Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents

Kerry Blinston[1] and Henri Blondelle[2]

Abstract

Like many other types of data in the energy industry, well data stored electronically can be divided into two categories: (1) data stored in relational or object databases that are highly structured and (2) data located in documents in various formats (TIFF, JPG, PDF, XLS, etc.) that are typically gathered in folders in a semistructured or unstructured form. Typically, these data break down into 20% structured data versus 80% semi- or unstructured data; this figure is in line with what is observed for other types of data across the industry. This situation affects the ability to make informed decisions since geoscientific software and risk-assessment analytic systems only operate on structured data. Current practices to extract data and metadata from unstructured documents involve a mainly manual and costly process. Data model limitations of the most prevalent databases are a further hindrance to the capture of unstructured data. We discuss a feasibility study to access the 11,500 well headers and 450,000 documents from the United Kingdom Continental Shelf (UKCS) that were released by Common Data Access Limited (CDAL — a wholly owned subsidiary of Oil and Gas UK, funded by 55 operators to share subsurface E&P data) as part of its 2016 Unstructured Data Challenge initiative. A cost-effective solution based on emerging machine learning technology "taught" and guided by data-management experts can support the reliable indexing and cataloging of these forms of data, paving the way for much more reliable E&P business decisions in the future.

Data is critical to E&P operations

Following the sharp fall in oil prices in 2014, it is now even more economically vital for the E&P industry to fully leverage existing information to maximize the recovery of producing fields and improve the performance of exploration drilling activities.

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, vast quantities of data and hundreds of thousands of files and documents have been collected over the past decades. Initiatives taken to optimize the management of existing fields and to evaluate new exploration opportunities can benefit from the knowledge gained during past decades as a result of earlier drilling and formation evaluation activity. In a study titled "The business value case for data management — A study," (Hawtin and Lecore, 2011) commissioned by CDAL and undertaken in collaboration with Schlumberger, one conclusion was that between 25% and 33% of oil company value creation can be reasonably estimated to be attributed to data. 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.

For member companies working in the UKCS, for example, CDAL releases well header and associated documents. Data in LAS, LIS, and DLIS file formats easily transfer into the various structured database systems and application repositories used by operators. However, vast quantities (76% of the CDAL document and file 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. These are typically stored in a tree of hierarchical folders that make it possible to locate them but not directly use their contents. Since the content of these files and documents includes information such as porosity, permeability, pressure tests, and geochemical analysis, geoscientists and E&P companies that do not effectively access this information introduce additional risk and uncertainty into their interpretations and models. To alleviate this problem, appointed data release agents of the UK's Oil and Gas Authority (OGA) have, over the past three decades, systematically manually extracted selected data from these documents and made them available to the industry as nonexclusive data products. Without this work, the ratio of structured to unstructured data is evaluated at 20/80. While it is probably not realistic to extract efficiently every piece of information from unstructured documents, achieving an extraction of 75% of that data would tilt the ratio of usable to unusable data from 20/80 to 80/20.

Until recently, the extraction of important data and metadata from unstructured files was done manually. Productivity was observed to be low, with a typical performance of 120 to 180 documents per day per person. If we consider the 450,000 UKCS documents released by CDAL as part of its 2016 Unstructured Data Challenge initiative, this would translate to 2500 to 3750 person days, or US$3 million to US$4.5 million. This type of work requires significant domain expertise to understand the data and documents and make correct decisions when faced with ambiguous data. The whole process would need to be repeated if, in the future, an additional set of data needed to be extracted from the same set of documents.

It is understandable that companies balk at making this sort of investment in time, resources, and money, with no clear metric of the added value at the time of project inception. Nevertheless, some companies have embarked on such projects. In time, they demonstrate value in terms of both time reductions in searching for and accessing data and in decision-making, where access to critical data, that would otherwise have been unavailable, influenced the decision. However, the time constraints of individual projects often preclude performing data extraction on a set of documents. Automation, operated by data-management experts, seems to be the only cost-effective and time-efficient solution to this problem, but the variability of the documents, the quality of

[1] CGG Data Management Services.
[2] Agile Data Decisions.

http://dx.doi.org/10.1190/tle360300257.1.

scanning, and the organization of material within reports or spreadsheets have been formidable challenges to achieving a computerized solution.

