geoconvention
Calgary • Canada • May 11-13 2026

Curating Continuous Monitoring Data for AI Workflows

Jess B. Kozman
Katalyst Data Management

Deniz Melek Paker
Texas A&M University

Meisong Yan
Enovate.ai

Summary

Large volume, continuous monitoring data from multi-resource subsurface energy assets present unique challenges for data curation, integration, and delivery to Artificial Intelligence (AI) augmented workflows. The value of subsurface monitoring data is substantially increased when it conforms to established industry standards for ingestion, enrichment and delivery. Robust AI applications require systematic identification and mitigation of data bias, incompleteness, and inconsistency, as even modest bias levels (e.g., >5–10% systematic sampling or labeling bias) can compromise model performance and decision reliability (Wang and Singh, 2021). Using industry-standard frameworks enables the implementation of embedded, end-to-end data management workflows that support the extended lifecycles of typical subsurface energy projects. Existing standards address multiple aspects of the metadata required to manage large volume subsurface monitoring datasets. Organizing these standards into a coherent, hierarchical metadata model provides a scalable foundation for integrating large volume data streams into AI augmented workflows.
The impact of complete and curated metadata is measurable in downstream model performance. Example model outputs improved could be optimized production in mature fields with extensive seismic coverage but limited well data, improved well optimization, more reliable directional well landing zone strategies in mature unconventional reservoirs, or probabilistic geostatistical modeling of two-phase multi-equilibrium commingled production with limited data available to a machine learning workflow for reducing uncertainty in history matching (Victoria et al., 2025). Workflows augmented with text to SQL models have demonstrated execution accuracy improvements on the order of 12 to 32% when metadata such as schema column descriptions were improved with LLM tools (Wretblad et al., 2024), execution time improvements correlated to the number of metadata components populated, and answer correctness improvements of up to 17.2 % from providing a symbolic language combination of metadata and business logic (Price, 2025).

Workflows

Published case studies show that AI-augmented workflows are not inherently constrained by the physical laws governing the system, which can lead to predictions that are physically inconsistent with the underlying system behavior (Karniadakis et al., 2021). To overcome this issue, Raissi et al. (2019) introduced physics-informed neural network (PINN) to solve the partial differential equations (PDEs) in engineering. Badawi and Gildin (2025) proposed a neural operator-based proxy model that includes well settings, locations, and permeability fields for reservoir simulation,

which can reduce the simulation training data set by 75% and still deliver predictions within 1% mean relative error of pressure and saturations. Automated interpretation of subsurface seismic properties can use reversed data-label pairs for log reconstruction, fault detection, etc. to train a machine learning model, particularly convolutional neural networks (CNNS) linking seismic acoustic impedance and estimations of rock properties (Abubakar et al., 2022). These workflows require assumptions about the sampling resolution of the input seismic data set and the data rules for identifying outliers in log curve values. Embedded data workflows that enable associating quantitative estimates of seismic resolution and data accuracy rules for log curves are included in optimum accepted industry standards for subsurface data management (Fig. 1).

Working Hierarchy of Metadata

Level	Hierarchy	Facets	Metrics & Artefacts	Example Standards
8	AI Readiness	Biases	Weightings	NIST AI Risk Management Framework
7	Fitness for Purpose, Lifecycle Curation	Lineage and Provenance, Ingest => Enrich => Consume	Business Process Maps, SIPOC	LEAN 6-SIGMA
6	Quality & Confidence	Technical Assurance	Reference Lists, Business Rules	PPDM, OSDU*
5	Organizational Literacy	Democratization vs. Security	Audits, Allometric Capability Maturity	SOC2
4	Custodianship	Propagation, Pervasiveness, Proliferation, Persistence	Observable Behavioural Characteristics	DMMM
3	Governance	Who, What, Why, When, Where & How	Knowledge Areas, Environmental Factors	DAMA
2	Deliverability	Find, Access, Interoperate, Reuse	Implementation Profiles	FAIR
1	Bibliographical	Copyright, Licensing	Citations	Dublin CORE

Subject matter domain specific

Figure 1. A hierarchy of facets of metadata, arranged bottom to top by increasing value and complexity of enablement for AI augmented workflows. The levels are associated with existing industry recognized data management standards and some techniques and methodologies commonly used to embed them with large volume energy and resource subsurface monitoring data.

