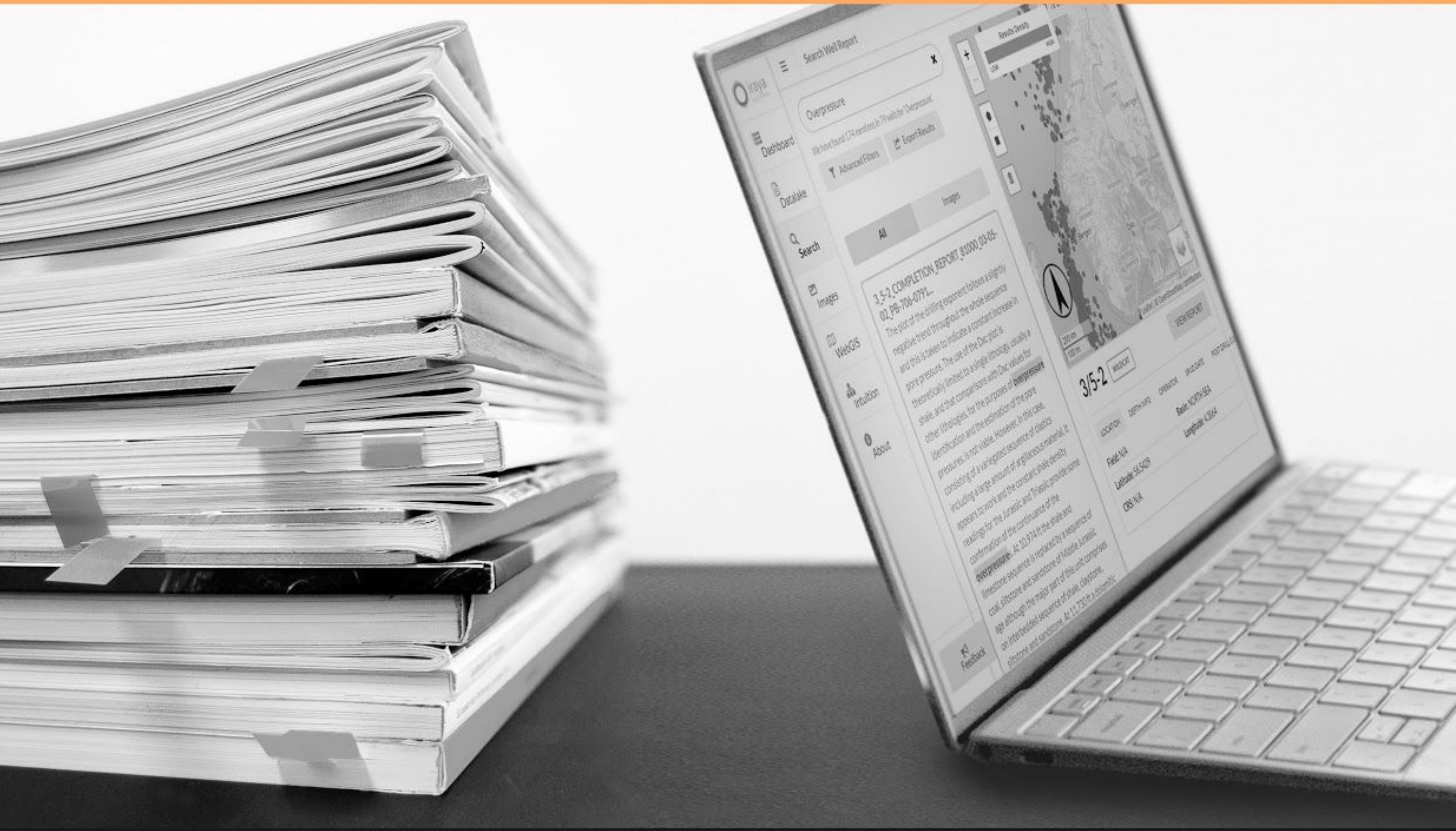
iraya | WHITE PAPER

Bringing Unstructured Engineering and Geoscience Data into a Fully Digitized World for Information Extraction (IE)

AUGUST 2020



We have found 174 mentions in 79 wells for 'Overpressure'.

3_5-2_COMPLETION_REPORT_81000_03-05-02_PB-706-0791...
The plot of the drilling exponent follows a slightly negative trend throughout the whole sequence and this is taken to indicate a constant increase in pore pressure. The use of the Dxc plot is theoretically limited to a single lithology, usually a shale, and that comparisons with Dxc values for other lithologies, for the purposes of overpressure identification and the estimation of the pore pressures, is not viable. However, in this case, consisting of a variegated sequence of clastics including a large amount of argillaceous material, it appears to work and the constant shale density readings for the Jurassic and Triassic provide some confirmation of the continuance of the overpressures. At 10,974 ft the shale and limestone sequence is replaced by a sequence of coal, siltstone and sandstone of Middle Jurassic age although the major part of this unit comprises an interbedded sequence of shale, claystone, siltstone and sandstone. At 11,730 ft a dolomitic

Dr Kim Gunn Maver & Francois Baillard

www.irayaenergies.com

Summary

It is difficult times for the oil and gas industry, which is why companies are implementing digital technologies to drive efficiency throughout the organization. But success has been limited and one key reason, which is often overlooked, is the inability to fully leverage data.

Oil and gas companies are awash with data from many different disciplines, the amount of data is growing exponentially and 80% is estimated to be unstructured (reports, presentations, spreadsheets etc). With a substantial part of existing and new data being unstructured, it means that the majority of the data generated and stored are to a large extent unusable, requiring engineers and geoscientists to spend significant time to find the information they need to make decisions.

An efficient and accurate transformation of unstructured data using Artificial Intelligence and Machine Learning will lead to improved performance of data analysis and Information Extraction.

Digitized unstructured data are ingested through a pipeline with workflows using Machine Learning techniques such as Natural Language Processing for text data and Deep Convolutional Neural Network for non-text data to provide an organized dataset.

The result of the unstructured data processing is made available anywhere through a data lake, which greatly accelerates the time to value using storage capabilities from distributed commoditized hardware and open-source software.

Organizing the unstructured data makes it possible to continuously organize the new unstructured data being produced. By breaking down the different data silos, freeing data and making data accessible through a single web-enabled interface, any data can instantly be identified, located and retrieved through a text and image search enabling new cross-functional use cases and higher order analysis can be performed.

A digital transformation of the oil and gas industry has been predicted to be able to provide 1 trillion USD value to oil and gas companies, reducing production costs by 10% and increasing recovery rates by 10%. Managing unstructured data has the potential to enhance both the productivity and competitiveness of a company by enabling exploration and production teams to gain insights through identification of patterns, relationships, and anomalies that can help them make better decisions faster, improving success and recovery rates, as well as bottom lines both short and long-term.