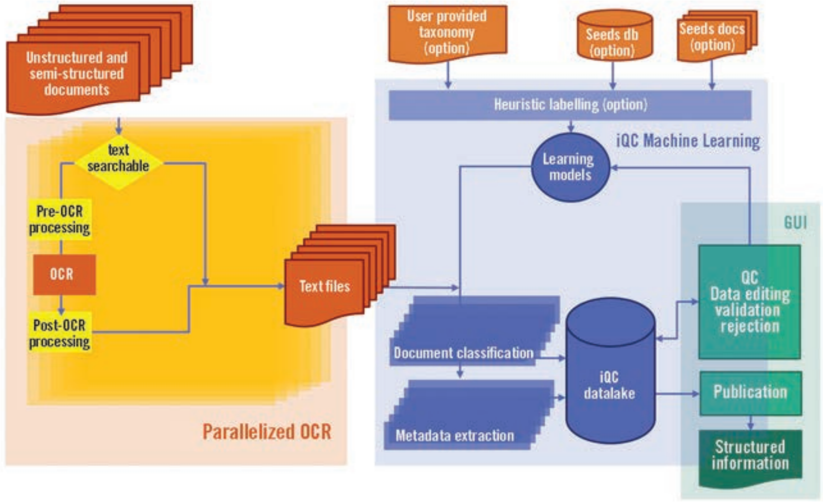


Figure 1. Machine learning system general flowchart.

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 behavior 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 cataloging for E&P files and documents, a machine learning system (MLS) was developed that could be taught and operated by data-management experts to handle the cataloging of data and indexing using metadata (Figure 1). 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 system's reliability and success ratio. Our primary goals were to significantly lower costs and ensure execution within a time span consistent with a project aiming at a review of a field or exploration area.

Two inputs are required to operate the MLS:

- A set of input data. In our example, the documents related to UKCS wells, which are a mixture of semistructured (LAS, LIS, DLIS), unstructured scans (TIFF, JPG, PDF), or unstructured documents (XLS, DOC).
- An initial learning model (ILM), which consists of a set of identified text patterns related to the metadata to be extracted and to the taxonomy used to classify the documents. The results of the automated extraction will improve in line with the number of tagged patterns available to the ILM. The content of the ILM is generated during the training phase of the process.

We will now review some technologies that are essential to the workflow, namely optical character recognition, indexing, cataloging, quality control (QC) and training, and the learning model.

Optical character recognition (OCR)

For scanned documents or nontext-searchable PDF, we first need to convert the image into text wherever text can be identified and properly interpreted into characters. This technology has been around for several decades and is mature. Our MLS integrates a commercial third-party product, considered to be one of the best in this domain. Given that well documents pose particular challenges to OCR systems, due to their complexity we developed both preprocessing and postprocessing steps to assist in generating text files that retained some of the layout information needed by the machine learning process to apply pattern recognition.

Classification based on indexing

The MLS starts by building a document classification using the contents of documents in the data set. The quality of this automated classification was demonstrated in the results obtained during our participation in the 2016 Unstructured Data Challenge organized by CDAL to compare available methods and technologies. Our automated process analyzes the content of documents to build its classification, and in this case, we were able to compare to the classification produced by many North Sea operators over decades of collaboration within CDAL. The CDAL taxonomy that has been used over time to classify the documents consists of 11 classes that segregate data and document types or formats (e.g., scanned logs, scanned reports, digital logs) — in other words, the "containers of the information" — and 67 subclasses that segregate specific data domains (e.g., core analysis, well tests, cementation reports) — in other words, the "content." Since some "containers" are specific to some "content," the combination of the classes and subclasses creates 80 categories on which the MLS has been trained to automatically classify the documents (Figure 2). We ran our MLS on the wells from a few UKCS quadrants (quads 132 and 8) and produced an index that was a match in 85% of all cases. The 15% failure rate was due to underrepresented categories that did not have enough items to be identified. Further training or a modification to the taxonomy would resolve these issues (Figure 3).