Machine learning object classification workflows are used to extract metadata from reports and documents that can define detailed business context relating to monitored subsurface assets. Specialized Core Analysis (SCAL) reports for example, can also contain mineral identification and abundance, thin section, or geochemical data discoverable by AI augmented workflows (Hewitt and Gadrbouh, 2023). Properly classified and processed images can provide input to models for predicting pore and grain type and behaviors in a multi-resource unconventional subsurface asset. Enriched metadata workflows have been developed for ingestion usage on industry standards-based technology agnostic systems of record. Complex relationships in large volumes of data can require machine learning workflows for unconventional reservoirs. Proper metadata population can help to overcome initial poor understanding of physical parameters such as

production drives that differ from nearby, underlying or shallower conventional drilling targets (Prochnow et al., 2022).

The curation of large volume monitoring data in the context of an end-to-end workflow can support two-way computational control with a digital twin (Gahlot et al., 2024) to optimize operating parameters, inform mission-critical decisions and reduce risks. This involves inferences from recursively trained AI models using field observations and current state samples, updated as new field data is ingested. The models must be able to use multi-model time-series data, and uncertainty metadata must be stored separately online for interoperability with the digital twin even when the full set of input data is stored in nearline or offline intelligently tiered cloud storage (Wilkinson, 2016). The workflows must accommodate further applications for edge computing (Mondanos et al., 2024), and high throughput data ingestion pipelines.

Sleipner Workflow Case Study

The Sleipner CO2 Storage project is an extensively documented case study showing auditable repeat monitoring measurements of plume evolution reducing uncertainty in containment and conformance. The location of the field is shown on Fig. 2.

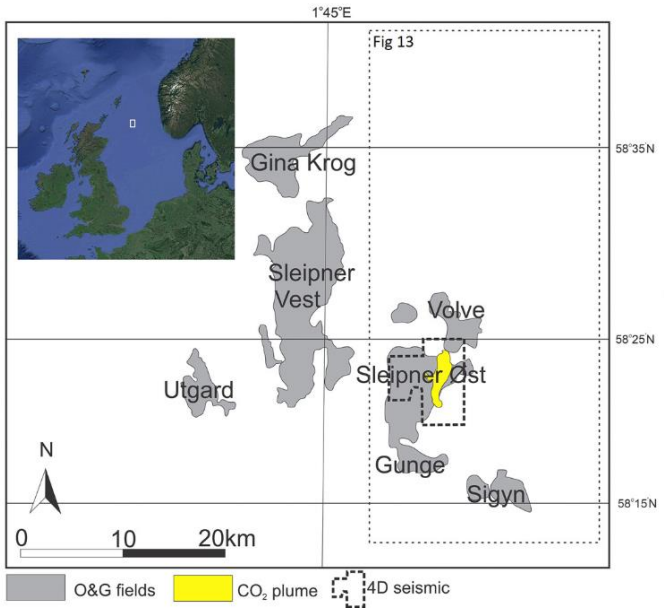


Figure 2. Location of the Sleipner Ost gas field. The repeat 4D seismic surveys have created large volumes of monitoring data. These datasets present opportunities for metadata enrichment that will reduce uncertainty when the data is used in deep learning or other AI-augmented modeling workflows (Sheng et al., 2023). Metadata can help to characterize biases such as scarcity of well log data, or inform weightings to mitigate the effects of noise or interference on seismic data quality across vintages.

With an injection history from 1996, it is also an example of where metadata can improve the AI-readiness of the large volume data from its extensive monitoring program of repeated 3D/4D

seismic surveys and complementary gravimetric studies. The uniquely rich, longitudinal subsurface monitoring dataset has become a reference point for CCS site characterization, risk management, and data-driven interpretation workflows, and can be a key study volume for characterizing, measuring, and enriching AI-readiness.

Repeated 3D/4D seismic complemented by repeat gravity, for example, could benefit from quantitative recorded geospatial uncertainty carried into the joint seismic and gravimetric inversion (Furre et al., 2024). as well as addressing the challenges due to CO2 injection into formation as shown in Fig. 3 (Furre et al., 2024; Rutqvist, 2012; Ringrose et al., 2013). The inversion had a direct engineering impact by turning plume behavior from an assumed model outcome into observed, continuously updated constraints on reservoir storage performance. Seismic images repeatedly mapped the plume geometry and its migration pathways, while joint seismic–gravity interpretation helped quantify changes in mobile "free" CO_2 mass and density. Improved metadata documentation of uncertainty would further strengthen the containment assurance and loop closure between injection operations and subsurface response.

