geoconvention
Calgary • Canada • May 12-14 2025

# Realistic Use of Generative AI for Subsurface Data Queries

*Lorena Pelegrin, Iron Mountain*
*Jess Kozman, Katalyst Data Management*

## Summary

There have been unrealistic expectations for the ability of modern technologies such as generative AI to automatically assist with location of business-critical data for subsurface resource projects. Initial tests with commercially available tools quickly revealed that extensive data quality management and standardization, rigorous indexing, and collection of curated metadata attributes are all critical to success. The rapidly expanding availability of newly digitized legacy subsurface data combined with rapid uptake of technologies for continuous data acquisition are putting pressure on workflows originally designed to support one time usage of subsurface datasets in a project-centric environment.

## Method and Data Workflow

We began with the usage of machine learning and convolutional neural networks to assist with the extraction of metadata from legacy subsurface documents such as raster image well logs, seismic survey headers, and airborne geophysics acquisition reports, supporting both oil and gas and mining initiatives (Gallant et al., 2023). This provided experience with the requirements for standardization and the use of label-value pairs to assist in an Augmented Intelligence workflow that increased the efficiency of human indexers processing thousands of documents for ingestion into industry standard databases. Those data stores were then accessed through map-based filtering and selection tools and more flexible elastic search keyword location for indexed documents. These workflows could however still require extensive user time to move through multiple downloaded documents to find a specific piece of critical subsurface information. This can be especially true when the workflows became saturated with large volumes of data and usage patterns that require rigorous application of industry accepted optimal practices for digital data management (Figure 1) for projects with decadal scales (Mavroeidi and Rattenbury, 2022).The effectiveness of the data curation steps can be measured by quantitative measures of compliance with implementation profiles for low latency in finding, accessing, interoperating with, and reusing data (Kinkaid and Shepherd, 2020).

Recent advances have added the use of a contextual natural language query tool on top of the existing data store to both decrease data decision latency in the cycle times for finding and accessing data, and to increase the interoperability and reusability of specific information by assisting end users in locating the source, lineage and provenance of located data. These aspects will continue to become more critical to digital data sets matching the reality of the subsurface as more legacy is re-used to explore and evaluate, and then appraise and asses, the suitability of geologic reservoirs and pore space for low-carbon initiatives. The data used in these studies has been selected from a working data set of over 100 petabytes of global subsurface data to represent data types currently under management for decision support in projects including natural hydrogen exploration, geothermal energy, carbon and energy storage in geologic reservoirs, geohazard estimation for location of renewable energy infrastructure such as wind turbines and solar farms, and geologic disposal of byproducts from zero-carbon nuclear power generation.

