Towards digital companions for earth modelers: combining data science and natural language processing to facilitate geomodelling workflows

Antoine Bouziat[1], François Cokelaer[1], Renaud Divies[1], Sylvain Desroziers[1], Mathieu Feraille[1] and Jeanneth Bouziat

[1] *IFP Energies nouvelles, 1-4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France*

September 2020

Abstract

With the commercial success of Amazon Alexa or Apple Siri, the term *digital companion* has been popularized beyond scientific and technological circles to reach the general public. It usually refers to devices designed to assist their owners in various routine tasks, based on machine learning and natural language processing (NLP) approaches. In this study, we explore the extension of this concept to the geomodelling domain. We appraise the potential of recent data science and NLP technologies to accelerate and democratize several steps of the geomodelling routine, from mining literature for parameters to browsing physical simulation results. We notably illustrate their value focusing on basin modelling in a petroleum exploration context.

Firstly, we assess text entity extraction techniques to automatically collect information about hydrocarbon source rocks in scientific papers. We build a dedicated ontology representing the conceptual relationships between source rock features, and we manually annotate 127 papers accordingly. Then we train an industrial deep learning model to autonomously retrieve similar features from a larger bibliographic corpus. The pieces of information extracted and their relationships are stored in a specific graph data base. Eventually, to ease the graph exploration, we build an interpreter of natural language queries and interface it in a web application. As a result, operational geoscientists can select modelling hypotheses and discuss simulation results from a wide literature in an efficient and intuitive fashion.

Secondly, we evaluate innovative ways to facilitate the analysis of physical simulation results. We start with linking a 4D basin model with interactive data visualization dashboards specifically tailored to highlight the maturity evolution of the source rocks through their geological history. Then we further democratize the process by training a full conversational engine to interpret natural language queries, to browse the simulation results for an answer and to provide a relevant data visualization. The resulting tool relies on an industrial and fully interfaced cloud-based platform. At the end, users from diverse backgrounds can interact with the geomodel as fluently as they would chat with their friends or colleagues.

We consider these examples pave the way for fully integrated earth-modelling companions, which could in the future assist many professionals in their geomodelling work.

Introduction

Artificial intelligence technologies are currently raising much interest, well beyond traditional scientific circles. Notably, the promotion of Apple Siri or Amazon Alexa publicized the concept of *digital companions*. It corresponds to devices relying on deep learning and natural language processing (NLP) methods, aiming to support their owners in several routine jobs. In this study, we investigate the extension of the notion to geomodelling activities, from two practical use cases in a basin modelling context.

