ECIM

ECIM & CDA

Joint Data Management Workshop

Aberdeen

30th November 2016

Digital Dividends from Subsurface Data: Data Science meets the Unstructured Data Challenge

Proceedings

Contents

1 Workshop Overview

1.1 Background

CDA was established by the UK oil industry in 1995 with the aim of sharing the costs and the benefits associated with managing subsurface exploration and production data through collaborative working.

ECIM – the E&P Community for Data and Information Management – is a non-profit foundation established by the industry in Norway to promote the data and information management profession and best practice within the E&P Industry, most notably through its annual two-day conference in Haugesund, Norway, perhaps the premier data management event outside the United States.

ECIM and CDA have agreed to work together in Europe to facilitate and encourage data management professionals to enhance their professional and technical competence, primarily through development, community building, and establishing a Body of Knowledge for the discipline.

As part of this work, both organisations have undertaken to organise a series of workshops on themes of immediate relevance to industry data managers based in Europe.

These are the proceedings of the third such workshop, held in Aberdeen on 30th November 2016.

1.2 Workshop Purpose

Modern data science techniques are widely regarded as being of some value when applied to the sub-surface data domain, but the source of that value is not clear. Case studies are scarce, potentially because those organisations that have realised value from data science techniques are constrained by commercial confidentiality restrictions in talking about them.

In an attempt to address this issue, and generate evidence for the value of data science when applied in the sub-surface, CDA announced an 'Unstructured Data Challenge', in which it would make available, under a non-disclosure agreement, a substantial fraction of its data holdings with the explicit aim of enabling a public demonstration of the value of data science and data analytical techniques.

The Data Challenge proved of wide industry interest, and of the forty companies that requested more information, nine decided to progress to the next stage: execution of a defined project involving CDA's data holding that aligned with the challenge purpose. CDA did not charge for participation in the challenge, nor did it pay any participant to take part.

Each of the nine companies received the same set of data – roughly 3.5Tb of well data, documents (some machine readable, but the majority as scanned images), and seismic documentation (but no trace data), supplemented by an export of CDA's well and seismic header database – and set to work over the summer of 2016, to develop, demonstrate, and validate their data science capabilities. Detailed statistics regarding the data provided are given in Appendix C: Data Provided to Participants.

The initial outcomes of the Challenge were announced by CDA at ECIM's conference in September, and the full results were presented in detail by the participants at this workshop, hosted by CDA and ECIM, on the 30th November 2016, at Village Hotel in Aberdeen.

1.3 Workshop Programme

Start	End	
11:45	12:20	**Sandwich Lunch, Tea, Coffee, and Registration**
12:25	12:30	**Welcome and Introduction** Malcolm Fleming, Chief Executive, CDA
12:30	12:50	**NDB:** Ed Evans
NDB Data Discovery: Using analytics to produce new data		
12:50	13:10	**Flare IM:** David Camden
Content Classification and Analogue Identification based on Text Analytics		
13:10	13:30	**Hampton Data Services:** Wally Jakubowicz
Autonomous Multi Faceted Metadata Capture from Image, Text and Industry Standard File Formats, and Classification to Multiple Taxonomies		
13:30	13:40	**Q&A: Presenters 1-3**
13:40	14:00	**Refreshments and Networking**
14:00	14:20	**Independent Data Services:** Colin Dawson
Visual Data Mining: Unlocking the hidden value of unstructured data		
14:20	14:40	**AGR Software:** Håkon Snøtun
From Unstructured to Contextualized Data: All data becomes Big if you fail to learn from it		
14:40	15:00	**Agile Data Decisions:** Henri Blondelle
Running iQC on the CDA Unstructured Data Asset: Why extract information from unstructured documents when a structured database does exist?		
15:00	15:10	**Q&A: Presenters 4-6**
15:10	15:30	**Refreshments and Networking**
15:30	15:50	**KADME:** Gianluca Monachese
Practical Applications of Data Analytics Techniques Using Unstructured Data: Results of the application of existing functionality		
15:50	16:10	**Cray Inc.:** Maria Mackey
Commencing an Analytics Workflow on the CDA Dataset: Understanding a large, old and varied dataset for an analytics pipeline.		
16:10	16:30	**Schlumberger:** Paul Coles
CDA Challenge: Using Analytics to Derive Additional Value from Unstructured Data – Initial Findings		
16:30	16:40	**Q&A: Presenters 7-9**
16:40	17:00	**Panel Discussion Session**
17:00	17:15	**Conclusions & Next Steps: Led by Malcolm Fleming**
17:15	-	**Workshop close & 'grab a beer'**

1.4 Workshop Presentations

The presentations given by each of the speakers are available to registered workshop attendees only at the ECIM Workshop website:

www.ecim.no/download301116

1.5 Speakers

CDA and ECIM were pleased to welcome the following speakers to the Workshop:

- Ed Evans, NDB
- Dave Camden, Flare IM
- Simon Fisher, Hampton Data Services
- Colin Dawson, Independent Data Services
- Håkon Snøtun, AGR Software
- Henri Blondelle, Agile Data Decisions
- Gianluca Monachese, KADME
- Maria Mackey, Cray Inc.
- Paul Coles, Schlumberger

Full speaker biographies can be found in Appendix B: Speaker Biographies.