Overview

Introduction

It is difficult times for the oil and gas industry with increasing price volatility, alternative forms of energy becoming increasingly popular, and the price of a barrel of oil remaining low. In the face of these challenges, oil and gas companies are implementing digital technologies to drive efficiency throughout the value chain. But success has been limited. One key reason that is often overlooked: An inability to fully leverage data[1].

Data explosion

Oil and gas companies are awash with data from many different disciplines. With the digitization of the oil and gas industry and Internet of Things (IoT), where digital sensors are placed throughout the exploration and production chain, the amount of data is growing exponentially and is estimated to double every 12 to 18 months by Bill Braun (CIO of Chevron)[2]. As the engagement in the geoscience and engineering domain also continues to grow within governmental agencies and scientific organizations and the number of related computational models expands, an overwhelming amount of geoscience text and image data are also being generated in a variety of digital forms and in numerous languages[3][4][5][6].

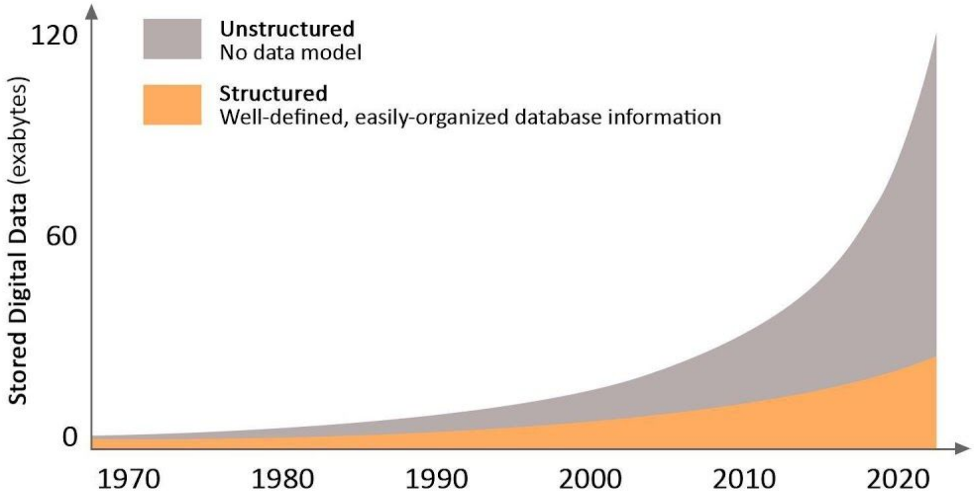


Figure 1: The distribution of structured and unstructured data.

[1] Santamarta, S., Forbes, P., Gandhi, R. and Bechauf, M., 2019: Big oil, big data, big value. Boston Consulting Group. https://www.bcg.com/publications/2019/big-oil-data-value.aspx.

[2] Crooks, E., 2018: Drillers turn to big data in the hunt for more, cheaper oil. https://www.ft.com/content/19234982-0cbb-11e8-8eb7-42f857ea9f09

[3] Lima, L. A., Görnitz, N., Varella, L. E., Vellasco, M., Müller, K.-R., and Nakajima, S., 2017): Porosity estimation by semi-supervised learning with sparsely available labeled samples. Computers & Geosciences, 106, pp 33–48.

[4] Wu, L., Xue, L., Li, C., Lv, X., Chen, Z., Jiang, B., and Xie, Z., 2017): A knowledge-driven geospatially enabled framework for geological big data. ISPRS International Journal of Geo-Information, 6(6), 166.

[5] Xiao, F., Chen, Z., Chen, J., and Zhou, Y., 2016: A batch sliding window method for local singularity mapping and its application for geochemical anomaly identification. Computers & Geosciences, 90, pp 189–201.

[6] Zheng, J., Fu, L., Ma, X., and Fox, P., 2015: SEM+: Tool for discovering concept mapping in Earth science related domain. Earth Science Informatics, 8(1), pp 95–102.

Business research and advisory company Gartner predicted in 2019 that the data volume will grow 800% over the next 5 years and, of the new data generated, 80% of it will be unstructured data[7] (reports, presentations, spreadsheets etc), which corresponds to a prediction by IDC in 2011 (Figure 1)[8]. A variety of sources have made similar predictions for unstructured data representing from 70% to 90% of all data [9] [10] [11]. As an example, 76% of the United Kingdom Continental Shelf (UKCS) data released through Common Data Access Limited (CDAL) was unstructured[12].

The oil and gas industry is among the most data intensive industries in the world, and while over the last decade the industry has become very effective and efficient in managing, storing, and sharing structured data, the industry has failed to find cost-effective ways to manage and use unstructured data. Critically, it is this unstructured data that contains much of the information needed to support key investment decisions[13] [14].

With a substantial part of existing and new data being unstructured, the majority of the data generated and stored are to a large extent unusable. Therefore, most users either don't know what data they have or simply cannot find it. And the oil and gas industry is no different[15]. While this is understandable as managing unstructured data takes time, money, effort and expertise, it results in companies wasting money by making ill-informed investment decisions[16].

Potential opportunities and solutions for Big Data are obstructed by the huge volume and variety of data. Gartner reports that the growing unstructured-data problem has less to do with storage than with accessibility[17]. According to GE research less than 1% of the collected data are used by oil and gas companies in decision making[18]. Within exploration activities it is estimated that barely 5% of the seismic data that has been gathered at great expense are put to use in the decision-making processes[19].

[7] Walker, A., 2019: Oil and gas has a problem with unstructured data. Data Science and Digital Engineering in Upstream Oil and Gas. https://pubs.spe.org/en/dsde/dsde-article-detail-page/?art=5676.

[8] IDC, 2011: Digital Universe study. https://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

[9] Gandomi, A., Haider, M., 2015: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 35 (2): 137–144. June.

[10] Schneider, C., 2016: The biggest data challenges that you might not even know you have. IBM. https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

[11] Agile DD, 2016: Structured information delivers reliable decisions. https://www.agiledd.ai/case-studies.

[12] Blinston, K. and Blondelle, H., 2017: Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents. The Leading Edge, March, p 64-68.

[13] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[14] Tankimovich, M. R., 2018: Big Data in the oil and gas industry: A promising courtship. Thesis, University of Texas at Austin.

[15] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[16] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[17] Walker, A., 2019: Oil and gas has a problem with unstructured data. Data Science and Digital Engineering in Upstream Oil and Gas. https://pubs.spe.org/en/dsde/dsde-article-detail-page/?art=5676.

[18] RDS Consulting, 2020: Changing the upstream oil and gas through AI and analytics. https://blogs.opentext.com/ai-analytics-power-change-upstream-oil-gas/.

[19] Economist, 2017: Data drilling. Oil struggles to enter the digital age. Ulrich Spiesshofer, CEO ABB, April.

Geoscientists and engineers searching for information

These vast quantities of data are stored in unstructured formats, meaning that it cannot be easily searched and utilized by geoscientists and engineers. Two phrases that are echoed across the industry are, “I don’t know what I don’t know” and “if we only knew what we know.”[20].