1. Assisted text mining for geological information

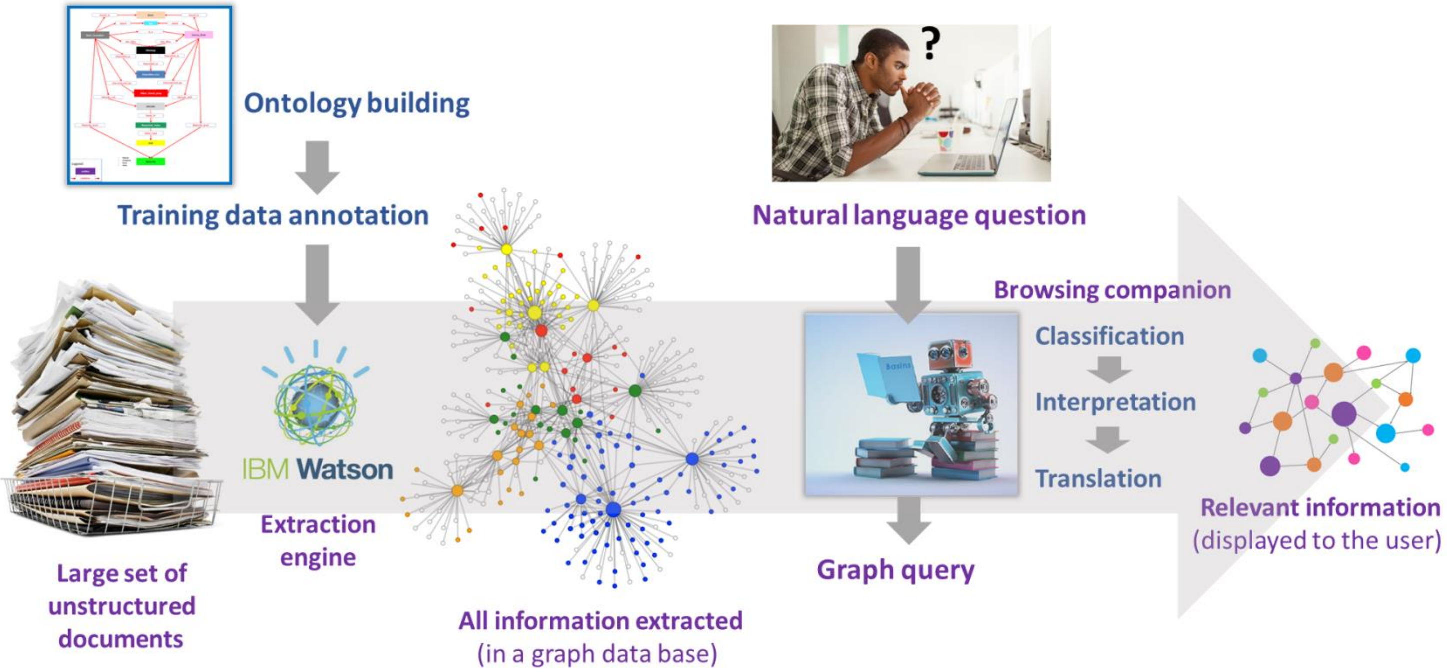
Our first use-case is an assisted text mining dedicated to geoscientific information. The expression *text mining* usually describes the ability of computer programs to retrieve meaningful technical information from a corpus of textual documents. This ability could be of great help for geomodelers, either to summarize the published knowledge on a specific area, to initialize the parameters of physical simulations or to compare their results with the literature. However, a specific effort in adapting and training the algorithms is required to handle geological concepts and vocabulary.

1.1. Demo project: source rocks characterization

In this demonstration project, we assessed the capability of automated text mining technologies for source rock characterization in a context of petroleum system analysis. The first objective was to automatically extract information on source rocks, such as their lithology, age, deposition environment, organic content or maturity, from academic articles. The second objective was to provide this information to geologists in a format which can be fluently browsed in an operational context driven by geological questions. Parts of this work were already presented in recent conferences (Guichet *et al.* 2019, Bouziat *et al.* 2020).

1.2. General workflow

For this use-case, our digital companion approach relies on a general workflow, illustrated on Figure 1.



Ontology building
Training data annotation
Large set of unstructured documents
IBM Watson
Extraction engine
All information extracted (in a graph data base)
Natural language question
Browsing companion
Classification
Interpretation
Translation
Graph query
Relevant information (displayed to the user)

Figure 1: General workflow for assisted mining of geological information in unstructured documents

Ontology building – Our approach is supervised and the first step of the workflow is to build a dedicated ontology. This ontology models the main features of the geoscientific language used to describe source rocks in our documents, in terms of concepts and relationships between concepts. The ontology building is a key step as it embeds the geological knowledge input into the text mining system.