2 Presentation: NDB – Ed Evans

2.1 Presentation Overview

Ed began his presentation by emphasising NDB's focus on business purpose at the root of all analytical endeavours – to ensure that in the pursuit of data analytics, the sources of value of data, and interpreted data in particular are understood, communicated to leadership, and exploited through effective integration of professional data managers within a broader business team.

Prior to the announcement of the Challenge, NDB had been tasked with analysis of data from the Norwegian side of the North Sea, with the aim of identifying information on oil and gas shows. The process followed – a manual, laborious exercise – looked at well information in detail, integrated this into a GIS database, and resulting in a co-visualisation of shows data with existing NCS and client data.

NDB began their exploration of the CDA data set by developing a set of tools that could automate the creation of shows data, as per the Norwegian client requirements, for the UKCS. Through developed of a scripted approach to data ingestion, analysis, and summarisation, NDB was able to reduce the amount of time taken for shows analysis from 20 days per 100 wells to 4 – a five-fold improvement.

Ed's thoughts then turned to use of data science techniques to improve the means by which subsurface data might be organised, and considered a number of approaches prototyped using mind-maps specific to the particular circumstances, and particular business purpose for which the organisational approach was intended. This flexible approach to categorisation enables the development of specific solutions to particular business issues encountered, whether around geological interpretation, well planning, or higher level industry issues, such as reduction in drilling costs, or development of options to exploit the Palaeozoic.

This suggested that the approach to categorisation to be adopted should be more business focussed, and purpose oriented than a static, classification-based taxonomy. Ed introduced the work of the Optique Project, which seeks to describe sources of information of relevance to a geoscientist using a shared ontology – a layer of meaning associated with the data, that enables independent information systems to share the same definition and concept of a well, or a seismic survey – and hence enable geoscience search queries to be executed across independent systems, and the results integrated into a single, holistic view.

The Optique Project[1], an EU project funded by Statoil and numerous industry partners, and based at the University of Oslo, has now come to an end, and its results are being taken forward by Sirius Labs[2], also at the University of Oslo, with a view of facilitating industry adoption, commercial exploitation, and onward development of the Optique project deliverables.

Ed concluded by noting that as part of its Autumn Statement, the UK Government is making funds available through Innovate UK[3] for transformational digital initiatives – suggesting that there is money out there for organisations looking to develop and implement current data science techniques – and ontological approaches – within the oil and gas industry.

[1] See http://optique-project.eu/

[2] See http://sirius-labs.no/

[3] See https://www.gov.uk/government/organisations/innovate-uk

3 Presentation: Flare IM – Dave Camden

3.1 Presentation Overview

Dave gave a presentation of two parts, the first considering the application of Flare's taxonomy to the documents in the data set; and the second looking at novel uses of text analytics to measure similarities between documents in the data set, based on linguistic analysis only.

Flare's research made use of the machine readable sub-set of well documents provided. Documents requiring OCR were not considered, and hence offer a substantial extra source of information and value, should a sufficiently reliable OCR process render them machine readable.

3.1.1 Use of the Flare Taxonomy

Flare applied its E&P Taxonomy to the challenge dataset, to investigate the fit between the taxonomy and the body of documents stored within UKOilandGasData, to determine if certain documents could not be classified (and hence indicate areas where the taxonomy could be improved), and to provide a view on the level of accuracy with which the CS8 classification has been applied within UKOilandGasData.

In preparing each document for indexing, Flare's processing pipeline identified synonyms for oil and gas terminology (including a number of new synonyms not previously noted); eliminated common words (e.g. a, the, in, etc.), and applied a stemmer to render equal the words of the same root, but rendered using different parts of speech (e.g. drilled, drilling, etc.). This pre-processing assists in further stages of the text analytics pipeline where word proximity and co-occurrence are considered.

The result was a taxonomy that mapped fairly well to CS8, though indicating a 20% disagreement between the two. Mapping to PON9 requirements also proved useful, on a well by well basis, as a potential future support in identifying where regulatory submission requirements had, or had not been met.

Classification within the Flare taxonomy also improved awareness of the availability of relevant data, by enabling tree-based visualisation of the kinds of documents that match text searches, and so providing a very graphical view of what is available. Through analysis of synonyms (e.g. an EOWR, a FWR, and an EWR are all End of Well Reports) and application of geological understanding (e.g. that limestone is a carbonate), further improvements in search relevance were made.

In concluding this section of Flare's work, Dave summarised by observing that the work had provided an effective stress test of its E&P Taxonomy, and surfaced a staggering number of synonyms used within the oil and gas industry in the last 50 years. He noted that future efforts in this area would be based on graph database technology, rather than traditional relational databases, as the ability to assign weights between data relationships provides more effective support for the development of machine learning techniques.

3.1.2 Linguistic Analysis of Well Documents

In the second part of his presentation, Dave shared the results of work on the analysis of the linguistic content of the data set, and the ability to calculate the similarity between document concepts on the basis that you 'can tell a word by the company it keeps'.