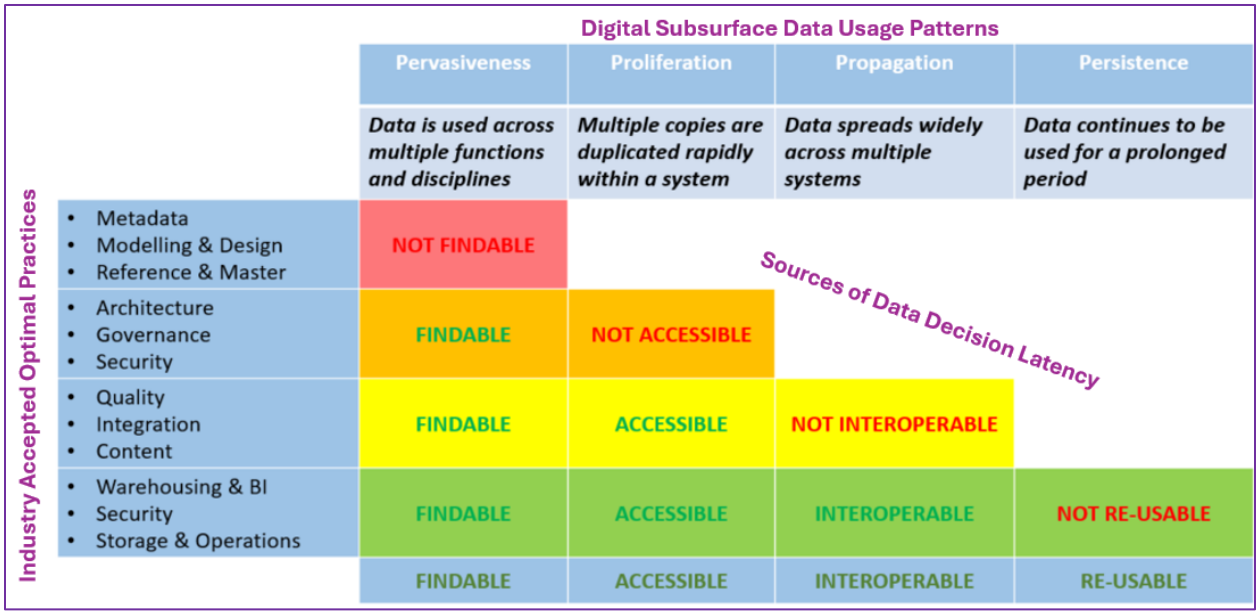


*Figure 1. Relationship between industry accepted optimal digital data management practices (Chatfield and Selbach, 2011), usage patterns of large volume digital subsurface data, and the resultant potential for increased data decision latency as measured in FAIR implementations.*

## Observations and Conclusions

We have observed that when dealing with the multiple document formats and layouts from historical subsurface data sets such as those held by government regulatory agencies, metadata extraction and classification into industry standard taxonomies using semi-automate routines can still be ambiguous or misleading (Blake et al., 2025). In the newly integrated hybrid data workflow, an Intelligent Document Processing step is added to more quickly configure the LLM without the need for large volume labeled datasets. An integrated AI/ML processing pipeline uses a generative AI algorithm to improve traditional machine learning classification routines and improve quality and repeatability of generative AI prompt responses (Figure 2). We ran tests demonstrating natural language requests for data relevant to evaluation of a subsurface reservoir for CO2 geologic storage, and for identifying potential data challenges when processing seismic data for either monitoring of CO2 plume migration in the subsurface, or evaluating subsea geohazards for placement of offshore wind turbines (Mueller and Smith, 2019). In the latter case we have previously been involved with data searches that were hampered by lack of properly curated metadata for older geotechnical datasets.

The tested workflow used a Large Language Model (LLM), Retrieval Augmented Generation (RAG) architecture, the maintained metadata, and knowledge graphs to rapidly present relevant information from a dataset consisting of hundreds of well completion and seismic support documents in the public domain. The AI “guard rails” provided by the ability of RAG architectures to capture and cite sources have been critical to the integrity needed for industry to use generative AI to support critical business decisions (Friedman, 2025). Live tests showed that searches for unique data elements in a large corpus that would have previously taken weeks to resolve in unstructured data stores such as shared file structures of hierarchical directories, could be completed with an auditable and retrievable workflow in a few minutes.

