CCUS: 4186419

Augmented Data Management for Subsurface CCUS Data Sets

Rhys Blake[2], Jess B. Kozman*[1], James Lamb[1], Lorena Pelegrin[3], 1. Katalyst Data Management – U.S., 2. Katalyst Data Management – U.K., 3. Iron Mountain – U.S

Abstract

Expectations are high for generative AI and machine learning workflows increasing potential profitability of CCUS enterprises (Yao et al., 2023). Advanced data management technologies are enabling the re-use of subsurface technical knowledge contained in unstructured files from decades of hydrocarbon exploration. Documenting the financial benefits from these new workflows requires data technology readiness across the CCUS data lifecycle. We are building and expanding on existing data workflows for using artificial and convolutional neural networks to find information in legacy documents that can predict physical properties of potential CO2 injection reservoirs. Similar information from unstructured reports and files can be used during the appraisal stage of CCUS projects to support evaluation of the mechanical stability of geologic trapping mechanisms. These data workflows and techniques can also be applied during the measurement, monitoring and verification (MMV) stage to manage the data for tracking CO2 plume migration or leakage during injection cycles. However, challenges remain in finding and accessing the most relevant unstructured data, and misadventures and misapplication of these technologies in identifying critical data for business decisions can lead to delays in project execution.

Introduction

Existing usage of generative AI for unguided applications of text extraction and automated metadata population can lead to improper indexing of documents such as wellbore diagrams, which can be critical to understanding potential leakage risks during CO2 injection (DiGiulio, 2024). In one example for documents available from a U.S. state regulator, a natural language query on a non-curated dataset led to a result indicating the availability of pressure data from a well submitted before the first oil discovery in the state. Further investigation showed that the optical character recognition (OCR) part of the workflow

had interpreted a poor-quality document image incorrectly (**Figure 1**). We demonstrate how AI enhanced and augmented applications that keep a human subject matter expert in the loop can improve the efficiency and usability of data through preparation, collection, and indexing of critical metadata. We then show the value of a natural language processing interface for question-and-answer style knowledge access.

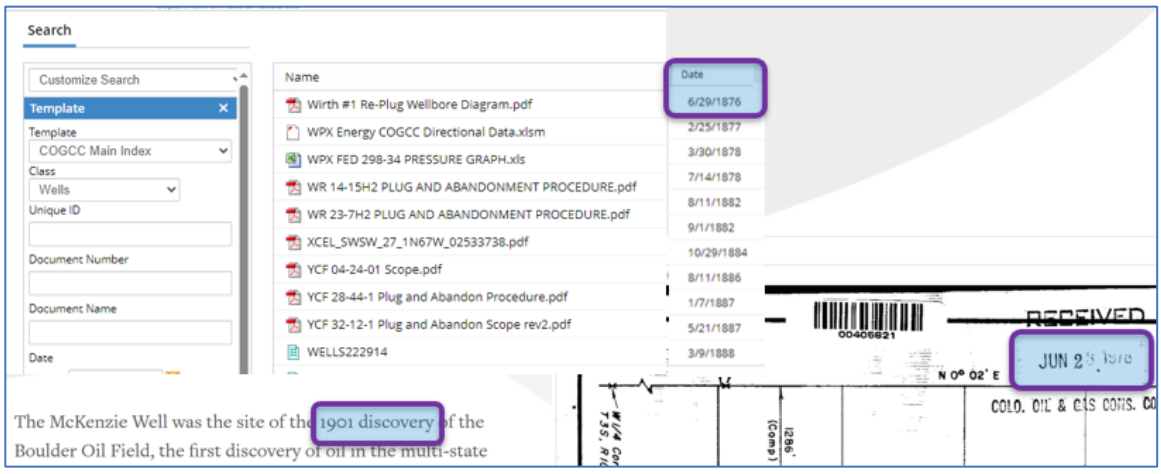

Figure 1. Example of legacy document data with incorrect dates from un-curated OCR application.