This 'distributional hypothesis', backed up by detailed work in the calculation of 300 parameter linguistic fingerprints, enabled similarities between geological terms to be determined purely by the context in which the terms appear within the document set, with

minimal human intervention – important, as there is now too much content within an average E&P organisation for humans to become involved in the analysis and categorisation of new and historic data.

Flare's use of a two-layer neural network in the classification process marks the beginning of their work towards creation of search-based applications, moving forward from rules-driven expert systems, towards more general approaches in which configuration happens through a learning process involving digestion of curated training sets.

The result is a classification engine that does not suffer from human cognitive bias, and hence is able to generate non-intuitive, challenging results from queries that have the potential to lead to new insights.

As a starting point in their experiments on linguistic analysis, Flare created a tool to identify Formation Analogues, through which users can identify similarities between geological formation based on parameters including age, lithology, depositional environment, etc., or solely based on a series of search terms provided.

While the Formation Analogues tool is not yet productised, it provides an intriguing view into what is possible through the calculation of linguistic similarity, particularly as the approach is readily applicable to other domains where a sufficient document base exists, such as plays, prospects, and field.

3.2 Conclusions and Q&A

Dave was asked if Flare's approach used off-the-shelf, or custom developed tools. He replied that they had used both, but focussed in their technology choices on the business process, and the configuration of the tools available.

He also responded to a question on the broader applicability of the techniques he described by observing that much of the problem in the operation of machine learning technologies such as that used in the Formation Analogues tool is in the availability of computing power, suggesting it is important to choose problems that scale with the availability of computing resources available: there is no such thing as intelligent machines – just an illusion created by the application of large amounts of computing power!

4 Presentation: Hampton Data Services – Simon Fisher

4.1 Presentation Overview

Hampton Data Services (HDS) took on the challenge along with a partner, Zorroa (a US-based business intelligence company with experience in film and media industries). They utilised a convolutional neural network (CNN) to perform image analysis, along with fuzzy text searching and text classification, to classify report pages within confidence thresholds. For this study they defined a scope that excluded very large documents, that performed analysis on log headers only, not the entire log, and that excluded complex images (such as VSP).

The business questions prioritised were firstly whether a report's title and sub-type truly indicate its content, and how a data owner might determine if their log data catalogue is valid and complete. Their approach treated all logs and report pages as images, and in the first instance used document-level statistics and metadata to begin classifying items; e.g. a 'large bulk' of image content is likely to be a log. This can then be confirmed as part of 'training' the CNN.

They also employed text analytics to identify phrases and score reports against the occurrence of those phrases; however, the optical character recognition (OCR) process resulted in many misreads, so they resorted to 'fuzzy searching' (expanding the search parameters to include partial matches), which results in more 'noise' to be filtered out. Nevertheless, in this way the data was scored and ranked based on its actual content, not just its title. Their results showed that some documents had been classified incorrectly originally, while those classified under 'general' codes could be assigned more precisely.

For the second business question, they made use of metadata from the structured log data provided, to plot well operations against dates and depths. This provides a visual indicator of operations and therefore expected data as the well progresses. However many of the log (.lis) files were missing depths and dates, affecting the outcome of the plots. Future work to include metadata from the unstructured reports and scanned/image logs will greatly improve results. Assuming the missing metadata can be populated or validated from other data types, they aim to create an overall QC well view.

Machine learning and image recognition enables the classification of data items based directly on content, with the support of metadata. This has potential for applying classification schema more accurately, and enabling data items to be assigned multiple classification codes. With additional training datasets HDS expects the CNN tool to be developed further and improve results.

5 Presentation: Independent Data Services – Colin Dawson

5.1 Presentation Overview

Colin Dawson presented IDS's submission to the data challenge.

The aim of the solution was to find required information that was hidden in the unstructured data in a timely manner or to "unlock the hidden value of unstructured data".

IDS's approach was to mine and search the data using open source and affordable solutions that were accessible through a web interface. This would reduce the time spent on searching for data, reducing operational spend and improve planning.

There was a three step process deployed to achieve the objectives:

1. Process the PDF and Word document data to make it machine readable then convert logs etc. to machine readable formats. This was achieved using LogStash to read the data, transforming the inputs and adding structure as required using a Grok ruleset, and then outputting the results in a consistent JSON format, suitable for ingestion into a search engine.
2. Enable text search on all data to make it searchable. Elastic search was used as the search software.
3. Present the data in a visualisation tool, Kibana, where it is easily digested.

The result of this process was the ability to search for anything within the unstructured documents. Colin presented an example of searching end of well reports for the term "stuck pipe". The solution presented the results of the search on a map display, all those wells whose end of well report contained "stuck pipe". The results were able to be cross-referenced with available stratigraphic data.

The challenges encountered by IDS were based around adding structure to the unstructured data and the process of OCR-ing the documents. Having non-standardised well reports etc. made the process more difficult than had they been one standard format. Many documents could not be OCR-ed. The real time sink was not the processing of the data itself, rather, preparing the data for processing.