It has been reported that engineers and geoscientists spend over half of their time in searching and assembling data[21]. Astrin Hanne Larsen (CIO at Equinor) has stated that 80% of employee time in the industry is spent looking through unstructured data to make informed decisions[22]. For mineral exploration, similar numbers are reported by Robin Lee Fell (Director, Strategic Technology Solutions at Goldcorp Inc) stating that 80% of the time is spent by exploration geologists searching for and manipulating data while only 20% is available for analysing data[23]. IDC puts the number for searching data at 30% across industries overall indicating the oil and gas industry is far behind[24].

The value in unstructured data

Data is normally not part of an oil and gas companies balance sheet, but is regarded as a cost, which is understandable as a substantial part of the data is unstructured and nearly impossible to locate and use in decision making.

Information extraction (IE) processes can be used to extract knowledge in the form of entities, relations, facts, terms, and other types of information, which helps the data processing pipeline to prepare the data for analysis. If unstructured data was of minimal importance, it wouldn’t really matter how much of it there was. But there is substantial value in unstructured data as will be described in more detail later in this article.

Geoscience technical articles are a good example of unstructured data, which have been successfully released because structured access has been provided. It has been widely recognized that the value of open and persistent data grows as they become discoverable, citable, re-usable, integrated, and linked with other data[25]. Interest in Big Data in geoscience is therefore increasing as can be seen in figure 2[26] as a variety of detailed data about geological topics and geoscience knowledge are buried in the geoscience literature[27].

[20] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[21] Brulé, M. R., 2015: The Data Reservoir: How Big Data Technologies Advance Data Management and Analytics in E&P. SPE Digital Energy Conference and Exhibition.

[22] Walker, A., 2019: Oil and gas has a problem with unstructured data. Data Science and Digital Engineering in Upstream Oil and Gas. https://pubs.spe.org/en/dsde/dsde-article-detail-page/?art=5676.

[23] Northern Miner, 2018: Using Big Data and AI for smarter mineral exploration. Round Table. http://www.northernminer.com/wp-content/uploads/2018/03/1-8_IBM-Mar22FinalDE.pdf

[24] Walker, A., 2019: Oil and gas has a problem with unstructured data. Data Science and Digital Engineering in Upstream Oil and Gas. https://pubs.spe.org/en/dsde/dsde-article-detail-page/?art=5676.

[25] Lehnert, K., & Hsu, L., 2015: The new paradigm of data publication. Elements, 11(5), pp 368–369.

[26] Chen, S. and Chen B., 2019: Practices, Challenges and Prospects of Big Data Curation: A Case Study in Geoscience. International Journal of Digital Curation 2020, Vol. 14, (1), pp 275–291.

[27] Qiu, Q., Xie, Z., Wu, L. and Tao, L, 2019: GNER: A generative model for geological named entity recognition without labeled data using deep learning. Earth and Space Science, 6, 931–946. https://doi.org/10.1029/2019EA000610

The efficient and accurate transformation of unstructured data using Artificial Intelligence (AI) and Machine Learning (ML) will lead to improved performance of data analysis and IE[28], which will in turn improve decision making, be cost efficient and increase oil and gas production.

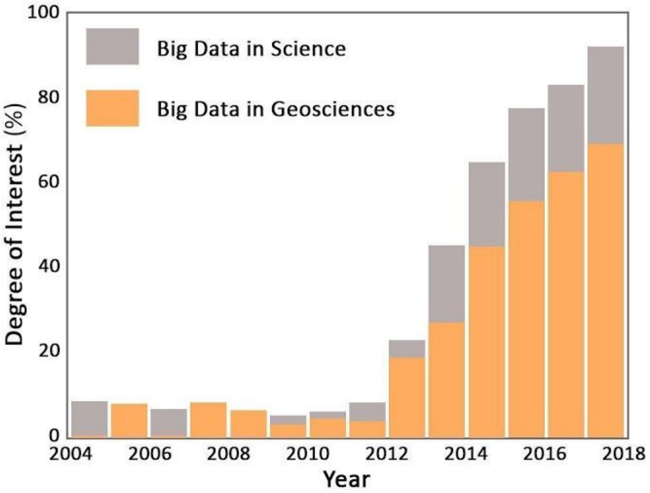


Figure 2: Degree of interest in Big Data in science and geoscience (based on Google Trends and acquired in January 2019).

The following sections will describe unstructured data within the oil and gas industry, review how the unstructured data can be processed and organised and then demonstrate how value and new knowledge can be efficiently extracted (Figure 3).

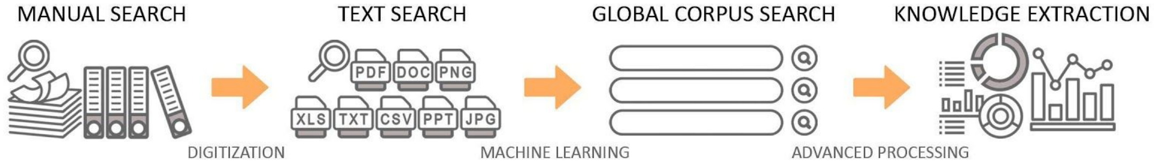


Figure 3: Unstructured data ingestion, digestion and IE.

What are unstructured data?

Data definition

Data are divided into 3 groups as shown in Table 1; Unstructured data, semi-structured data and structured data.

As semi-structured and especially structured data are already well organised, they are already suitable for advanced data analytics using technologies such as AI and ML. Due to their direct applicability, these data can be associated to shallow AI, where very specific repetitive tasks will be performed faster and often with better accuracy than possible by a specialist. Examples of such applications in the oil and gas industry are seismic interpretation or log facies clustering.

[28] Adnan; K. and Akbar, R. 2019: Limitations of information extraction methods and techniques for heterogeneous unstructured big data. International Journal of Engineering Business Management, Volume 11: pp 1–23

Unstructured data is not immediately ready for data analytics and therefore requires a wider breadth of tools and technologies to be fully utilized. This will be typically associated to a deeper level of AI.