Methods

We show that a generative AI enabled query tool using an agent with contextual intelligence can reduce time spent locating a specific item of information in a large corpus of unstructured data. We use the example of a geotechnical end user wanting to find specific porosity data from a public domain and open-file document collection comprising documents associated with over 100 seismic surveys and 300 wells (**Figure 1**). In this example we show how traditional search and filter options can locate a group of selected wells in less than 50 milliseconds running the search as a hosted service. However, even with indexed data in an industry standard database structure, using elastic and/or Boolean search, a user would face the daunting task of then potentially opening multiple 50-page well completion documents to find the requested results. The augmented subsurface discovery tool interprets a natural language query related to the availability of porosity data in multiple documents, its relation to existing production, the description of the reservoir, and the depths of formation evaluations. It then points the user to a specific fractional porosity entry on a table on page 3 of a 30-page report for a specific wellbore. Our tests showed that a natural language query tool directed users to usable information in a few seconds, for searches that otherwise would have taken a knowledgeable technical end user of the data weeks. Comparative examples show this time savings only when accessing a globally standardized large volume data store curated and indexed with industry standard processes. We found that the data also needed multiple passes of curation, indexing, and quality control before the subsurface discovery tool was able to produce consistent and reproducible responses. Having the ability to rapidly and efficiently identify the original source of this type of information will more effectively support other machine learning workflows for site screening of potential reservoirs for geological storage of CO2 (Wang, 2024). With some of these methods being used to create automated composite scores and cluster analyses combining factors such as location, capacity, and injectivity potential (Leng et al., 2023), it will be critical to have a documented audit trail of the information's provenance and lineage.

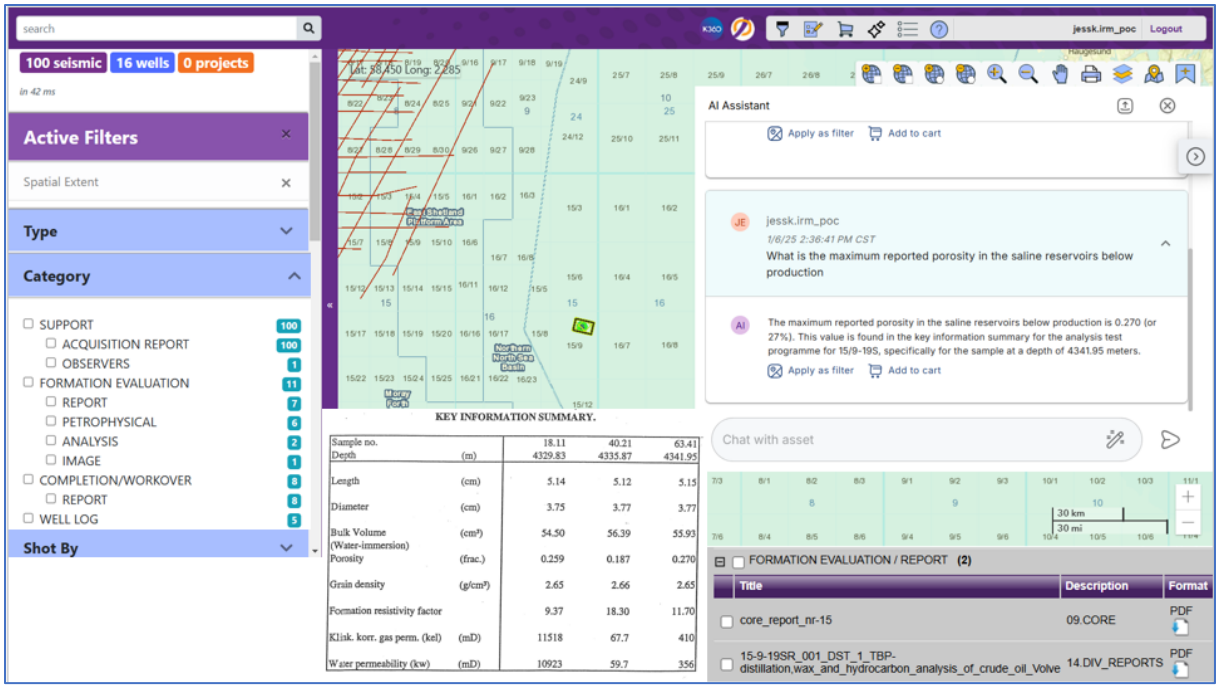

Figure 2. Map based view the dataset and a natural language contextualized Q&A query response pointing to key data in a multi-page report.