5.2 Conclusions

With the loss of domain knowledge resulting from the industry downturn, the value that can be extracted from the data will rely heavily on how accessible it is and how efficiently it can be accessed. IDS want to be on the crest of the data science wave to address this issue.

Unstructured data mining is affordable using open sourced technologies. The use of several data sources (CDA, NPD, OGA stratigraphy etc.) was key to maximising value.

Colin noted that CS8 has no provision for workover data, suggesting the standard, and the PON9 Basic Set should be expanded to encompass all the data types generated during a well's lifecycle, and which must be archived after a well is plugged and abandoned.

6 Presentation: AGR Software – Håkon Snøtun

6.1 Presentation Overview

Håkon began his presentation by highlighting the main source of value from historical data – its use in making predictions regarding future events: enabling companies to move up the data value pyramid, to use data to make good decisions, rather than to enable the quality of decision making to be evaluated in hindsight.

Within the software world, Håkon observed that there is no shortage of technology to choose from. The challenge is to identify the tools that are relevant to the problem at hand, rather than those with familiar, impressive brands behind them.

In its work on the challenge dataset, AGR focussed on visualising the results of text analytics (Håkon showed Anscombe's quartet to emphasise the importance of good data visualisation in developing a proper understanding of one's data), with a particular focus on the content of end of well reports. AGR's work aimed to assist the process of digesting these reports, which may be 650+ pages in length, to enable a high level view of the purpose, status, and outcome of a well to be ascertained in a glance.

AGR passed the end-of-well reports in the dataset through a text processing pipeline, including OCR, Lucene for lemmatisation and stemming, and then use of Apache tools to extract relevant paragraphs, and other chunks of text. Subsequent focussing on titles, chapter summaries, and paragraph headings enabled automatic creation of well summaries, through which context on the well, and the good and bad experiences encountered while drilling it could be extracted, classified, stored, and visualised: moving the information within from available to accessible; and from accessible to contextualised.

6.2 Conclusions

Through improved accessibility of contextualised data, well planners are able to understand rapidly what is known and not known about previous attempts to drill similar wells, enabling future well designs to be improved. As Håkon highlighted through his example on the survivability of aircraft during the second world war, awareness of the data missing from your model can be crucial in making good decisions.

Håkon also offered suggestions for regulators to improve the accessibility and utility of regulatory data submissions for data science purposes. Notably, he suggested a movement away from PDF as a format for data submission, requesting that structured data be submitted in a structured format instead to preserve its machine readability; that narrative documents be broken into standard sections to support future readability; and for data to be made available earlier, and more freely, to support community feedback in the shared effort to improve data quality.

7 Presentation: Agile Data Decisions – Henri Blondelle

7.1 Presentation Overview

Henri started by outlining the main issue to solve, namely that of enabling value to be extracted from unstructured data to the same extent as value can be extracted from structured data. Current estimates are that in the CDA dataset, about 20% is well structured. This data is easy to mine but the value that can be extracted is already well known and of a known richness determined by the structure. Unstructured data however is more challenging to mine but can be very rich in content and potential value. Many other industries have been faced with this problem. The prize is potentially huge. Increasing the structured content by even a small amount can have big benefits. To move all the unstructured data to a structured database by traditional means would require a massive human effort and would need the expertise of a data and domain expert. However, the new technologies emerging in the analytics and machine learning sphere can help reduce the burden of cataloguing (extraction of metadata from the content).

Machine Learning systems can save money by automating metadata extraction, can reduce the time between data acquisition and the final decision and can reduce risk in the final decisions by using more verified information (improving quality control). Pattern recognition techniques taken from photographic image analysis can be applied to text and data elements of scanned images. This enhances traditional OCR (optical character recognition) and can 'learn' based on a statistical approach combined with some user inputs to train the system.

Agile approached the CDA Data Challenge using the iQC tool. This tool uses a user defined taxonomy and the ability to QC results and feed them back through again to train the machine. The learning model is constantly updating to improve results and can work on structured and unstructured data.

The results from the CDA Challenge were very encouraging. Agile noted that the CDA taxonomy was easy to map to and raised the possibility of associating the same document to multiple CS-8 codes. This could have profound effects on the CDA dataset. For example, it might show that a well is actually more complete than we currently believe it to be because at the moment each CS-8 code requires a distinct data item to be loaded to demonstrate completeness. The process could be used to clean-up and re-catalogue the existing CDA data and also be used as part of the data submission workflow for new data.

7.2 Conclusions

More fine tuning of the machine learning model is required. Start small scale and scale up. Ideally a test dataset such as a bulk submission project or re-cataloguing exercise to demonstrate the value and effectiveness of the tool – Agile would welcome approaches and suggestions for pilot projects in this regard.

8 Presentation: KADME – Gianluca Monachese

8.1 Presentation Overview

KADME approached the challenge using Whereoil, their data integration platform for structured and unstructured databases. They digitised scanned documents via OCR, and then combined the structured information in CDA's catalogues with the indexed content of the documents. They then geotagged any document that contained coordinate information, and were able to generate 'heatmaps' to show the geographic distribution of search results.