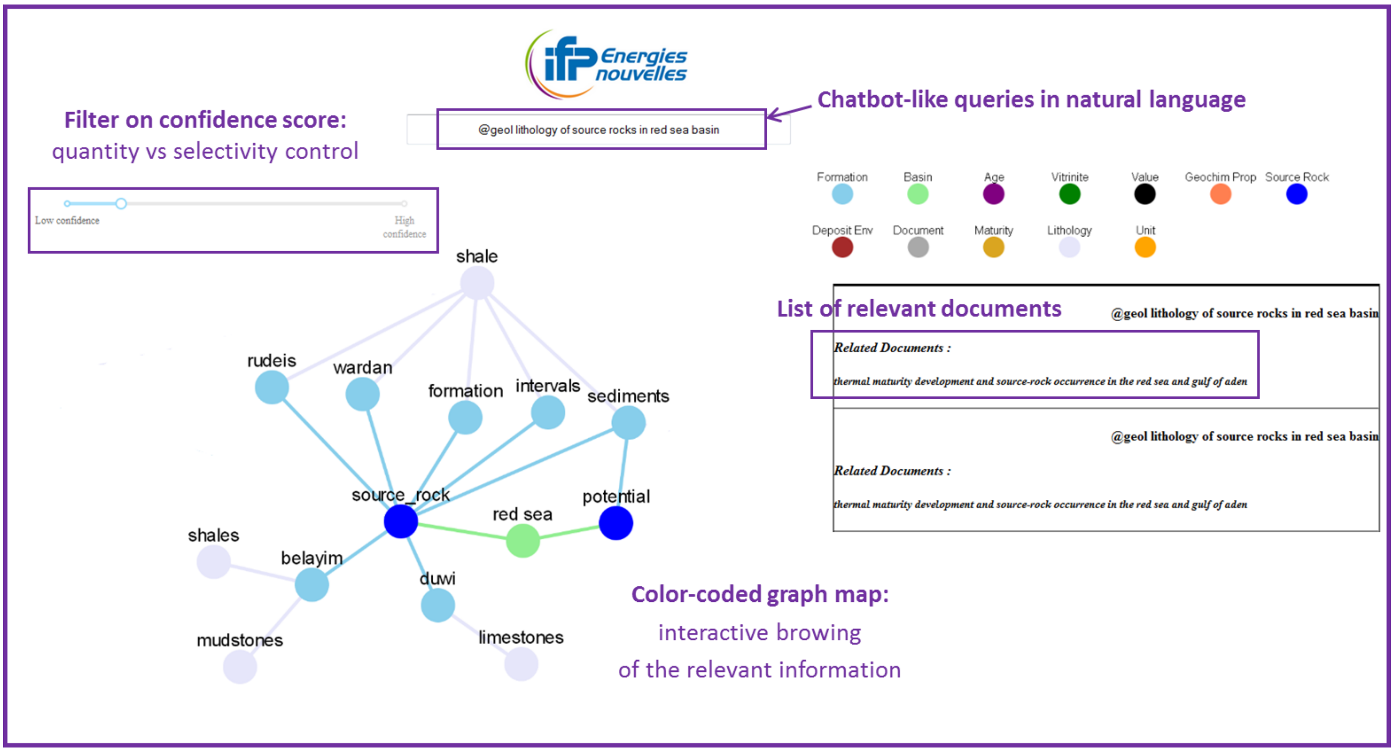
Training documents annotation – Once the ontology is built, training data is generated by annotating a representative set of textual documents. This annotation phase corresponds to mapping the conceptual ontology model to actual written sentences. The quality of this annotation will directly reflect on the quality of the future information extraction. It should then be closely monitored by subject matter experts, in our case petroleum geologists. 127 training documents were annotated in our demo project.

Information extraction – The annotated data can now be used to train an extraction model. In this project we used a commercial model : IBM Watson (High 2012). Once trained, the extraction model is applied on new documents. In this project, a few hundreds of articles were processed.

Graph data-base storage – The previous step usually generates a large collection of information pieces extracted from the documents, which can be hard to browse manually. We propose to store them in a dedicated graph data-base, built following the conceptual ontology. This will facilitate further selection of the pieces relevant to a specific geological question, and their display in a geologically meaningful format.

Natural language translation – At this stage of the workflow, the objective is to assist geoscientists in efficiently exploring the knowledge data base extracted from the documents. We propose to train a deep learning model to translate natural language questions into numerical graph queries. These queries will retrieve from the whole graph the specific data relevant to answer the input questions. To optimize the quality of this translation, we recommend to use a goal-oriented chatbot strategy (Ilievski *et al.* 2018). In practice, the questions are first classified into several pre-defined families, then more precise word analyses are carried out.

Information display – Finally, for each geological question, the relevant part of the knowledge graph can be displayed as an answer. We propose to manage this display through a dedicated web application, designed to maximize interactivity in the exploration of the text mining results by the operational geoscientists (Section 1.3).