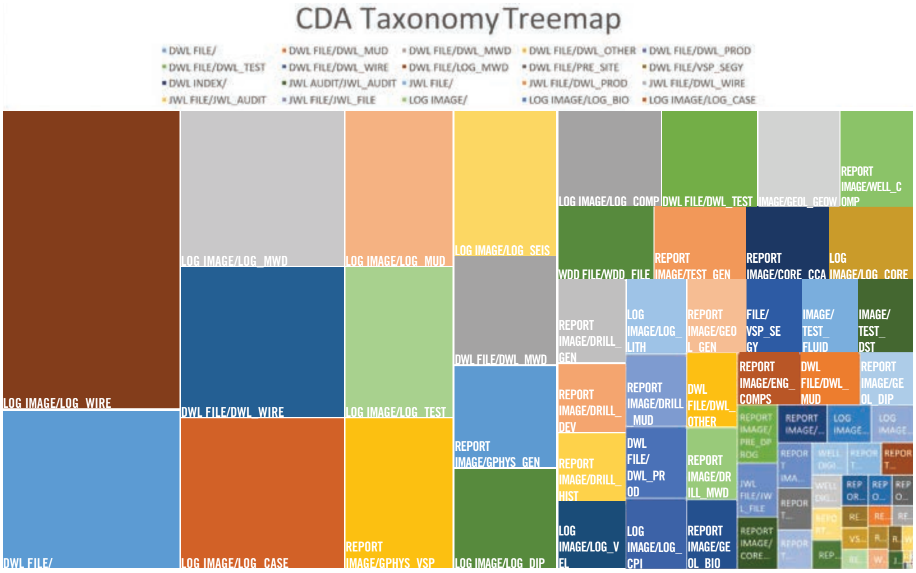


Figure 2. Treemap displaying the CDA taxonomy used to classify the well-related documents.

Generating the catalog

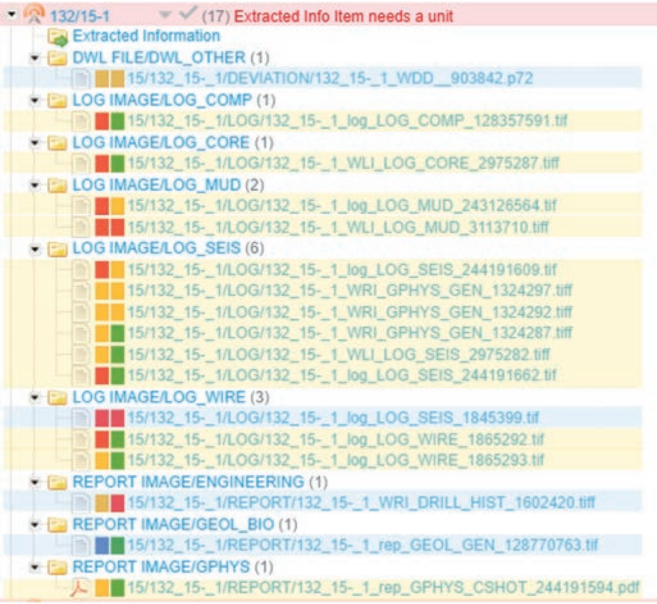
Following the indexing of each document based on content, the MLS attempts to recognize the metadata chosen by the operator as defined in the learning model. The detection process matches as closely as possible a group of characters from the original document with a specific pattern within the learning model. We perform this task using a support vector machine (SVM) (Vapnik, 1999). Several methods for information extraction exist (Su et al., 2015; Gogar et al., 2016; Zhong et al., 2016), but we have found a sparse text representation applied in conjunction with SVMs to be appropriate, especially when only a small number of documents are available for training.

The detection process starts by identifying metadata candidates in the native-format text-searchable file or one that results from OCR processing. Using "total drillers depth" as an example, each positive float value in the text file is considered as a candidate for that metadata value. A set of features is associated with each of the candidates. Examples of features to be considered are character font, page number, the count of similar candidates in the same document, and the surrounding words and their Euclidian distance. The features enable the location of each of the metadata candidates within a hyperspace similar to the one of the learning model. The role of the SVM, at this level, is to distinguish between the "good" and "bad" candidates by defining the best boundary (linear or polynomial) between two areas of the hyperspace where the candidates are located (Figure 4).