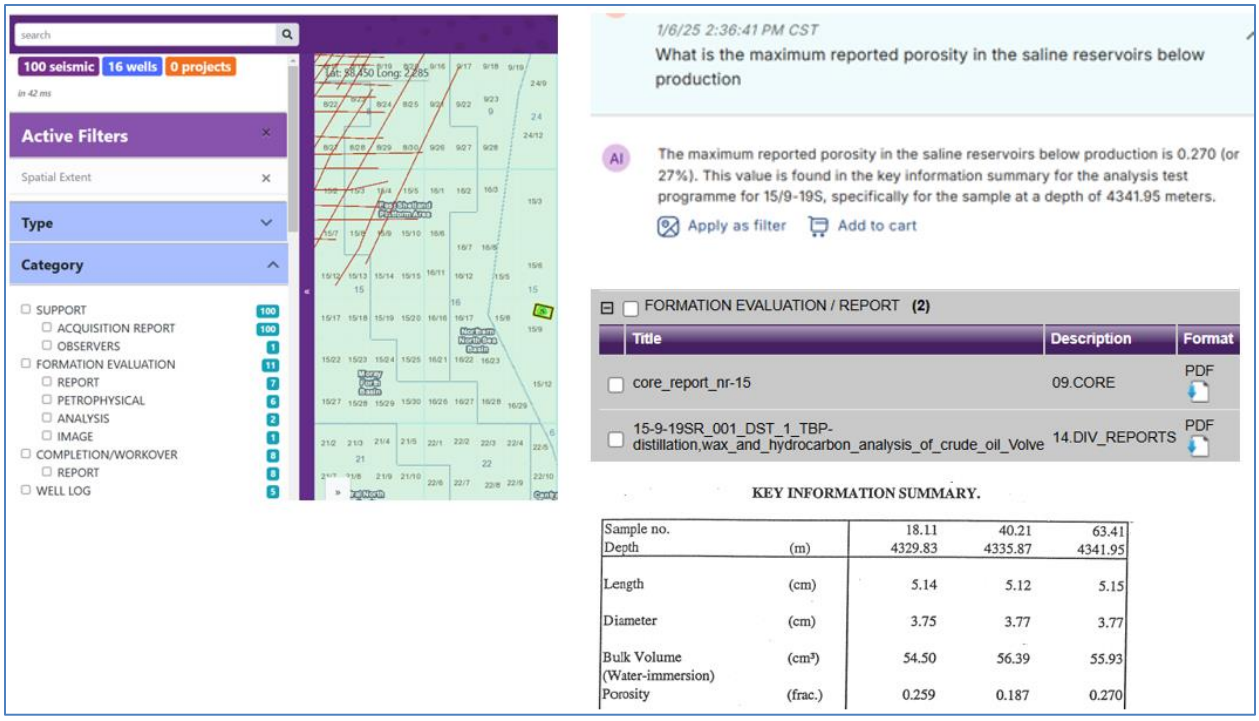


*Figure 2. Contextual Generative AI query workflow for locating business critical geotechnical data for a CO2 storage evaluation. Left; map based query for relevant data types. Right, from top down; Natural language query and specific response with document and depth reference, retrieval of document from industry standard classification taxonomy, and location of information in a table view from page 3 of a 30-page document.*

## Novel/Additive Information

We have demonstrated that a rigorous approach to the pre-conditioning, classification and delivery of large subsurface digital data sets is required to obtain effective results from the application of advanced data management tools such as generative AI. As demand for historical and legacy data to support new energy projects increases, we expect to see more applications of these type of hybrid workflows, with demonstrable decreases in cycle times for locating key business critical decision support information. While generative AI tools can assist with these reductions in cycle times, they do not replace the application of optimum industry accepted practices for data management and curation.

## Acknowledgements

We wish to acknowledge the extensive technical expertise of global development and consulting teams at both Iron Mountain and Katalyst Data Management for the progress made on this work. We also thank both organizations for the permission to publish and discuss the results of the project.

**References**

Blake, R., Kozman, J., Lamb, J. and Pelegrin, L., 2025. “Augmented Data Management for Subsurface CCUS Data Sets”, Society of Petroleum Engineers, Carbon Capture, Utilization, and Storage conference (CCUS) DOI 10.15530/ccus-2025-4186419

Chatfield, T., Selbach, R., 2011. “Data Management for Data Stewards. Data Management Training Workshop”, Bureau of Land Management (BLM).

Friedman, B., 2025. “Avoiding Missteps and Misconceptions from AI in the Energy Sector”, American Association of Petroleum Geologists, AAPG Explorer: 68199, Janauary 2025, https://explorer.aapg.org/story/articleid/68202/avoiding-missteps-and-misconceptions-from-ai-in-the-energy-sector

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

Kinkaid, D. and Shepherd, A., 2020. “Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles”, Geoscience Data Journal, 2022;9:177–186. https://doi.org/10.1002/gdj3.120

Mavroeidi, M. and Rattenbury, M.S., 2022. “FAIR Principles applied to high-value geoscience datasets”, Lower Hutt, N.Z.: GNS Science. GNS Science report 2021/62. 39 p.; doi: 10.21420/88HQ-9792

Mueller, R. and Smith, K., 2019. “Geotechnical and Geophysical Desktop Study to Support Offshore Wind Energy Development in the New York Bight - Final Report”, New York State Energy Research and Development Authority, NYSERDA Contract 135752, https://www.nyserda.ny.gov/-/media/Project/Nyserda/Files/Programs/Offshore-Wind/19-19-Geotechnical-and-Geophysical-Desktop-Study-to-Support-Offshore-Wind-Energy-Development.pdf