Filter on confidence score:
quantity vs selectivity control

Chatbot-like queries in natural language

@geol lithology of source rocks in red sea basin

List of relevant documents

@geol lithology of source rocks in red sea basin

Related Documents :

thermal maturity development and source-rock occurrence in the red sea and gulf of aden

@geol lithology of source rocks in red sea basin

Related Documents :

thermal maturity development and source-rock occurrence in the red sea and gulf of aden

Color-coded graph map:
interactive browing
of the relevant information

Figure 2: Interface of a web application designed to browse the text mining results

1.3. Web application

A web application interface was specially designed and developed to assist geoscientists in the exploration of the text mining results (Figure 2). From an input question in natural language, a list of relevant documents in the corpus is given to the user, as well as the information extracted in these documents answering the question. This information is displayed as a color-coded graph following the ontological model used for the extraction. Finally, a confidence filter is provided to control the quantity and selectivity of the information displayed. This confidence filter relies on a confidence score associated to each piece of information extracted by the text mining engine, this score reflecting whether the source sentence was statistically similar to the training data.

1.4. Examples

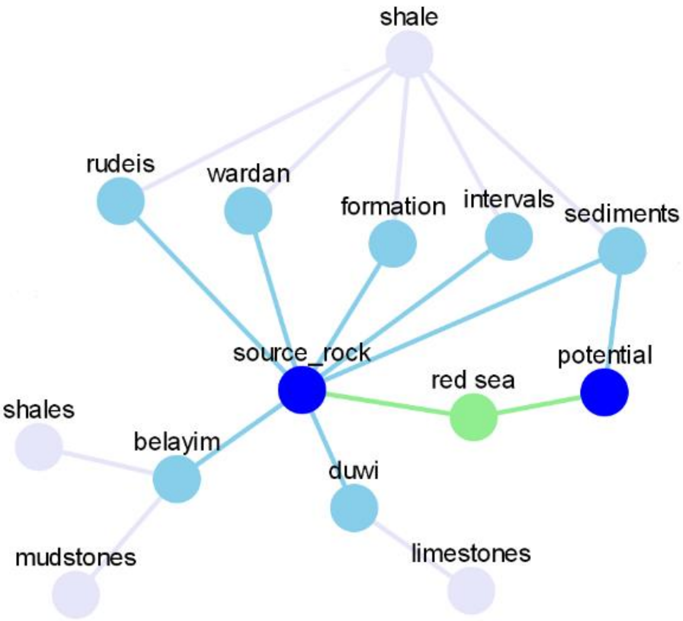


*Figure 3: Information automatically extracted and selected as relevant to the question “*lithology of source rocks in the red sea basin?”

In this section, we illustrate the results of the demonstration project with two examples of geological questions.

Figure 3 shows the piece of graph answering the question “*lithology of source rocks in the red sea basin?*”. With the color code derived from our conceptual source rock ontology, the green node corresponds to basin information. We can then see that the question was correctly understood by the automated translator, as the words “*red sea*” were retrieved. The dark blue nodes correspond to words designating source rocks, which were extracted in sentences about the red sea basin. Obviously the words “*source rocks*” were retrieved. More interestingly, the word “*potential*” was also highlighted. It is true petroleum geologists can use this word in expressions as “*oil potential*” or “*gas potential*” to refer to source rocks. It is worth noting this was not explicitly input into the extraction engine, but only statistically determined from the ontology, the training data and the sentence structure. The light blue nodes correspond to names of geological formations, which were associated to the source rock synonyms and the red sea basin in the documents processed. The remarkable thing here is that the system extracted very generic words like “*formation*”, “*intervals*” or “*sediments*” along very specific ones like “*duwi*”, “*wardan*”, “*rudeis*” or “*belayim*”. The latter words are actual local names of source rock formations in the red sea basin (Barnard *et al.* 1992). Last, the grey nodes correspond to lithological information about these formations. They hold the answer to the initial question. We learn that, while Rudeis and Wardan formations are only mentioned as shales, the Duwi one is reported as calcareous. At this stage, the user can integrate this information into his basin model, read the source document, or complete the corpus to process more articles.