Work to weight or define the more discriminant features to better characterize the candidate has been done but must be improved. Our initial results validated the fact that the surrounding text is a good feature to classify the candidates (e.g., coordinate values are frequently close to the words "lat," "latitude," "long," "longitude," *x*, *y*, etc.), but features such as the page number are also very useful at discriminating between candidates (we know intuitively that the first occurrence of a well name in a well report typically occurs on the first few pages).

One benefit of this method is that each identification is awarded a detection confidence factor. This value is defined by the inverse of the Euclidian distance between the model and the candidates. In the event of a candidate being considered as "bad" by the SVM, its confidence factor will be negative. To facilitate a review of the results and to help in the QC process, each metadata item is stored with its confidence factor and the location within the document from which it was extracted.

As we did for the document classification, we compared the automated cataloging done by the MLS with the manual cataloging done over time by the operator on the two UKCS quadrants. Eighty percent of the results matched for the automated detection with a confidence level of above 50. This was equally true for alphabetical metadata, such as the operator name, reference datum, and well status, as it was for numerical values such as coordinates, water depth, or total depth.



```
▾ 132/15-1   ✓ (17) Extracted Info Item needs a unit
    Extracted Information
  ▾ DWL FILE/DWL_OTHER (1)
      15/132_15-_1/DEVIATION/132_15-_1_WDD__903842.p72
  ▾ LOG IMAGE/LOG_COMP (1)
      15/132_15-_1/LOG/132_15-_1_log_LOG_COMP_128357591.tif
  ▾ LOG IMAGE/LOG_CORE (1)
      15/132_15-_1/LOG/132_15-_1_WLI_LOG_CORE_2975287.tiff
  ▾ LOG IMAGE/LOG_MUD (2)
      15/132_15-_1/LOG/132_15-_1_log_LOG_MUD_243126564.tif
      15/132_15-_1/LOG/132_15-_1_WLI_LOG_MUD_3113710.tiff
  ▾ LOG IMAGE/LOG_SEIS (6)
      15/132_15-_1/LOG/132_15-_1_log_LOG_SEIS_244191609.tif
      15/132_15-_1/LOG/132_15-_1_WRI_GPHYS_GEN_1324297.tiff
      15/132_15-_1/LOG/132_15-_1_WRI_GPHYS_GEN_1324292.tiff
      15/132_15-_1/LOG/132_15-_1_WRI_GPHYS_GEN_1324287.tiff
      15/132_15-_1/LOG/132_15-_1_WLI_LOG_SEIS_2975282.tiff
      15/132_15-_1/LOG/132_15-_1_log_LOG_SEIS_244191662.tif
  ▾ LOG IMAGE/LOG_WIRE (3)
      15/132_15-_1/LOG/132_15-_1_log_LOG_SEIS_1845399.tif
      15/132_15-_1/LOG/132_15-_1_log_LOG_WIRE_1865292.tif
      15/132_15-_1/LOG/132_15-_1_log_LOG_WIRE_1865293.tif
  ▾ REPORT IMAGE/ENGINEERING (1)
      15/132_15-_1/REPORT/132_15-_1_WRI_DRILL_HIST_1602420.tiff
  ▾ REPORT IMAGE/GEOL_BIO (1)
      15/132_15-_1/REPORT/132_15-_1_rep_GEOL_GEN_128770763.tif
  ▾ REPORT IMAGE/GPHYS (1)
      15/132_15-_1/REPORT/132_15-_1_rep_GPHYS_CSHOT_244191594.pdf
```

Figure 3. Automatic indexing by a machine learning system.

In the event of a discrepancy between the manual extraction and the MLS extraction, the cause was found to be either an erroneous manual extraction (5% of cases) or an erroneous detection (15% of cases).