Unstructured can be divided into 4 main classes:

- Text
- Images
- Audio
- Video

TABLE 1: DATA TYPES

TYPE	DESCRIPTION	TECHNOLOGY REQUIREMENTS
Structured data	Structured data usually resides in relational databases (RDBMS). Elaborate metadata is often associated to the data itself. Text strings of variable length like “names” are contained in records, making it a simple matter to search. Data may be human- or machine-generated as long as the data is stored within an RDBMS structure. This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date.	Shallow AI
Semi-structured data	Semi-structured data is a form of structured data that does not obey the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data, also known as self-describing structure. In semi-structured data, the entities belonging to the same class may have different attributes even though they are grouped together, and the attributes’ order is not important.	Shallow AI
Unstructured data	Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured data is often text-heavy and also contains data such as dates, numbers and images. Limited or no metadata is associated to the data. This results in irregularities and ambiguities, which makes it difficult to process using traditional programs as compared to data stored in fielded form in databases or annotated in documents.	Deep AI

The unstructured data challenge

An Unstructured Data Challenge Survey for a world-wide range of oil and gas companies reveals significant and common internal challenges in managing unstructured data[29]:

[29] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

- High variation in information types and sources (e.g., internally vs. externally generated)
- Inconsistency in data quality
- Variation in nomenclature and units to reconcile
- Number of disparate data containers in disparate locations
- Significant variation of file formats to reconcile
- Diverse data models
- Non-homogeneous ownership with diverse collection strategies
- Uncoordinated access rights and access models
- Unconfirmed, inaccurate, expired, or retracted information
- Duplicate files and lack of version control

Furthermore free text in the geoscience literature domain is often recorded in either structured or unstructured forms (e.g., technical reports, geological reports, books, and other types of reports), therefore also posing challenges for engineers and geoscientists who need to effectively manage, share, analyse, and reuse all these online data[30] [31] [32].

By processing the unstructured data, the goal is to create organised data, which is easily accessible and applicable for IE.

How to process unstructured data?

Initial processing of unstructured data has been focused on digitization, data storage and simple search functions to utilize this "dark" data. By utilizing advances in supervised/unsupervised ML, active learning and cloud computing it is possible to further organize these data and undertake IE (Figure 4).

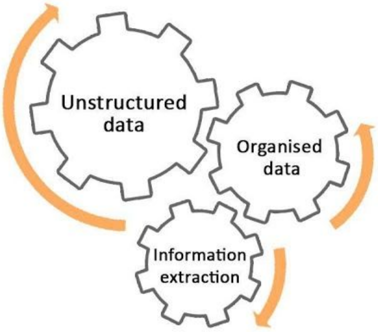


Figure 4: Processing flow.

Digitized unstructured data are ingested through a pipeline with workflows using ML techniques such as Natural Language Processing (NLP) for text data and Deep Convolutional Neural Network (DCNN) for non-text data to provide an organised dataset.

[30] Cernuzzi, L., & Pane, J., 2014: Toward open government in Paraguay, IT Professional, Vol. 16, pp 62–64.

[31] Ma, X., 2017: Linked Geoscience Data in practice: Where W3C standards meet domain knowledge, data visualization and OGC standards. Earth Science India, 10(4), pp 429–441.

[32] Wang, C., Ma, X., Chen, J., and Chen, J., 2018: Information extraction and knowledge graph construction from geoscience literature. Computers & Geosciences, 112, pp 112–120.

A workflow for automatically extracting information from the documents consists of a set of algorithms to identify blocks/segments within a document, after which, supervised machine learning, is used to classify the document segments as either text or non-text.

Optical Character Recognition (OCR) can be applied to the text segments to convert them into editable text. In a separate data pipeline, the non-text components such as images and tables can be tagged and using convolutional neural networks (CNN) the machine learns to auto classify different image types[33].

The result of the unstructured data processing is then made available through a data lake. This greatly accelerates the time to value because data lakes derive their storage capabilities from distributed commoditized hardware and open-source software, rather than from legacy systems, and they can scale to enormous capacity. Also, structured and unstructured information can be stored in infinite combinations[34].

It is important to note that when unstructured data has been processed to create an organised dataset, this is not directly comparable to a native structured dataset. Unstructured data consist of a very diverse range of information and relationships (soft data), whereas structured data are less diverse, but each dataset is voluminous and digital by birth (hard data). The organised soft dataset includes associated new metadata defining new knowledge and insights, which will become more easily accessible and through IE, made directly available to the end user.

Developing a training dataset

The quality and quantity of the training data has as much to do with the success of a data project as the algorithms themselves and is therefore the cornerstone of the whole process. Even if a vast amount of well-structured data is stored, it might not be labelled or only partially labelled in a way that the data can be used as a training dataset for prediction. For most ML initiatives, a big part of the effort of the data scientists is preconditioning the data and making sure each data point is annotated accordingly. Given the volumes of data to process, an active learning methodology is conducted allowing rapid model iterations and fast prototyping. The data scientist can therefore not only focus on labelling random data but on labelling a portion of the data, which will significantly improve the accuracy of the ML model.

The final output of such an exercise is an annotated training dataset, which is statistical representative of the unlabelled data we want to predict or annotate in the future production stage. In addition, and equally important, this process provides additional insights and a deeper understanding of what the data represents. This is particularly useful not only for the data scientist but also for the geoscientist and engineer to allow identification of general trends, different statistical distributions and localized anomalies embedded within the dataset making it possible to find a “needle in a haystack”. The process and path towards the solution may in many cases be as useful as the solution itself.

As a result, the ML systems have the advantage of retaining knowledge within the system itself and thereby allowing for more efficient processes as the system is scaling up. The long-term result is

[33] Maver, K. G., Hernandez, N., M., Baillard, F. and Cooper, R., 2020: Processing of unstructured geoscience and engineering information for instant access and extraction of new knowledge. First Break, vol 28, June, pp 59-64.

[34] Santamarta, S., Forbes, P., Gandhi, R. and Bechauf, M., 2019: Big oil, big data, big value. Boston Consulting Group. https://www.bcg.com/publications/2019/big-oil-data-value.aspx

knowledge aggregation via a learning system, which results in new insights and improved processing speed way beyond what a single specialist can do.

Cross-disciplinary approach

Going digital is not easy, new capabilities are required. It is not just a question of shuffling people around and thinking every engineer can now be a data scientist[35]. In general organizations needs cross-function capabilities of subject-matter experts as well as experts having a strong domain knowledge for establishing an efficient delivery engine[36]. As machines are performing repetitive and replicable tasks better and better, domain experts are not limited anymore to the platform or software needed to perform this task. Domain experts are seen more and more as gatekeepers to ensure the machine performs as expected and that new “human” insights or contexts are constantly fed to the machine to ensure knowledge is captured in a collaborative and scalable solution.

What is the value of extracted information from unstructured data?

The new tech trend in a 2018 Deloitte review states that the aim of data ingestion is to “free” the unstructured data; to make information accessible, understandable, and actionable across business units, departments and geographies, which is increasingly a requirement as data expands in both volume and complexity[37]. By freeing data from silos, a flexible data platform also enables new cross-functional use cases[38].

Data are only important if it influences decisions that have business impact within an organisation. According to the 2016 Accenture and Microsoft oil and gas digital trends survey, two-thirds of oil and gas professionals reported that the use of analytics is one of the most important capabilities for transforming their companies. In the same survey, 56 percent of respondents indicated that big data and digital capabilities will enable them to make faster, better decisions[39].

