and segment the variety of tables existing in the geoscience documents.

Recently, the speech recognition has made tremendous progress using combination of graph and machine learning to crack (to digitize the oral language) and make possible the real time translation. A document can be compared to a speech where the lines are the phonemes and the paragraph the words. The tables being a specific class of paragraph.

Supported by several French O&G operators and contractors, we are conducting an R&D project to develop models able to segment documents as a graph of paragraph (including tables) and to segment tables as a graphs of cells, column header, nested header, title and footer, etc.

The result of this effort will be published as an open source library to be beneficial for all our industry and beyond.
Lessons learnt
ML has proven to be very useful in increasing the data manager capability to index subsurface documents rapidly and at low cost, hence to provide more data and information for business decision. With five years of experience in this domain, we have been able to define some key points to ensure the success of an automatic indexing project:

- Put the data stewards in the centre of the system. They know about the taxonomy and the semantic of the documents and can interact with the model if a user interface allows to do so in a zero-coding mode.
- Manage the models. Models are new objects in the data management world but they should be managed like other and described with metadata describing their source of training data, performance and versioning. It is a point to make an efficient data model for one task it is another to keep the model improving with the user experience.
- Be ready to combine several machine learning techniques to build more efficient models with NLP, text based ML, computer vision, graphs. This flexibility is necessary to face the huge variety of information requested by the end users
- Models are not the alpha and omega of the data management. They are just tools among other to push the data to the data consumers. Therefore, they have to be integrated in existing workflows using restAPI or other communication protocols.

Conclusion
Geosciences documents are very complex, diverse and very rich in information. First attempt to extract automatically their content at a large scale has already provided very promising results and we have begun the journey to make the unstructured documents a source of knowledge to build databases, bots, analytical and finally reduce the risk of our business decisions.

References
1- Alexey, A.B., Windows and Linux version of Darknet Yolo v3 & v2 Neural Networks for object detection (Tensor Cores are used) https://github.com/AlexeyAB/darknet
2- Blinston, K., H. Blondelle, Machine Learning Systems open up access to large volumes of valuable information lying dormant in unstructured documents in The Leading Edge, March 2017