We were able to verify that the percentage of incorrect detections decreases as the learning model matures during the QC phase following result validation and refutation. This confirms that the longer the MLS is used in production the better the results become.

In addition to QC of the existing CDAL well-header database using the unstructured data set, we saw the possibility of extracting more metadata than has been done in the past. Metadata, such as the location of the casing shoe, were extracted with success on the same two quads. This is a good illustration of the flexibility of an MLS to adapt to changing industry needs in terms of metadata extraction as the life cycle progresses from exploration to decommissioning.

Training and QC

The confidence factor is a critical element of the user interface geared toward QC. We initially use a simple color code to indicate whether the confidence level for each item is above or below the threshold of "acceptable" confidence. Those flagged below the level then can be investigated, and corrections can be made. The QC interface also allows for the grouping of extracted items by well bore, so that it is possible to compare the value extracted for a particular data object from different documents and files. It is not uncommon for metadata items that, in principle, should be identical across all documents to have differing representations and values.

On checking data and metadata values below the threshold of acceptable confidence in the original files and documents, the data-management expert has a choice of several actions:

- Validation: the reading had low confidence but is in effect correct. The learning model will be updated with additional contextual elements that will result in a higher confidence level should a similar layout be encountered in future extraction runs.
- Refutation: the metadata is wrongly assigned or irrelevant; again, the context around the error is memorized in the learning model, such that similar patterns will be ignored in future extraction runs.
- Deletion: the detected metadata and associated patterns are abandoned and therefore will not be used positively or negatively in the next extraction runs.
- Alternate selection: if the desired metadata is available elsewhere in the document, the user indicates that location, and the learning model memorizes the context such that it will positively identify the correct metadata in future runs.

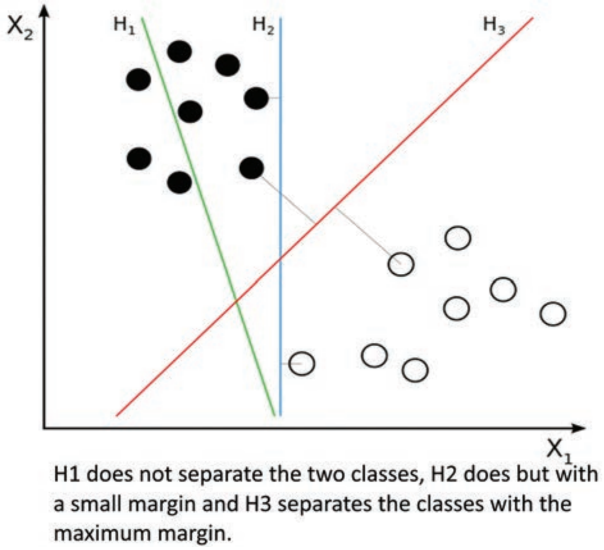

X₂
H₁
H₂
H₃
X₁
H1 does not separate the two classes, H2 does but with a small margin and H3 separates the classes with the maximum margin.

Figure 4. Search for the best boundary (vector) to separate candidates in a feature hyperspace of two dimensions (from https://commons.wikimedia.org/wiki/File:Svm_separating_hyperplanes.png).

Note that the QC step is highly desirable in the early days of applying the learning model to a particular set of data. As confidence levels increase through a number of QC cycles, a point can be reached where the percentage of high-confidence extractions is sufficient to be able to skip the QC process. This is a case-by-case decision as there is a risk involved in skipping the QC steps altogether. However, the reality of project deadlines and limited resources can dictate such a decision.

The evolving learning model

The learning model is the cornerstone of any machine learning system. It embodies the patterns, both positive and negative, that drive metadata identification. The learning model is the sum of all the past experiences made with the MLS and all the QC work performed by data-management experts to correct and improve results.

Our MLS uses an independent learning model for each data item. This explains why, in our system, different metadata items show different confidence levels: this can be attributed to greater complexity, higher variability, or fewer instances from one data item to another. These factors affect the speed with which a learning model matures over a number of runs.