They also carried out automated QC, using intelligent search tools to map information to a defined ontology. In addition, they extracted information directly from the structured curve data, which itself served as a useful baseline for training algorithms to identify quality issues or indeed confirm accurate data.

8.2 Conclusions

It was highlighted that relevant domain expertise is vital in order to drive the technology efficiently. Their involvement in the SIRIUS project (Centre for Scalable Data Access in the Oil and Gas Domain) was also noted as a beneficial resource of research and competence to pilot further development of this work.

9 Presentation: Cray Inc. – Maria Mackey

9.1 Presentation Overview

Maria Mackey presented Cray's experience of the data challenge. Although Cray did not provide any results from the data challenge, there were several useful insights gained from their time with CDA data.

Cray partnered with industry partners NDB and Venture to scale up their respective solutions with Cray resources. Cray have access to powerful super computers as well as in-house software but note that in order to deliver an effective analytics solution, skills are required from a variety of other areas, including data scientists, solutions architects, and experts in the business domain under analysis.

Cray identified that the time taken to get insight from data was traditionally measured in minutes and hours for batch analytics, whereas nowadays the demand is for insights in seconds or less for interactive analytics.

Cray used a combination of open sourced software (OpenStack) and Cray proprietary solutions in the challenge. This was paired with Cray's supercomputer hardware, Urika-GX, to mine and analyse CDA's data.

Cray's approach was to load and OCR text from a small set of well data then scale up the exercise to all the wells within the dataset. The objective of this was to obtain structured metadata from the unstructured data. This could then be used to investigate inconsistencies and identify new relationships.

The first phase of the project was to parse the data to identify what file types the CDA data contained. The results were returned in 10 minutes for the 490,000 input files and identified TIFF, PDF, text and other custom formats.

Cray then ran OCR on all data, in parallel, which took just over 2 hours. The results showed that 70% of the files were empty or unreadable by the software (Apache Tika), though this may be due to the inability of Tika to make sense of industry format well data files (e.g. LIS and LAS) – more detailed investigation is planned here.

Cray has not yet proceeded to the second phase of their program, to identify and investigate inconsistencies. A revised phase 2 has been drafted to investigate the issues they had parsing the data, identify improvements in data loading, convert good data to RDF format and then identify relationships through graph visualisation.

9.2 Conclusions

Analysis is usually the easiest and least time consuming part if requirements are provided. Data quality is key, as are solution architects.

Cray is open to approaches from academia and industry from those wishing to develop computational techniques within the oil and gas domain, for which free access to a Cray supercomputing environment may be of assistance.

10 Presentation: Schlumberger – Paul Coles

10.1 Presentation Overview

Paul presented the initial results of Schlumberger's application of its technology stack, and that of a number of its industry partners, to the CDA data set, using a variety of Cloud-based technologies, resulting in an analytics pipeline with the capability of taking geological data from a raw (albeit organised) form and transforming it into a petrophysical model suitable for review and quality assurance by a petrophysicist.

The pipeline begins with collation and harmonisation of the data once loaded into Schlumberger's system, including a number of steps to address raw data deficiencies – for example, use of a machine learning model to improve the accuracy with which log curves are classified, and identify common errors, such as incorrect assignment of units (e.g. feet, rather than metres). Schlumberger also used WIPRO's Holmes, and open source tools to OCR scanned images of well documents, classify them, and generate a searchable text index of the full document corpus.

The next step involved automation of the process of well log quality control, including application of environmental corrections, depth shifting, and merging and splicing, driven using a machine learning model executing within Google's TensorFlow environment, to generate a final set of log curves suitable for incorporation into a petrophysical model.

To ensure the petrophysical model output also made geological sense, Schlumberger applied constraints taken from the OGA's geological tops database, and from cuttings analysis, to ensure alignment between the model, the geological predictions it makes, and the geology that was actually observed while the well was drilled.

Finally, the resulting model was visualised using a variety of Schlumberger tools, to support regional formation mapping across the UKCS, to identify working hydrocarbon systems that appear underexplored, and to display correlations between well properties at the reservoir level, at scale, across every UKCS quad and block.

10.2 Conclusions

Analytics is an area of significant research and development activity within Schlumberger at present.

For example, the analytics pipeline developed for the challenge data set was applied to forty-six wells within the Piper field, resulting in automated generation of a geologically plausible model suitable for professional petrophysical review in just six hours of computation, as opposed to the six weeks that otherwise would be required for the workflow if it was executed manually.

11 Question & Answer Summary, Conclusions, and Next steps

11.1 Q&A

Q: How do you go about establishing relationships between documents in the data set provided?

A: We considered logging programmes and summaries to see what was carried out, and attempted to make the match between what's in the logs and what's in the reports.

Q: Did you use freely available tools and software in your work, or was most of your software developed purely in-house?

A: Yes, we used standard OCR toolkits and GIS systems. We also mixed the toolkits with prior knowledge, to very carefully bias or weight the results. But at the same time we need to be able to let go and make new discoveries through analysis results.

A: Spelling mistakes and misreads are an issue with OCR, but not so much with newer native documents.

Q: Can you pre-select the dataset that you analyse and therefore get rid of more noise?

