An Automated Information Retrieval Platform For Unstructured Well Data Utilizing Smart Machine Learning Algorithms Within A Hybrid Cloud Container

Dr. Kim Gunn Maver, Nina Marie Hernandez, Patrick Jay Lucañas, Juan Carlos Graciosa, Charmyne Mamador, Luis Caezar Ian Panganiban, Carl Yu, Marcus Gunn Maver

iraya
machine learning • geoscience

BIG DATA ANALYTICS

has long existed in the petroleum/energy industry.

EXPLORATION

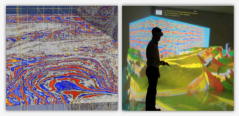
requires thorough data analysis and interpretation of large volumes of seismic, well and spatial data in various formats and vintages. Upstream projects related to new basin entries, exploration and appraisal programmes and wildcat well planning rely on good quality and large volume datasets – where available – to mitigate risks and make the best-informed decisions.

Rock physicists , Geophysicists ,Geologists, Production & Reservoir Engineers

go through **TERABYTES** of data on a daily basis

THE PROBLEM

about data mining is that it is TIME-CONSUMING and EXPENSIVE.



HIGHLY ADVANCED PLATFORMS for integration of well and seismic data exist.

UNSTRUCTURED DOCUMENTS in the form of well reports, seismic reports, exploration and production analytics, are one form of geoscience data that has been neglected by the industry.

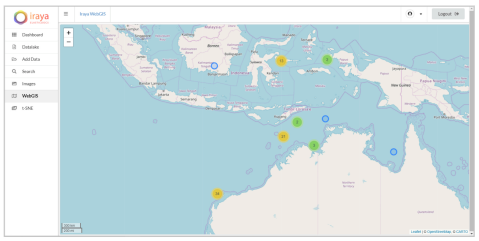
DECADES-WORTH OF GEOLOGICAL INSIGHTS from these documents become lost. Companies end up either redoing the exploration work, or reacquiring additional data which are both very costly.

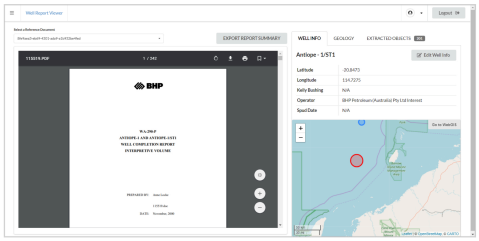
OUR SOLUTION

- A good data mining design as a strategy in exploration efficiency.
- TensorFlow **ElasticDocs** as fully tested pipeline of ML-enabled Data Mining Technique.
- Pre-trained Global Models on Geoscience Data.
- Web-based information for global access.
- Extracted geolocation provides spatial context to data.

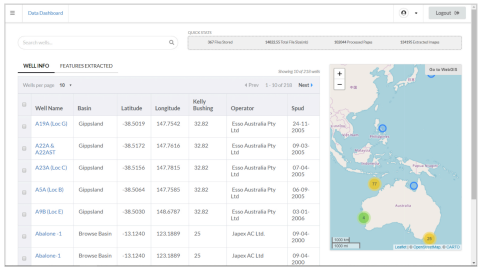
4 PILLARS OF ELASTICDOCS

Information Geolocation & Density

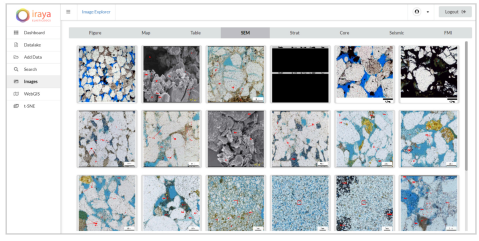




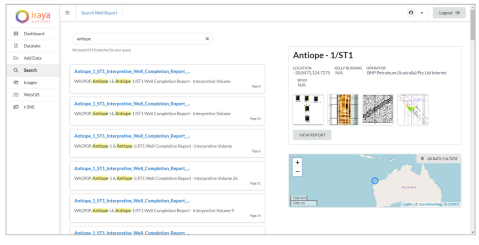
Metadata Extraction



Autoimage Classification & Classification



Global Elastic Search in Corpus



METHODOLOGY

ElasticDocs uses a sequence of processes which mimics the human experience of processing unstructured documents (Figure 2). It utilizes a hybrid data service architecture that uses both cloud and private servers (Figure 1).

ELASTICDOCS PLATFORM ARCHITECTURE

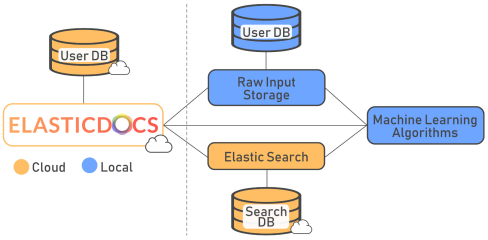


FIGURE 1. The architecture of the platform and its process indicating which types of servers are used.