Regarding the initial learning model — i.e., the inception of a new data item not previously part of the MLS runs — we have developed a robust methodology for two different approaches:

- User-guided inception: the operator adds the new data item to the classification and explicitly directs the MLS to the localization of the correct value in all the document and file types where it can be found; this process is quite time-consuming, but it creates an ILM with very detailed positive patterns. However, it will lack any negative patterns.
- A heuristic detection process of metadata with known values derived from, for example, an existing database compiled from documents that have been scanned and manually processed. This heuristic approach feeds the learning model faster and does so by creating not just positive patterns but also negative patterns. Time must be spent performing QC on the results, as there is a high probability that some false-positive patterns or false-negative patterns may have been generated in the process.

Running a machine learning system

Various stages of the typical workflow involve compute-intensive steps, most notably the OCR and the pattern detection calculations, which comprise large matrix inversions. The decision was taken early on in the design of the system to make the various computation steps independent of each other. The design of the MLS was done using the MapReduce architecture, which means that we achieve a linear relationship between computer power and execution speed. This makes the application a very good fit for execution in the cloud, where resources can be modulated according to data volumes. However, this benefit can tail off for the upload of very large collections, which will be very time-consuming. The cloud is therefore the best fit for small- to medium-sized collections. Larger collections are best addressed using local resources, preferably Hadoop or alternatively a high-performance-computing platform, or, in some instances, a single computer.

A hybrid version is in development that will combine the benefits of local execution for access to documents and OCR with execution of compute-intensive MLS tasks on a dedicated machine.

Early results from real data sets

In addition to the work completed for CDAL's 2016 Unstructured Data Challenge, the MLS system is also being used in-house. Basin and reservoir studies benefit from accelerated and reliable document indexing and the extraction of well information. Since the machine learning was initially trained on North Sea data and has been enriched by the results of the CDAL challenge, the best information extraction results are obtained in this region.

The MLS system has been used successfully 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 in early 2017.

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

Alternative solutions

The use of text-pattern detection by an MLS is not the only way to index and catalog large collections of files and documents. An established alternative is to perform full-text indexing after the OCR stage. In this way, the set of documents becomes fully searchable. The end user can search the list of documents containing a particular keyword. The resulting document list can be sorted according to relevance using dictionaries. In some cases, it is very useful, but this type of search cannot be extended to numerical values. Search queries such as "find documents for wells having a total depth greater than 10,000.00 ft" are not possible using this approach.

With an MLS able to detect metadata whatever its nature (keyword or variable numerical value), this type of search is made possible. In other words, machine learning allows the searching of documents without prior knowledge of keywords. For this reason, the MLS approach is superior to existing established full-text indexing.

Conclusions

Our machine learning system, bringing together an essential combination of data-management expertise and technology, significantly automates the very labor-intensive and therefore time-heavy and expensive process of manually cataloging and extracting data and metadata from a large number of 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 be 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. TLE

Corresponding author: kerry.blinston@cgg.com

References

Gogar, T., O. Hubacek, and J. Sedivy, 2016, Deep neural networks for web page information extraction: Artificial Intelligence Applications and Innovations, 154–163, http://dx.doi.org/10.1007/978-3-319-44944-9_14.

Hawtin, S., and D. Lecore, 2011, The business value case for data management – A study: Common Data Access Limited, http://cdal.com/wp-content/uploads/2015/09/Data-Management-Value-Study-Final-Report.pdf, accessed 27 January 2017.

Su, F., C. Rong, Q., Huang, J. Qiu, X. Shao, Z. Yue, and Q. Xie, 2015, Attribute extracting from Wikipedia pages in domain automatically: Information Technology and Intelligent Transportation Systems, 433–440, http://dx.doi.org/10.1007/978-3-319-38771-0_42.

Vapnik, V. N., 1999, An overview of statistical learning theory: IEEE Transactions on Neural Networks, **10**, no. 5, 988–999, http://dx.doi.org/10.1109/72.788640.

Zhong, B., J. Liu, Y. Du, Y. Liaozheng, and J. Pu, 2016, Extracting attributes of named entity from unstructured text with deep belief network: International Journal of Database Theory and Application **9.5**, no. 5, 187–196, http://dx.doi.org/10.14257/ijdta.2016.9.5.19.