Figure 4 shows the piece of graph answering the question "*vitrinite reflectance in the Paris basin for the Jurassic?*". As in the previous example, the green node corresponds to the target basin, here the Paris one. The light blue node corresponds to a geological formation, while the purple one corresponds to an age. The results here underline the interest of using an approach based on a conceptual ontology, rather than a more explicit dictionary-based one. Indeed, the more conceptual approach provides enough flexibility to have the word "*lias*" recognized as a geological formation when the word "*jurassic*" is recognized as an age. As it is common for geologists to refer to formations by their age, the classification could be the opposite in a sentence built differently. The dark green node corresponds to a vitrinite information, and finally the dark nodes correspond to numerical values. Here, one can see the Jurassic-aged source rocks in the Paris basin described in the documentary corpus are mature enough to have generated hydrocarbons, but that significant lateral variations are reported. This kind of literature mining could then be used for calibrating basin simulations. More generally, this example highlights the capacity of the proposed workflow to extract and contextualize numerical values, facilitating initiation of geomodel parameters and discussion of simulation results based on scientific literature.

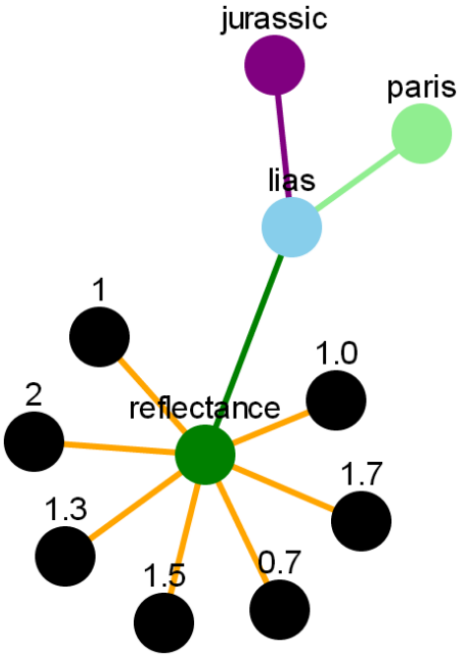


Figure 4: Information automatically extracted and selected as relevant to the question "vitrinite reflectance in the Paris basin for the Jurassic?"

2. Assisted analysis of numerical simulation results

Browsing and analyzing results of physical simulations run on geological models can be very tedious and involve complex post-processing. Typically, a geomodeler will snapshot some displays and compute a few graphs to be pasted in a slides show. Presenting the slides in a meeting, with colleagues, clients or partners, some questions will arise, necessitating new displays and graphs to be answered. But, without the model directly accessible, the geomodeler will need to come back to his desk, post-process the results again and set-up a new meeting to provide and discuss answers. Our feeling is that the whole process could be accelerated and optimized thanks to interactive analytic tools that could be accessed remotely in meetings or business trips. Additionally, such digital companions could support geomodel exploitation in fast decision-making contexts, while democratizing it for professionals without a strong geomodelling background.

2.1. Demo project: 4D basin model

In this demonstration project, we assessed the capability of two different data analysis tools to facilitate the exploration of numerical simulation results. In both cases, we used the same 4D synthetic basin model as test data. This model presents about 100 000 cells at present-day, 19 geological ages and 3 source rock layers. It can be considered as representative of classical basin models in petroleum exploration studies, as it was previously used to create tutorials of commercial basin modelling software.

2.2. First approach: interactive visualization dashboards

Our first approach is inspired from the *business intelligence* community and relies on interactive visualization dashboards. With such approach, the simulation results are first exported as large column-based ASCII files, so they can be accessed outside geomodelling software. Then they are connected to cloud-based interactive visualization dashboards (Figure 5). These dashboards consists in inter-connected data analysis widgets offering a powerful graphical overview of the main statistical patterns and multiple cross-filtering possibilities. Key scalar indicators can also be computed and monitored (Eckerson 2010).

In this demonstration project, a source rock maturation dashboard was built using Microsoft PowerBI platform (Ferrari and Russo 2016). It provides a graphical visualization of the volume of each maturity window in the three source rocks, as well as the spatial and statistical distributions of the vitrinite reflectance proxy simulated. These displays can be updated on-the-fly for each of the 19 geological ages modeled. Feedbacks from users highlight the time saved in post-processing tasks and the interactivity gained in business presentations in comparison with traditional slides shows.