Within a Data-to-Impact process, there are three limiting factors for adequate data decisions to be taken: (1) Internal complexity within an organization, (2) External constraints such as time or money, and (3) Individual limitations related to human cognitive skills[40]. The outcome of organising the unstructured data is managing these three limiting factors and results in the following value chain (Figure 5):

[35] Bonny, T., 2019: Is data the new currency? Unconventional operators go digital to help improve well productivity & operating efficiencies. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/process-and-operations/us-is-data-the-new-currency.pdf

[36] Shetty, P., 2018: A defined methodology better manages subsurface data. Oil and Gas Engineering. https://www.oilandgaseng.com/articles/a-defined-methodology-better-manages-subsurface-data/

[37] Briggs, B., 2018: Tech Trends 2018. The symphonic enterprise. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/be/Documents/technology/TechTrends-2018.pdf

[38] Santamarta, S., Forbes, P., Gandhi, R. and Bechauf, M., 2019: Big oil, big data, big value. Boston Consulting Group. https://www.bcg.com/publications/2019/big-oil-data-value.aspx

[39] Accenture, 2016: The 2016 Upstream Oil and Gas Digital Trends Survey. https://www.accenture.com/us-en/insight-2016-upstream-oil-gas-digital-trends-survey

[40] Baillard, F., 2020: Do you always need new data? Medium. https://medium.com/@fb_88498/do-you-always-need-new-data-96d4c0364a

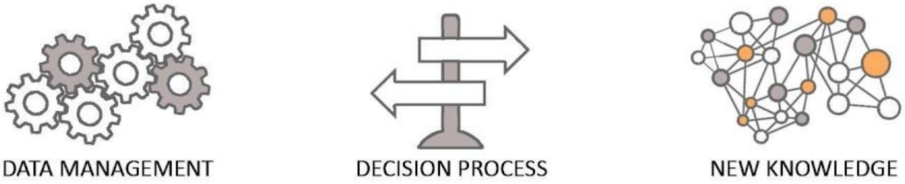


Figure 5: The value of processing unstructured data.

Data management

Data silos are a particular challenge in the oil and gas industry, where local storage of data and long, complex value chains mean individual businesses or even units within a business may lack a holistic view of the data they need to improve their operations[41].

Organizing the unstructured data in a data lake makes it possible to continuously organize the new unstructured data being produced, and as a result all data can be utilized in decision making. With consolidated unstructured data instantly accessible across the organization, cross-company collaboration is significantly strengthened.

An organization may decide to buy, acquire or record new data to improve the decision-making process. However, it may well be that the information required is already available within the organization, albeit in an unstructured form. Thus, a process of extracting knowledge from existing data may negate the considerable delay and cost associated with acquiring new data.[42].

Decision making

By ingesting both historic and future unstructured data, breaking down the different data silos, freeing the data and making the data accessible through a single web-enabled interface, any data can instantly be identified, located and retrieved through a text and image search enabling new cross-functional use cases[43].

Tools have been developed to ensure it is possible to retrieve relevant information when required and improve decision making:

- Intelligent full text search capabilities of the text corpus and images with a link to the original data point (report etc.)
- Structuring of all images for better overview and comparison
- Tables converted to .csv files for easy data retrieval
- Non-English translation functionality are possible for easier information digestion and full search capabilities in English
- Ask questions using a Chatbot and receive relevant answers

[41] Williams, J. and Strier, K, 2020: Is AI the fuel oil and gas needs? EY. https://www.ey.com/en_gl/oil-gas/is-ai-the-fuel-oil-and-gas-needs
[42] Baillard, F., 2020: Do you always need new data? Medium. https://medium.com/@fb_88498/do-you-always-need-new-data-96d4c0364a
[43] Santamarta, S., Forbes, P., Gandhi, R. and Bechauf, M., 2019: Big oil, big data, big value. Boston Consulting Group. https://www.bcg.com/publications/2019/big-oil-data-value.aspx

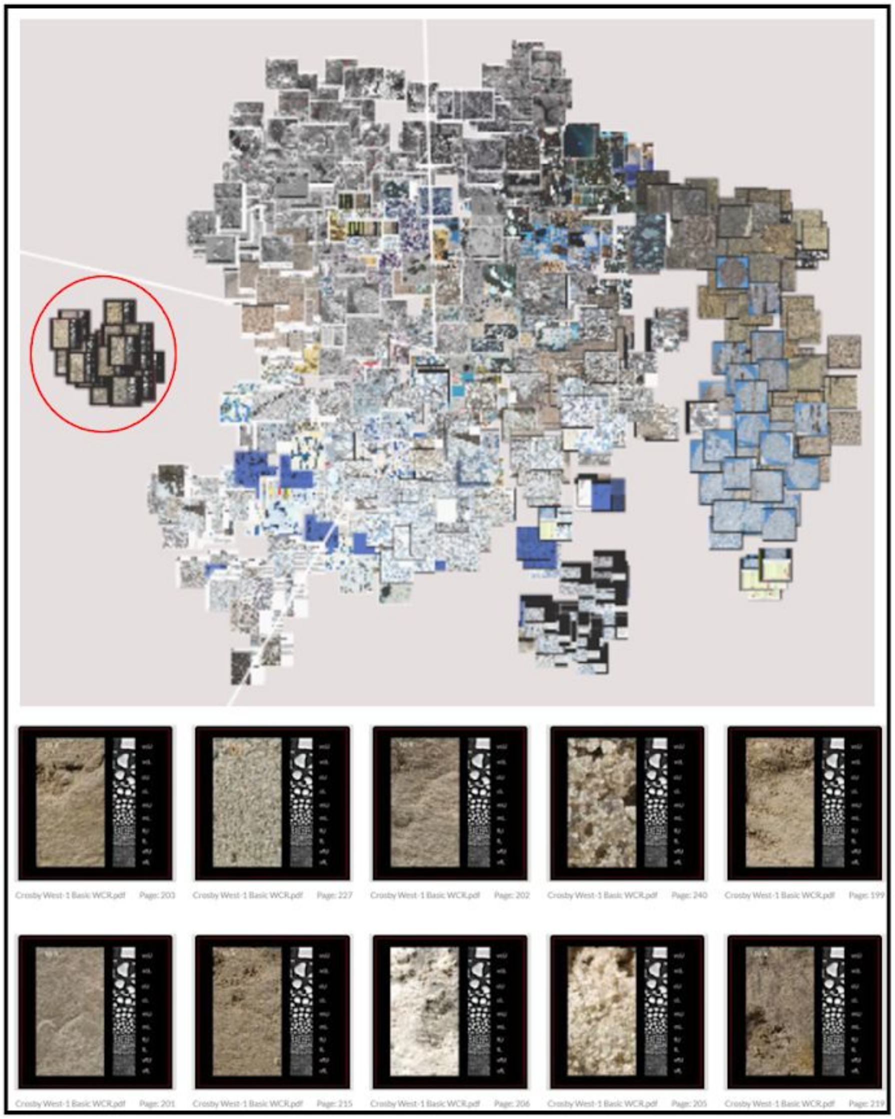

Crosby West-1 Basic WCR.pdf Page: 203
Crosby West-1 Basic WCR.pdf Page: 227
Crosby West-1 Basic WCR.pdf Page: 202
Crosby West-1 Basic WCR.pdf Page: 240
Crosby West-1 Basic WCR.pdf Page: 199
Crosby West-1 Basic WCR.pdf Page: 201
Crosby West-1 Basic WCR.pdf Page: 215
Crosby West-1 Basic WCR.pdf Page: 206
Crosby West-1 Basic WCR.pdf Page: 205
Crosby West-1 Basic WCR.pdf Page: 219

Figure 6: A representative cluster of thin section images.