Use of a Large Language Model (LLM), Retrieval Augmented Generation (RAG) architecture, the maintained metadata, and knowledge graphs avoids dataset bias pitfalls and other misuses of machine learning on non-curated datasets. In the integrated AI/ML processing pipeline a generative AI algorithm is used to improve traditional machine learning classification routines. Where metadata extraction is not satisfactory or classifications into industry standard taxonomies are still ambiguous, an Intelligent Document Processing step is added the more quickly configures the LLM without the need for large volume labeled datasets. This also improves the quality and repeatability of the outputs, especially when dealing with multiple formats as found in many historical subsurface data sets.

Results

Machine learning techniques also improve the initial indexing of documents into an industry standard database with consistent taxonomies and reference lists (Gallant et al., 2023). This workflow was originally applied to raster log images and seismic survey headers to increase the efficiency of semi-automated population of standardized metadata for map-based queries. We then applied the same workflow by creating label-value pairs for a client selected set of up to 21 client selected priority metadata attributes associated with airborne geophysics data (**Figure 3**), which can assist with basin scale evaluation of reservoirs for geologic storage. For this workflow, an automated workflow was used that delivered high value candidate label value pairs to a human indexer, improving their efficiency by up to two times. The manifest payloads of standardized metadata were then used to ingest the documents into a cloud native, industry standards-based, and technology agnostic data platform for use by analytics and interpretation application, including further machine learning workflows. In the case of the thousands of documents in this study, manual metadata extraction times of months were reduced to days by the application of the augmented intelligence workflow.

Discussion

Q&A query results were observed from searches requested on quality-controlled digital subsurface data files used as a major input to interpretation and modeling workflows. This data included legacy well completion, logging and core analysis reports, seismic processing parameters and results, and other linked, indexed and classified metadata and schemas. One key lesson learned is that rigorous data governance during ingestion and indexing is a required success factor for generating relevant access results. Recent working sessions with applied geoscience users conclude that AI assisted agents can perform searches that previously took weeks in minutes. Topics of our test searches have been validated as well through discussion of recent use cases for ML/AI in CCUS projects by existing users of curated proprietary datasets (Barlow and Shahi, 2024). These case studies also demonstrate that understanding the value of having information to support decisions during evaluation and monitoring can improve overall management of the subsurface resource (Dias, 2024). Quantitative measures of the potential contribution of information to future decision-making processes can be balanced against the cost-effectiveness of acquiring that information from legacy data sources. Work is ongoing with geotechnical end users of subsurface data to use these new types of comprehensive uncertainty analysis which consider the limitations of imperfect or incomplete data like that obtained from historical well or seismic data sets. Ultimately the concept of penalization through data decision latency will allow a rigorous examination of the time to a positive return on investment for applications of generative AI in advanced data management strategies.