The first step in the process is the digitization of data. This is where unstructured dataset inputs like .pdf or .docx file formats are converted into an editable format.

For the extracted images, VGG-16 is a very deep convolutional neural network architecture with 16 weight layers using around 138 million parameters (Simonyan et al., 2014). This was used to develop a model to automatically classify between tables, charts, stratigraphic charts, maps, seismic, core samples and SEMs within each document. Figure 3 shows the accuracy of the model produced by the neural network against 5000 test images. Recall shows the accuracy of the model at detecting relevant elements while precision shows the exactness of model's positive classification. F1-score is the harmonic mean of recall and precision.

F1-SCORES OF THE IMAGE CLASSIFICATION

	PRECISION	RECALL	F1-SCORE
MAP	0.83	0.96	0.89
SEISMIC	1.00	0.95	0.97
CORE	0.89	0.98	0.94
SEM	0.95	0.93	0.94
OTHERS	0.91	0.73	0.81

FIGURE 3. F1-scores of the image classification

Figure 4 shows the distinction of seismic images against the other images within the t-SNE visualization. The visualization helps the user to automatically identify relevant information such as seismic images, SEM, maps and core sample images.

ELASTICDOCS MACHINE LEARNING SEQUENCE

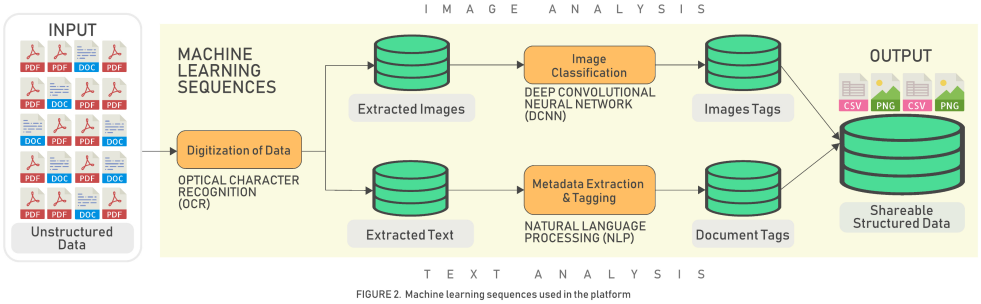


FIGURE 2. Machine learning sequences used in the platform

Users can upload data into the platform which are stored into the object storage. ElasticDocs platform triggers the machine learning algorithm workflows, pulls data from the storage, and implements the necessary processing. Generated outputs are forwarded through to Elasticsearch which is then exposed to the user. ElasticDocs processes unstructured datasets into a condensed format in which only specific or important information are stored.

The conversion uses Optical Character Recognition (OCR) where the machine identifies each character in the image (Smith., 2007) and output results depending on the user's preference. Hough Line transform (Saha et al., 2010) and Non-Local Means Denoising algorithms (Buades et al., 2011) is used to improve the detection of text and non-text areas of the documents.

After digitization, the next is the detection of the important information within the digitized texts. This process is the metadata extraction and tagging. It utilizes Natural Language Processing (NLP) to tokenize each digitized text and identify terms of significant value.

Named Entity Recognition (NER) is then performed to create a model (Ratinov et al., 2009) that extracts the metadata like wellname, basin, permit, operator, well classification, latitude, longitude, spud date and kelly bushing on each summary page.

t-SNE 3D VISUALIZATION

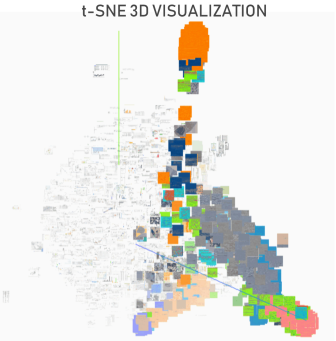


FIGURE 4. t-SNE 3D visualization of images extracted from the unstructured datasets

It also identifies which images are more identical or similar to each other which is useful when comparing seismic images.

CONCLUSION

ElasticDocs is an effective platform that automatically reads and understands hundreds or thousand of technical documents with little or no human supervision by leveraging machine learning techniques.

The aim of the platform is to ease the identification of the important parts of each of the documents from an unstructured dataset.

It automatically extract significant information and display the extracted dataset to the user to aid them in decision making.

- Buades, A., Coll, B., & Morel. J.M., 2011. Non-Local Means Denoising. Image Processing on Line, 1, pp. 208-212.
- Ratinov, L., & Roth, D. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147-155.
- Simonyan, K. & Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. arXiv preprint arXiv:1409.1556
- Saha, S., Basu, S., Nasipuri, M. & Kr. Basu, D. 2010. A Hough Transform based Technique for Text Segmentation. Journal of Computing. 2. pp. 134-141.
- Smith,R. 2007. An Overview of the Tesseract OCR Engine, Proc. International Conference on Document Analysis and Recognition.