Classified images can be organized using t-Distribute Stochastic Neighbor Embedding (t-SNE), which is a dimension reduction technique suited for visualizing high-dimensional data[44]. This tool has significant application in data-intensive fields, for example in bio-medical fields for analysis of human genetic data and prognostic clustering of tumour sub-populations[45].

Typical engineering and geoscience challenges are finding analogues, similarities and unique image clusters. As an example, a t-SNE is used to visualise a dataset consisting of several hundred well reports. Without any prior knowledge, the t-SNE visualization is effective in clustering images with similar

[44] Maaten, L. v. d. and Hinton, G., 2008: Visualizing Data using t-SNE. Journal of Machine Learning Research, vol 14.

[45] Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J.,Reinders, M. J. T., Walch, A., McDonnell, L. A., and Lelieveldt, B. P. F., 2016: Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. PNAS October 25, 113 (43), pp 12244-12249.

features regardless of their geospatial proximity (Data Source: Geoscience Australia). As each image has been tagged, the origin is fully traceable back to the specific page in the original report.

A visualization of thin sections shows how the images have been clustered and one image cluster in particular stands out and can be further investigated (Figure 6).

In cases where an exploration area of interest has been identified, it is possible to search the database for information available in the same area. Two map images are clustered together with the same map but with two different wells highlighted (Figure 7).

If an image is of interest but the origin is unknown, the image can be ingested and any identical or similar images can be identified.

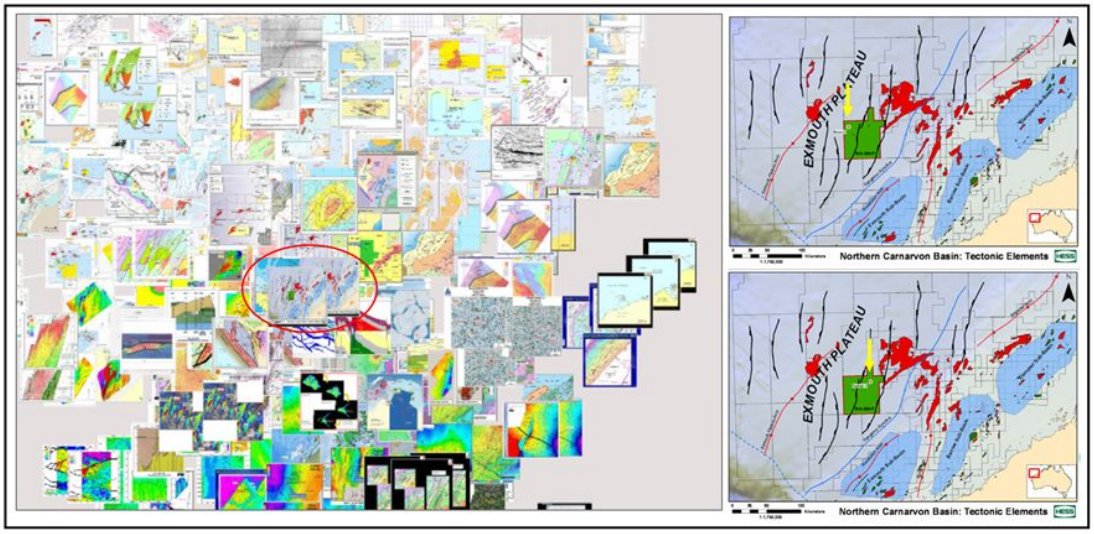

EXMOUTH PLATEAU
Northern Carnarvon Basin: Tectonic Elements
HESS
EXMOUTH PLATEAU
Northern Carnarvon Basin: Tectonic Elements
HESS

Figure 7: An area of interest has been defined and other maps which have information about that specific area can be identified (The yellow arrows highlight the differences in the two maps).

Advances in NLP enables translation in any language. An explorationist can go to South America, or Africa, in search of hydrocarbons, mine through troves of engineering and geoscience documents written in Spanish, or French, and get real-time machine-aided translations in English. This is breaking down language barriers and opening doors for new opportunities in foreign countries.

IE

Unstructured Big Data is difficult or even impossible to access manually in order to search for geoscience and engineering relationships and knowledge.

While it is probably not realistic to extract efficiently every piece of information from unstructured documents, achieving an extraction rate of 75% for example, would tilt the ratio of usable to unusable data from 20/80 to 80/20[46].

[46] Blinston, K. and Blondelle, H., 2017: Machine learning systems open up access to large volumes of valuable information lying dormant in unstructured documents. The Leading Edge, March, pp 64-68.

When data analytics is applied to organised data, higher order analysis can be performed. A knowledge graph can make the exploration history in area instantly available, the geo-density of information is promptly accessible, and microscopic analysis is immediately available from all wells in an area, to mention a few examples[47].

The 4 V's of big data

Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information. Big Data was originally described by the 4 V's (Velocity, Variety, Veracity and Volume) and more recently one more V has been added (Value) as well as the term Complexity.

Based on the original 4V description of Big Data it is clear, that processing and organizing of unstructured data can manage and overcome the issues related to the 4V's as listed in Table 2[48] [49].

TABLE 2: 4V's OF BIG DATA

	DEFINITION	BENEFITS
Volume	The main characteristic that makes data "big" is the sheer volume	Millions of pages can be processed on a monthly basis and the process is easily parallelized
Variety	The difference in data types and formats	More than 100 different format types can be automatically processed
Velocity	The frequency of incoming data that needs to be processed	It is estimated that data ingestion is up to 40 times faster than a manual approach, which will further improve as capabilities and training data are enhanced
Veracity	What is the trustworthiness of the data	Automated workflows can verify the data, remove duplicates and identify issues: With good training data, processing accuracies approaching 99% can be achieved

Business value from organizing the unstructured data

Digital transformation value

A digital transformation of the oil and gas industry has been predicted by the World Economic Forum to be able to provide 1 trillion USD value to oil and gas companies[50] and Barclay Equity Research predicts a 10% production cost reduction and a 10% recovery increase[51]. This is probably why 50% of oil and gas executives have begun using AI to solve challenges in their organizations according to an EY survey[52].

[47] Maver, K. G., Hernandez, N., M., Baillard, F. and Cooper, R., 2020: Processing of unstructured geoscience and engineering information for instant access and extraction of new knowledge. First Break, vol 28, June, pp 59-64.