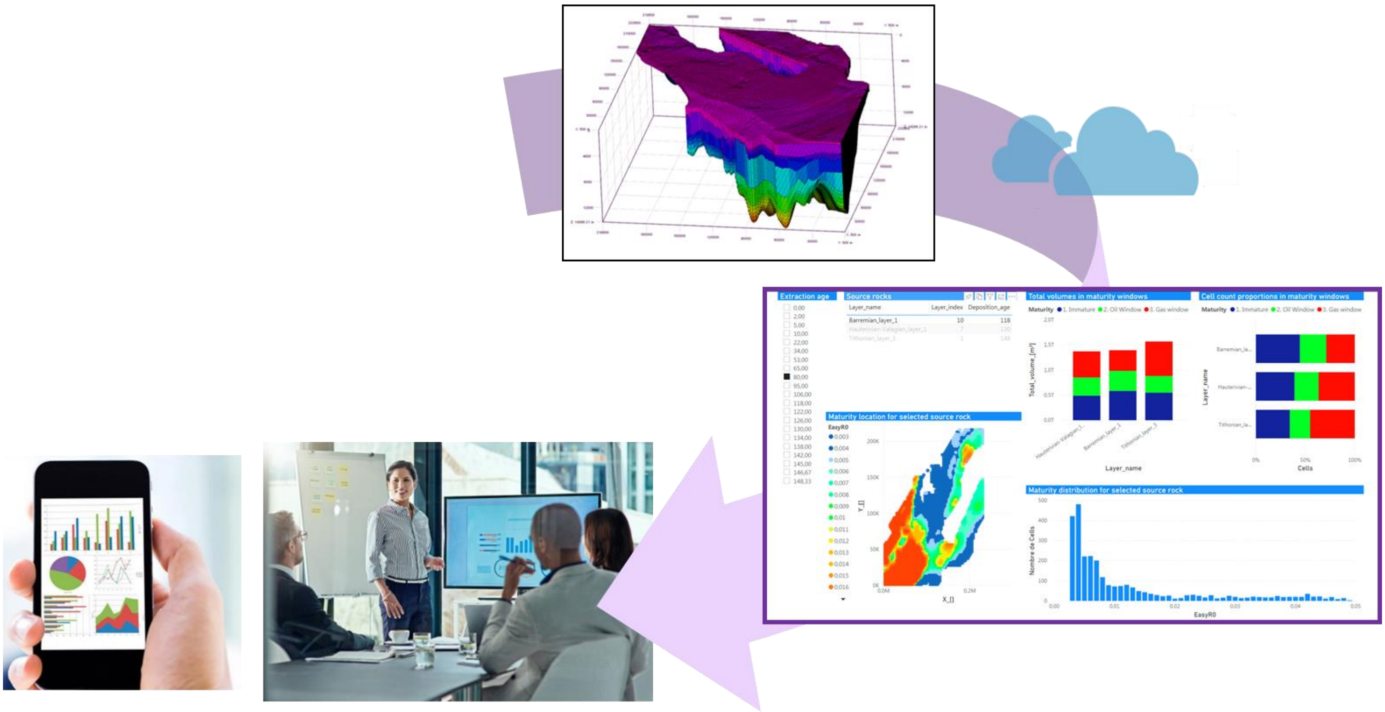
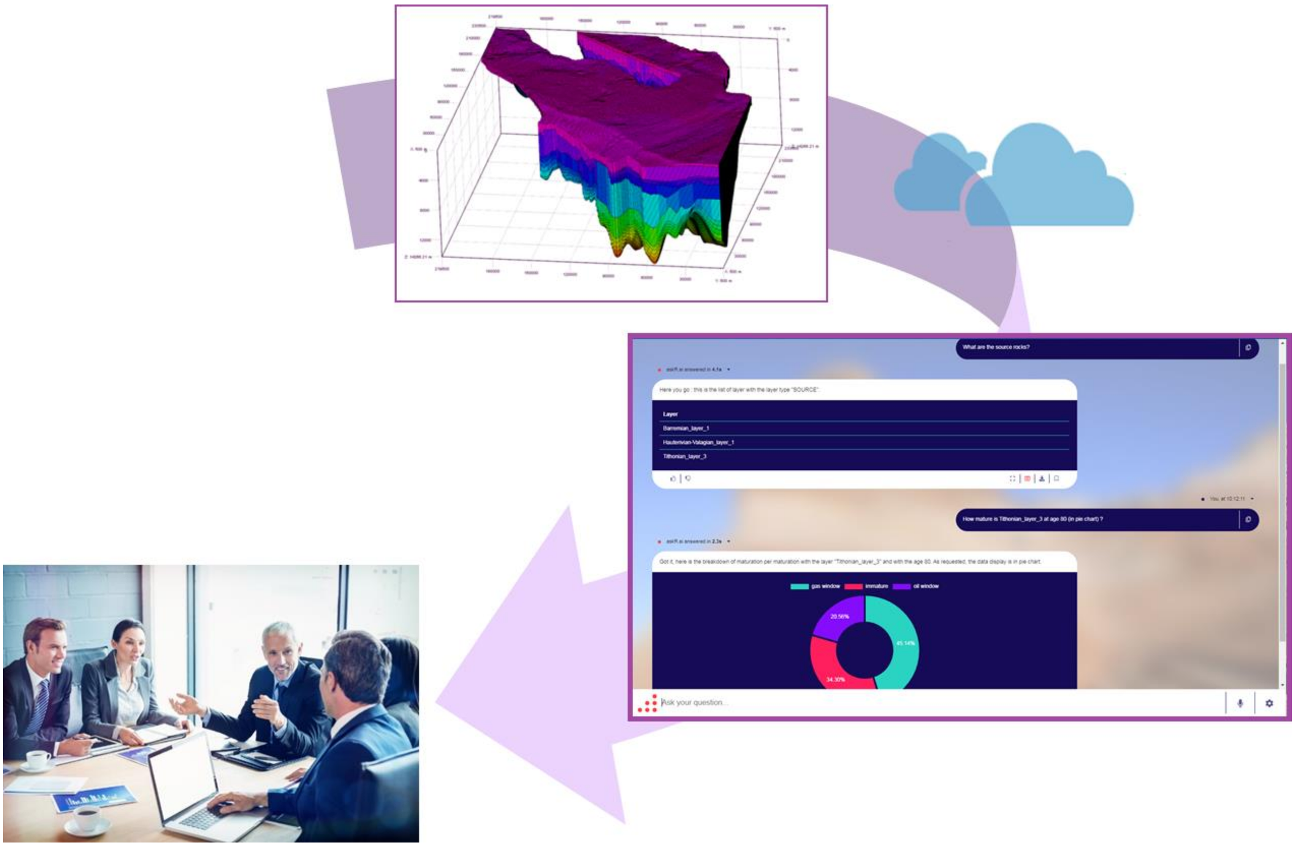