A: We can attempt to do this in some situations, e.g. decommissioning data, where certain specific data types are clearly identifiable.

A: Some data types are fairly standard, e.g. well test analysis; but there could be a wide variety of information within a PowerPoint presentation. We should focus on the known standard data types to improve consistency.

Q: There has been comment on poor quality scanned images – could the tools flag these, for CDA to alert its members to identify better copies or to re-scan the items?

A: If the item or well is of value, it would be worth spending the time and effort to do so. The difficulty with scanned images is not necessarily quality, but also that text can be within set lines, or be inside a box, or vertical. We had to pre-process the data get the text out and then feed it into the machine learning tool.

Q: There is a difference between legacy data and point forward data. The OGA should determine the requirements for point forward data. As for legacy, OCR is just one method. Secondly, you talked about converting files to JSON – how confident are you that the conversion is accurate and preserves quality?

A: It's impossible to say if a conversion is good or bad, but the source provenance and metadata can be retained, which can provide a confidence score and indicate a level of data quality.

A: A final well report is not one thing to all parties. It's hard to determine the value overall, so an option is to extract what's of value at the moment, and retain the master copy to be revisited.

A: Looking at CDA final well reports, it's clear that some of the data there came from structured databases – so as a start the unstructured data could point you to the structured data.

Q: How did you view the role of standards in completing this work?

A: They played a huge role. The way companies categorise basic things can be very different. Standardisation definitely has a role in structuring datasets, and importantly open standards. There were a lot of proprietary standards in this dataset.

Q: There is a lot of unstructured, un-standardised data here, on which you managed to perform analytics. Do we then not need standards anymore?

A: Regardless of the type of dataset or repository, structure needs to be applied in some way, and standards greatly help with that – regulation can really help drive standards and structured data reporting.

A: We've seen here that a lot of data just didn't parse or achieve OCR, so that is locked potential. Is there enough money, and will, to unlock it?

A: Standards and regulations are connected. What we've seen a lot of is structured data being 'deconstructed' into unstructured data for the purpose of reporting.

Q: Do you think the standards should be defined by a committee of people, or by the machines?

A: The standards are perhaps slightly less important; what is important is what sort of information is being reported.

A: If you have a machine-learning standard that's ready to go, then use it. After all, standards have been driven by people up to now, and we still haven't resolved the issue.

A: A balance of both is required.

A: But who owns the machines?

Q: There is a sense that somebody should do something about this. Where should we look for leadership?

A: It comes from the biggest need. And currently in the UK that is MER (maximising economic recovery).

Comment: At the time that a lot of the CDA document scanning was performed, PDF was an immature format. We can't go back to scan the documents, but can the machines be tuned to cope and account for shortcomings in the older standards?

Q: A lot of what you've done is to structure the metadata. What will it now take to mine the knowledge and interpretation in there; to find analogues, etc.?

A: There is probably not too long to wait. This is likely an attainable goal in the next few years. It's also a matter of selecting parameters, and a machine can learn to do this.

A: If a human looks at a poor scan they can still recognise what it is – so, soon a machine can learn to do this too.

A: Also, this project was undertaken in a very short time span, so there is much more potential.

Comment: Scanned images, data science, image capture and contextualisation have all come a long way. There could be solutions to work through the 'poor' images. There are solutions out there, and I would encourage all parties to engage with academia to find them.

11.2 Conclusions and Next Steps

Malcolm Fleming closed the workshop by re-emphasising the need echoed by all the participants that data science activities be performed in search of answers to important business questions.

Information Management activities are now encompassing 'information exploitation' – unlocking the value of unstructured data, seeking business intelligence through analytics, and ensuring all information, both recent and historic is appropriately assessed and exploited – leading to better, more accurately risked decision making.

To support this work, it is important that data is made available in formats suitable for analytics (minimising the effort each participant expended in data preparation), an area in which regulatory intervention may be helpful. Regulators and organisations such as CDA should also consider those enabling steps that need only be performed once, on behalf of industry, (for example, basic data transformations such as effective OCR, and scanned image processing) to make national data archives more amenable to analytical techniques, and lower the cost of entry for new entrants into the sub-surface analytical space.

Malcolm noted that as ever, and as emphasised by the wide range of open software used by the presenters, technology itself is not the biggest challenge. Rather, industry should focus on the provision of data at a known, acceptable level of quality, in a usable format, thereby reducing or eliminating data preparation barriers that might otherwise dissuade companies from developing and innovating using UKCS data.

Finally, Malcolm observed the need for effective leadership in this area, to ensure real progress is made at the level of urgency required under MER UK.

OGA representatives at the workshop stated that this is something the OGA can and does want to help with, to reduce the burden of the decommissioning process, as well as to support new exploration and appraisal activity. There are huge opportunities here. Progress will initially be incremental, but as we demonstrate value, the uptake of data science and analytics within E&P will hugely increase.