[48] Maver, K. G., Hernandez, N., M., Baillard, F. and Cooper, R., 2020: Processing of unstructured geoscience and engineering information for instant access and extraction of new knowledge. First Break, vol 28, June, pp 59-64.

[49] Baillard, F., Maver, K. G. and Hernandez, N. M., 2019: A new way of handling unstructured data in the age of digitalization. EAGE Subsurface Intelligence Workshop, 8-9 December, Manama, Bahrain.

[50] World Economic Forum, 2017: Digital Transformation Initiative. Oil and Gas Industry. White paper. https://reports.weforum.org/digital-transformation/wp-content/blogs.dir/94/mp/files/pages/files/dti-oil-and-gas-industry-white-paper.pdf

[51] Barclays, 2020: North America Oilfield Services & Equipment. Frac to the Future; Oil's Digital Rebirth. https://novilabs.com/wp-content/uploads/2020/02/Barclays_Frac-to-the-Future-Oils-Digital-Rebirth_01152020.pdf

[52] EY, 2020: Applying AI in oil and gas. EY. https://www.ey.com/en_jo/applying-ai-in-oil-and-gas

92% of oil and gas companies are either currently investing in or planning to do AI in the next 2 years according to an EY survey[53] and a similar response was the result of the DNV-GL 2020 oil and gas outlook regarding maintaining or increasing digitization investment in 2020[54].

It has been widely recognized that the value of open and persistent data grows as they become discoverable, citable, re-usable, integrated, and linked with other data[55].

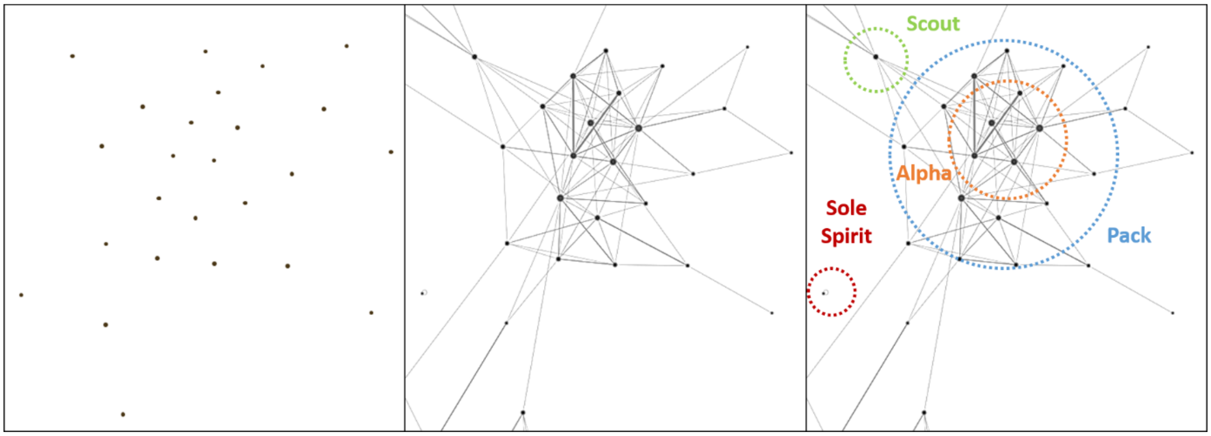
Management of unstructured data has the potential to enhance both the productivity and competitiveness of a company by enabling exploration and production teams to gain insights through identification of patterns, relationships, and anomalies that can help them make better decisions faster, improving success and recovery rates, as well as bottom lines in both the short and long-term[56]. Deploying technologies, such as ML and NLP, with domain knowledge, can further help companies to better analyse data, identify patterns and predict outcomes, leading to more informed decisions[57].

With the continued ingestion and aggregation of more unstructured data, ML processes will improve and thereby resulting in better data analytics providing a continuous exponential knowledge explosion across geography, organization and disciplines (1+1=3).

IE examples

By organizing unstructured data, Parsley Energy has made it possible for staff to spend less time looking for data and more time for analysis, resulting in 60% faster exploration and production decisions due the enhanced management of well information[58].

A knowledge graph can visualize the dependencies between wells and illustrate development in a basin over time. For example, regional exploration in a large basin with sub-basins results in more than 160 wells covering nearly 50 years of drilling history.



[53] EY, 2020: Applying AI in oil and gas. EY. https://www.ey.com/en_jo/applying-ai-in-oil-and-gas

[54] DNV-GL, 2020: New directions, complex choices. The outlook for the oil and gas industry in 2020. https://industryoutlook.dnvgl.com/2020

[55] Lehnert, K., and Hsu, L., 2015: The new paradigm of data publication. Elements, 11(5), pp 368–369.

[56] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[57] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[58] Chouhan, C and Woerdle, C., 2018: Managing well files and the unstructured data dilemma. Hart Energy. https://www.hartenergy.com/exclusives/managing-well-files-and-unstructured-data-dilemma-176405.

Figure 8: Knowledge graph. Left map; well locations. Middle map; well importance by size of well location and well interdependence by thickness of the lines between the wells. Right map; resulting interpretation of the drilling history.

The analysis is constructed by interrogating each well in the database and investigating the degree of correlation with the remaining wells. An example of such a knowledge graph is shown in Figure 8 for one basin, where the size of each well node corresponds to the number of links with other wells and the line thickness connecting two wells illustrates their level of inter-dependency[59].

Using the analogue of a wolfpack the knowledge graph shows a level of hierarchy and organization among the wells:

- The “alpha(s)” are the wells in the centre of the pack. They are essential for the understanding of the cluster and highly connected to the rest of the cluster. They provide a good analogue to the rest of the “pack” of wells and represent the “type” wells for the basin.
- The “scout(s)” are the wells connecting two different well clusters. They provide an essential link between clusters, geographical regions, geological basins etc.
- The “sole spirit(s)” are the wells that are not connected to any other cluster. It could represent a new exploration opportunity in an unknown territory.
- The “pack” are the wells that are part of a cluster of the wells without any particular distinction but can be used to analyse step-out near-field exploration effort.

[59] Hernandez, N. M., Baillard, F. and Maver, K. G., 2019: An effective G&G exploration strategy inspired by a wolfpack. Abstract, Force meeting, September, Stavanger, Norway.