Extraction age
Source rocks
Layer_name
Layer_index
Deposition_age
Barremian_layer_1
Hauterivian-Valagian_layer_1
Tithonian_layer_1
Total volumes in maturity windows
Maturity ● 1. Immature ● 2. Oil Window ● 3. Gas window
Total_volume_[m³]
Layer_name
Cell count proportions in maturity windows
Cells
Maturity location for selected source rock
EasyR0
Y_[]
X_[]
Maturity distribution for selected source rock
Nombre de Cells
EasyR0

Figure 5: Remote interactive analysis of geomodelling results with a cloud-based dashboard (conceptual workflow)

2.3. Second approach: conversational engine

A limitation of the dashboarding approach is that displays must be defined beforehand. The various plots are inter-connected and multiple data filters can be applied on-the-fly to update them, but it is not immediate to add a new widget to an existing dashboard. To overcome this limitation, we assessed completing the dashboards with a conversational engine, able to translate natural language questions into data queries and provide new plots interactively (Figure 6). To do so, we relied on the commercial ASKR technology (https://en.askr.ai), into which we integrated geoscientific concepts and vocabulary. The resulting tool is a fully web-based platform which can be connected to simulation results exported as a column-based ASCII file. The users can type questions formulated in natural language as in classical messaging applications. They are translated by a sophisticated NLP model into numerical queries, applied on the simulation results data. Finally, the answer is returned to the user as a scalar value or a new visualization plot. Previous questions are recorded to integrate contextual information in the NLP analysis, leading to dialog-like interactions.