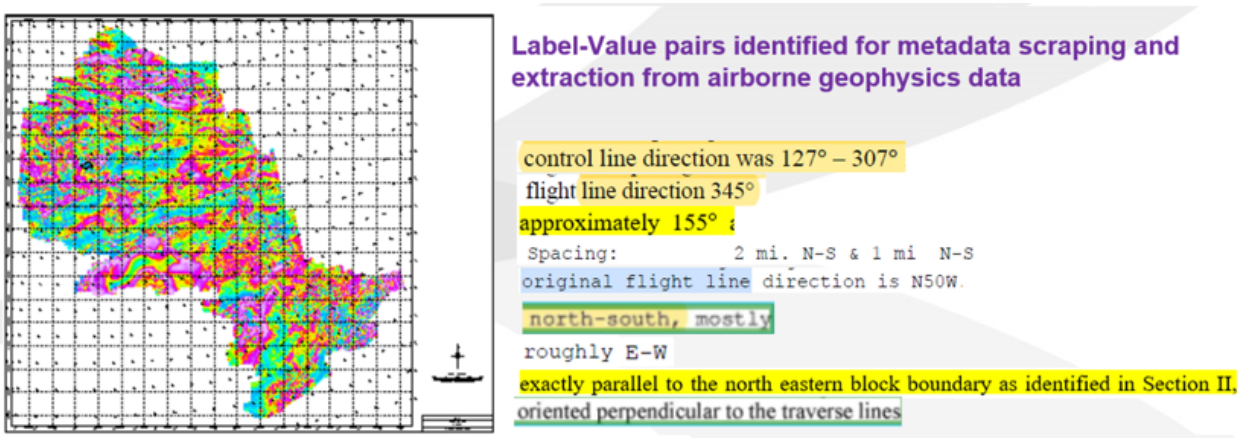

Figure 3. Example of the variations in how a single metadata attribute such as “flight line direction” can be represented in unstructured report documents, making standard keyword searches impractical for large datasets.

Conclusions

The use of advanced generative AI and natural language processing concepts can contribute to a quantifiable and measurable reduction in time spent accessing legacy data as a major input to the evaluation of the subsurface as a multi-resource asset for CCUS enterprises. We show that data decision latency times can be reduced by factors up to 84% by reducing the need for users to open multiple documents in sequentially compressed files to find a single keyword or value required to support business decisions. Calculations of a positive return on investment are based on enhanced access capability for candidate seismic volumes for advanced processing and well data supporting CO2 injectivity. We believe these advanced data management and metadata extraction methodologies will continue to streamline the business processes that support more efficient evaluation and development of potential subsurface reservoirs for carbon storage.

References

Barlow, H. and Shahi, S.S., 2024. "Technical Report - State of the Art: CCS Technologies 2024", Global CCS Institute, Washington, D.C. https://www.globalccsinstitute.com/wp-content/uploads/2024/08/Report-CCS-Technologies-Compendium-2024-1.pdf

Dias, R., 2024. "Improving Resource Management through a Value of Information (VoI) Methodology to Support Seismic Monitoring Decisions", Society of Exploration Geophysics, *4D Forum, Insights to Actions – A Global Forum*, November 2024. https://seg.org/wp-content/uploads/2024/09/4D-Forum-Technical-Program.pdf

DiGiulio, D., 2024. "Understanding, Evaluating, and Remediating Leakage from Abandoned Oil and Gas Wells During Geological Storage of Carbon Dioxide", Center for Applied Environmental Science, Environmental Integrity Project, https://environmentalintegrity.org/wp-content/uploads/2024/03/20240318_DiGiulio_report_Final.pdf

Gallant, S., Patel, N. and Zaheri, S., 2023. "Artificial Intelligence and Machine Learning in Sustainable Energy", 8th International Congress of the Brazilian Geophysical Society, Rio de Janeiro, October 2023, SBGf - Sociedade Brasileira de Geofísica, https://sbgf.org.br/mysbgf/eventos/expanded_abstracts/18th_CISBGf/7647966b7343c29048673252e490f736SBGF%20Abstract%20-%20KDM%20-%20AI-ML.pdf

Leng, J, Wang, H. and Hosseini, S., 2023. "A Data Analytics and Machine Learning Study on Site Screening of CO2 Geological Storage in Depleted Oil and Gas Reservoirs in the Gulf of Mexico", Society of Petroleum Engineers, Annual Technical Conference and Exhibition, DOI: 10.2118/214866-MS. https://www.researchgate.net/publication/374560631

Wang, K., 2024. "Geological Carbon Storage Site Screening Using Machine Learning", American Association of Petroleum Geologists, Geological Carbon Storage Site Screening Using Machine Learning Workshop, https://www.aapg.org/career/training/in-person/workshops/workshop-details/articleid/67704/generative-ai-machine-learning-and-analytics-for-subsurface-energy?srsltid=AfmBOor3yVxxO182b4c_oofqnF3aF6E4ag7-zm8Hucanb3X4X4Vr9zFj#program

Yao, P., Yu, Z., Zhang, Y. and Xu, T., 2023. "Application of machine learning in carbon capture and storage: An in-depth insight from the perspective of geoscience", *Fuel* 333(12):126296, DOI: 10.1016/j.fuel.2022.126296. https://ui.adsabs.harvard.edu/abs/2023Fuel..33326296Y/abstract