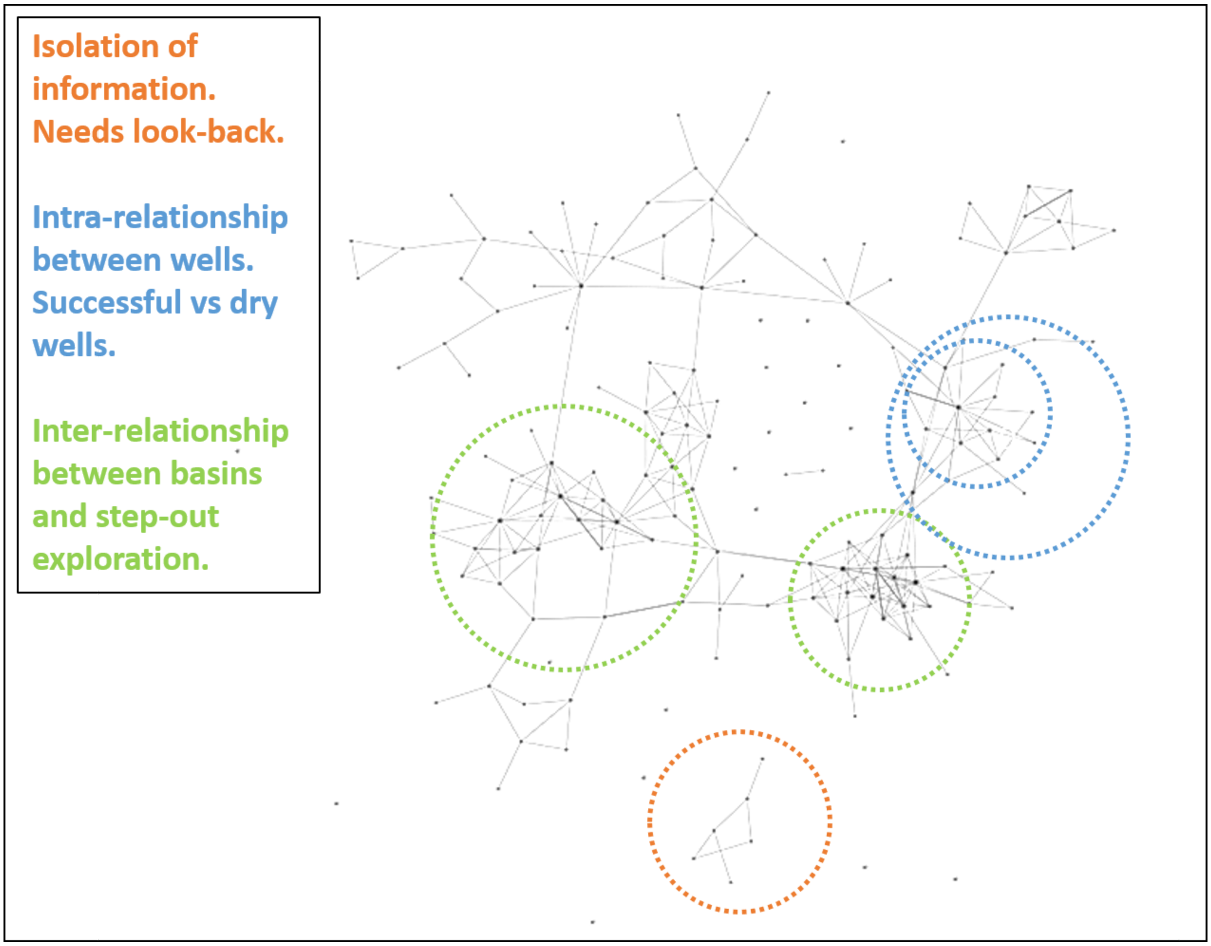

Isolation of information. Needs look-back.
Intra-relationship between wells. Successful vs dry wells.
Inter-relationship between basins and step-out exploration.

Figure 9: Interpretation of the drilling history regionally using the knowledge graph.

The wolfpack terminology can be extended regionally with additional sub-basins (Figure 9).

This classification from the knowledge graph provides the geoscientist and engineer with a new insight regarding where to focus the exploration effort based on the complex interrelationship between already drilled wells. For a new venture operation, it will provide unique insight, and speed-up the exploration effort dramatically as for this example, where nearly 50 years of drilling history from 160 wells were made available in 3 months[60].

Retaining and maintaining corporate knowledge

A GE oil and gas survey predicts that 50% of the oil and gas workforce will have retired by 2025 and there is no pipeline of younger workers to fill the gap[61] as the next generation of workers to replace them are less interested in a career in the oil and gas industry[62].

[60] Maver, K. G., Hernandez, N., M., Baillard, F. and Cooper, R., 2020: Processing of unstructured geoscience and engineering information for instant access and extraction of new knowledge. First Break, vol 28, June, pp 59-64.

[61] Annunziata, M., 2016: Digital Future of oil & gas & energy. Presentation. https://s3.amazonaws.com/dsg.files.app.content.prod/gereports/wp-content/uploads/2016/02/22094804/GE_Digital_Future_WP-02191611.pdf

[62] Williams, J. and Strier, K, 2020: Is AI the fuel oil and gas needs? EY. https://www.ey.com/en_gl/oil-gas/is-ai-the-fuel-oil-and-gas-needs

The combination of the big crew change coming in the oil and gas industry, mergers and acquisition changing the industry landscape and staff changes due to fluctuations in the business environment are making it harder to capture and maintain corporate knowledge. This challenge can be much better managed by having a process for making unstructured data easily captured, organized and accessible in a central cloud-native data lake. It is therefore critical that oil and gas companies take measures to invest in understanding what data and information they have and invest in managing that data so it can be discovered and used[63].

Conclusion

The biggest time and financial sink are not in the actual data processing, but in the evaluation and preparation of the content to be processed, and in the development and implementation of quality assurance and quality control processes. Overcoming the unstructured data challenge requires devising and implementing an effective and efficient processing pipeline, developing robust taxonomies, accessing extensive content to help 'train' and refine capabilities in NLP and semantics text analytics, and generating tools that support the discovery of the newly structured content that align or are integrated into the users workflows[64].

As recently stated by Schlumberger CEO, Olivier le Peuch "Our customers recognize the potential value of digital transformation, and are evolving the methods of finding, developing, and producing hydrocarbons. They will rely on increased volumes of data captured and consumed through integrated and reengineered workflows, allowing them to make faster and better decisions."[65].

Digitization and managing unstructured data will have a significant impact on the oil and gas industry in the future and commercially available web-enabled and cloud native tools can cost-effectively provide this digitization transformation.

In the competition for exploration and production opportunities increases, those oil and gas companies embracing digitization will gain enormous commercial advantages[66].

[63] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[64] McMellon, P., 2019: Unstructured data is risky business. R&D solutions for oil & gas. Elsevier. https://www.elsevier.com/__data/assets/pdf_file/0006/518181/Unstructured-Data-is-a-Risky-Business.pdf

[65] Peuch, O. L., 2020: JP Morgan 2020 Energy, Power & Renewables Conference, June. https://www.slb.com/newsroom/presentations/2020/le-peuch-speaks-at-jp-morgan-energy-conference-2020

[66] Maver K. G. Mamador, C. and Hernandez, M. H., 2020: Explore your unstructured E&P data for new value. Data Science and Digital Engineering in Upstream Oil and Gas. https://pubs.spe.org/en/dsde/dsde-article-detail-page/?art=7216