12 Appendix A: Workshop Attendees

Organisation	Delegate Name
Agile Data Decisions LLC	Henri Blondelle
AGR Software AS	Håkon Snøtun
AGR TRACS International	Lynn Smith
Animus Technology Ltd	Peter Taylor
BP	David Cox
BP	Isobel Emslie
BP	Niall Webster
BP	Tom Baird
Capgemini Norway	Tetyana Kholodna
CDA	Daniel Brown
CDA	Jackie Clapp
CDA	Malcolm Fleming
CDA	Richard Salway
CDA	Sakthi Norton
CDA	Terry Alexander
Cegal	John Sayer
Centrica	David Sneddon
Centrica	Greig Henderson
Centrica	Rachel Harrold
CGG Data Management Services	Kerry Blinston
Chevron	Charles Cook
Company Connecting	Paul Lindop
ConocoPhillips	Andrew Reader
ConocoPhillips	Ashley Dunlop
Cray Inc.	Maria Mackey
DataCo Ltd	Christopher Frost
Decomm Data Ltd	Jane Hodson
E&P Consulting ltd	Ian Kennedy
E&P Consulting ltd	Nick Gibson
ECIM	Reidar Kalvig
Fairfield Energy Limited	Kathy Strachan
Flare Solutions Ltd.	David Camden
Halliburton - Landmark	David Seymour
Hampton Data Services Ltd.	Simon Fisher
Hampton Data Services Ltd.	Waclaw Jakubowicz
Independent Data Services	Colin Dawson
ITF	Craig O'Brien
KADME AS	Gianluca Monachese
Leidos	Jennie Morrison
Luchelan Ltd	Alan Smith
Maersk Oil	Christine Mckay
Moveout Data	Philip Wild
NDB	Ed Evans

Organisation	Delegate Name
Nexen CNOOC	Katherine Grundy
NPD	Elin Aabø Lorentzen
NPD	Eric Toogood
Offshore Engineer	Elaine Maslin
Oil & Gas Authority	Carlo Procaccini
Oil & Gas Authority	Claire Black
Oil & Gas Authority	Nick Richardson
Oil & Gas Authority	Simon James
Oil & Gas UK	Katy Heidenreich
Premier Oil	Jonathan Pye
Schlumberger	David Smith
Schlumberger	Gerry McNeill
Schlumberger	Mike Smith
Schlumberger	Paul Coles
ScotlandIS	Allan Sutherland
Scottish Enterprise	David Smith
Scottish Enterprise	Steven Harrison
Shell	Bob Harrison
Shell	David Pert
Shell	Elin Marie Nicolaisen
Shell	Helen McGlen
Shell	Ian Jackson
Shell	Katie Izat
Statoil	Eirik Time
The Data Lab	Duncan Hart
ThinkTank Maths Limited (Edinburgh)	Cyrille Mathis
ThinkTank Maths Limited (Edinburgh)	George Weatherill
Troika International Limited	Audrey Hughes
University of Aberdeen	Andrew Starkey
University of Aberdeen	George Coghill
University of Aberdeen	Joe Chapman

13 Appendix B: Speaker Biographies and Presentation Abstracts

Speaker Biographies
Ed Evans, Managing Director, NDB
With a background in Geology and IT systems analysis Ed has been working in E&P IT systems for more than 25 years, at BG Group and at Landmark/Halliburton. Ed is one of the founders of NDB. Since 2004, NDB has worked with Oil Company Subsurface and IT teams to improve subsurface capability and functional excellence. This means optimising organisation, workflow, data and software toolkits. from their technical systems and data environment. Ed has delivered strategic consulting, defining Digital Strategy, Digital Governance and how to implement successfully, to many of the worlds' leading Oil and Gas companies including BP, Shell and ConocoPhillips amongst many others. Ed focuses on the value of technical systems and data management to the business.
David Camden, Director, Flare Solutions
David has over 40 years' global experience in E&P. He started as field engineer in Schlumberger in Iraq then petrophysicist and petroleum engineering manager in BG Group. David co-founded Flare in 1998 with oil company colleagues to concentrate on E&P Information management. He has worked on many client consulting projects including authoring the original NPD Blue Book. His focus in Flare is on E&P Taxonomy development, strategic IM consulting and IM solutions delivery.
Waclaw (Wally) Jakubowicz, Managing Director, Hampton Data Services Ltd.
Wally has 37+ years in Upstream Oil & Gas. He started as Field Engineer with SLB, then worked as a Geologist, Geophysicist and Petrophysicist also with SLB and independently. Over the past 25 years, he has been implementing E&P IM/DM projects for several Supermajor Oil Cos, many large to small E&P Cos, NOCs as well as major service companies and consultancies. He holds a BSc Geology, MSc Geophysics, DIC, and is a member of SPE, SPWLA, PESGB, LPS.
Colin Dawson, Program Manager, Independent Data Services
Colin worked offshore for 2 years before university and quickly realised that it wasn't a life for him. His first degree was in Computer Science, and he has a background in software development for the oilfield, specialising in analytics, benchmarking & data science. Currently he is the Program Manager for the Anova product suite.
Håkon Snøtun, Project Manager Software, AGR Software AS
Håkon Snøtun is the project leader for AGRs iQx Software Suite, and has worked as a software architect and developer for more than 10 years. He has a Master of Science from the Norwegian University of Science and Technology, an MBA from the same university and was a Visiting Fellow at MIT's Sloan School.
Henri Blondelle, Co-Founder / VP Sales and Marketing, Agile Data Decisions LLC
Henri is a geologist by education. He started his professional career in the '80s at the early days of the workstation development with BP then with CGG-Petrosystems. He is the recent co-founder of Agile Data Decisions, a start-up company dedicated to Machine Learning application for Subsurface Data Management.