What are the source rocks?

askR.ai answered in 4.1s

Here you go : this is the list of layer with the layer type "SOURCE".

Layer
Barremian_layer_1
Hauterivian-Valagian_layer_1
Tithonian_layer_3

You at 10:12:11

How mature is Tithonian_layer_3 at age 80 (in pie chart) ?

askR.ai answered in 2.3s

Got it, here is the breakdown of maturation per maturation with the layer "Tithonian_layer_3" and with the age 80. As requested, the data display is in pie chart.

Figure 6: Remote interactive analysis of geomodelling results with a cloud-based conversational engine (conceptual workflow)

This conversational approach was applied on our 4D test basin model. The range of questions successfully processed includes:

- *How mature was Tithonian source rock 95My ago?*
- *Average burial of oil window in Barremian through time?*
- *Location of highest overpressure in Eocene layer?*

Answer times between 1 and 10 seconds are observed. Feedbacks from users highlight the flexibility offered by the tool and its complementarity with the preset dashboards. The approach could be valuably extended to assist users in comparing different versions of a same geomodel, or in appraising the quality of a geomodel calibration to well data.

Conclusions

The use cases presented in this paper underline how geomodelling activities can concretely benefit from data science and NLP technologies. They display what we could name a "3A vision" where artificial intelligence is preferentially used to Assist humans in difficult tasks, Accelerate tedious processes and partly Automate repetitive actions. The use cases also pave the way for fully integrated geoscientific companions, which could in the future support many geoscientists in their daily work. However they also underline that dedicated innovation will be needed to embed as much geoscience knowledge as possible in the algorithms and adapt digital technologies to the geoscience context.

Code availability

The prototypes presented in this paper can be shared through the TELLUS community, a research-and-innovation partnership framework dedicated to the digital transformation of geoscientific activities. For more information, please contact the lead author at antoine.bouziat@ifpen.fr.

References

Barnard, P. C., Thompson, S., Bastow, M. A., Ducreux, C., & Mathurin, G. (1992). Thermal maturity development and source-rock occurrence in the Red Sea and Gulf of Aden. *Journal of Petroleum Geology*, 15, 173-186. https://doi.org/10.1111/j.1747-5457.1992.tb00961.x

Bouziat, A., Desroziers, S., Feraille, M., Lecomte, J., Divies, R., & Cokelaer, F. (2020). Deep Learning Applications to Unstructured Geological Data: From Rock Images Characterization to Scientific Literature Mining. In *First EAGE Digitalization Conference and Exhibition* (Vol. 2020, No. 1, pp. 1-5). European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202032047

Eckerson, W. W. (2010). *Performance dashboards: measuring, monitoring, and managing your business.* John Wiley & Sons. ISBN: 978-0470589830

Ferrari, A., & Russo, M. (2016). *Introducing Microsoft Power BI.* Microsoft Press. ISBN: 9781509302758

Guichet, X., Dubos-Sallée, N., Cacas-Stentz, M. C., Rahon, D., & Martinez, V. (2019). Efficient Access to Relevant Knowledge Extracted From Geoscience Literature Dedicated to Petroleum Basin Exploration by Using IBM Watson. In *2019 AAPG Annual Convention and Exhibition.* http://www.searchanddiscovery.com/abstracts/html/2019/ace2019/abstracts/1704.html

High, R. (2012). *The era of cognitive systems: An inside look at IBM Watson and how it works*. IBM Corporation, Redbooks. IBM Form: REDP-4955-00

Ilievski, V., Musat, C., Hossmann, A., & Baeriswyl, M. (2018). Goal-oriented chatbot dialog management bootstrapping with transfer learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 4115-4121). https://www.ijcai.org/Proceedings/2018/0572.pdf