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

Philip Neri[1]

[1]AgileDD

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 paper 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.