Speaker Biographies

Gianluca Monachese, Director Business Development, KADME AS

Gianluca Monachese is the Founder of KADME and its Director for Business Development. He has a Master in Geoscience from Italy and one in Computing for Geoscience awarded by the Nottingham Trent University and the British Geological Survey. He has 20 years of specific experience in Information Management in the oil industry and started KADME in 2002 with the vision of bringing modern technologies into the E&P information management business.

Maria Mackey, Energy Industry, Business Development EMEA & APAC, Cray Inc.

Maria has worked in upstream O&G as an independent seismic processing contractor and as a Geoscience Applications Consultant for Schlumberger. She has also worked as an O&G Systems Engineer for SGI, Sun Microsystems and Oracle and now promotes the advantages of Cray technologies in the Energy Sector.

Paul Coles, Business Development Manager, Schlumberger

Paul Coles is the Business Development Manager for Schlumberger Information Solutions and is responsible the global National Data Center Business. Paul acted as the Project Manager for the successful delivery of both the CDA Well and CDA Seismic DataStore Implementation Projects. Paul has previously held a number of Project Management and Services Delivery roles during 12 years with Schlumberger Information Solutions. Prior to joining Schlumberger Paul held Data Management positions with BP Exploration and Deminex Oil and Gas.

14 Appendix C: Data Provided to Participants

After signing a confidentiality agreement, each participant in the Data Challenge was provided a copy of the same data set, as detailed in the following table. Only released data was provided as a part of this exercise.

The data set was supplemented by an extract from the UKOilandGasData well and seismic headers (in a structured) format; and all data was provided in an hierarchical directory structure, as an aid to navigation and loading into participant data management systems.

Domain	Data Type	No. Items	Size [GB]
Wells [11,028]	REPORT IMAGE	132,612	670.9
	LOG IMAGE	222,940	1,233.1
	DWL FILE	97,263	1,167.7
	JWL FILE	2,508	7.7
	JWL AUDIT	313	0.3
	WDD FILE	7,857	0.5
	VSP FILE	5,538	51.7
	WELL DIGITAL SEISMIC	1,122	0.6
	WELL DIGITAL CORE	1,775	2.5
	WELL DIGITAL TEST	402	1.4
	TOTAL	*472,330*	*3,136.5*

Seismic - 2D [940]	Acquisition QC Report	245	1.5
	Acquisition Report	296	1.6
	Basemap	7	0.0
	Data Loading Form	117	0.0
	Data Loading QC Report	925	0.1
	Field Processing Report	1	0.0
	Field QC Output Listing	114	0.0
	Final Survey Report	29	1.4
	Navigation Report	26	1.6
	Observer Logs	335	6.5
	Operations Report	10	0.1
	Processing Report	184	5.6
	Reprocessing Report	10	0.3
	Survey Field Report	19	0.1
	Survey Notes	48	0.0
	Survey Shipment Report	2	0.0
	Survey Supervision Report	84	0.6
	Section Image	7	0.2
	Section Label	59	0.2
	Postplot Navigation	3,723	4.3
	Velocity	106	0.3
	TOTAL	*6,347*	*24.4*

	Acquisition Contracts and Correspondence	5	0.2

Domain	Data Type	No. Items	Size [GB]
Seismic - 3D/4D/OBS [330]	Acquisition QC Report	270	3.4
	Acquisition Report	249	2.5
	Data Licence and Trade Agreements	1	0.0
	Data Loading Form	37	0.0
	Data Loading QC Report	353	0.1
	Field Processing Report	2	0.0
	Field QC Output Listing	35	0.3
	Final Survey Report	19	0.1
	Nav Performance Report	15	0.2
	Navigation Report	155	1.5
	Observer Logs	150	7.4
	Operations Report	13	0.1
	Processing Report	288	5.9
	Reprocessing Report	6	0.1
	Survey Field Report	4	0.0
	Survey Notes	6	0.0
	Survey Positioning Review	7	0.1
	Survey Shipment Report	1	0.0
	Survey Supervision Report	35	0.5
	Section Image	9	0.1
	Section Label	1	0.0
	Postplot Navigation	501	247.0
	Velocity	122	90.3
	TOTAL	*2,284*	*359.8*

Seismic - Site [4]	AG_Section Image	1	0.0
	DG_Acquisition Report	30	1.2
	DG_Data Loading QC Report	4	0.0
	DG_Field QC Output Listing	6	0.0
	DG_Observer Logs	2	0.0
	DG_Survey Supervision Report	2	0.0
	DG_Section Image	70	0.2
	Postplot Navigation	5	0.0
	TOTAL	*120*	*1.4*