The Sleipner data set is a good example of how sampling bias can degrade the results of AI-augmented workflow such as neural network (NN)-based interpretation methods (Li and Li, 2021). A random subset sample of 3D cubes from baseline, time-lapse and label training volumes will necessarily have an oversampling of seismic parameters not representing the labeled CO2 plume, which slows model convergence.

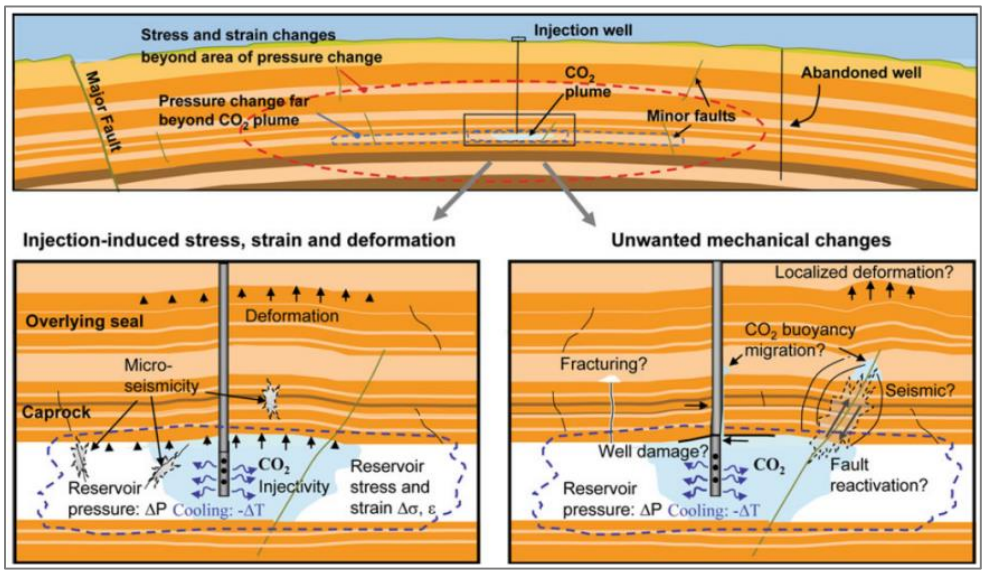


Figure 3. Geomechanical challenges of CO2 storage in deep formations (After Rutqvist, 2012).

The repeat large volume monitoring record forces iterative refinements of static and dynamic models of permeability architecture, baffle heterogeneity, and layer spreading. AI-augmented workflows could benefit from metadata reflecting how amplitude matching levels on repeat vintages were preserved in the reprocessing workflows. These workflows were used to improve

history matching and reduce uncertainty in forward predictions used for risk management and surveillance planning.

Use of the long-duration monitoring dataset as an engineering decision support system improves confidence in safe long-term storage, informs how monitoring toolboxes should be designed for offshore saline aquifer storage, and provides regulators and the operators including Equinor and partners with extensive, defensible, measurement-backed and data-driven evidence of plume evolution over decadal timeframes, rather than relying only on the simulation models. With the large amounts of time and effort from operators and regulators spent on acquiring and managing these large volume monitoring data sets, it is a prime candidate for field validating AI-readiness approaches to metadata storage and maintenance.

Continuous monitoring provides economic impacts that increase value of information, reduce cost of uncertainty, and help eliminate bias that can permeate, pervade, propagate, and persist in AI-augmented workflows. Repeatedly imaging and validating plume evolution with consistent and auditable measurement workflows at great cost should mandate a similar investment in sustainable custodianship of the resulting data. A positive return on investment in AI-ready and augmented data ingestion, enrichment and consumption workflows can be demonstrated by tracking containment and conformance audits with reduced uncertainty, lowering the probability of expensive contingency actions (additional wells, unplanned monitoring campaigns, remediation, or injection curtailment). AI-ready metadata population will strengthen the evidence base needed for long-term storage assurance.

Auditable data-driven confidence measures supported by robust and AI-ready metadata will help the overall CCS value chain more broadly in tax and economic regimes like Norway or Australia, where tax and incentive advantages can help to offset or exceed storage costs. Business context metadata will support key proof mechanisms that enable a value proposition over decades of site operation. As monitoring programs become more economically efficient over time, operators will want to record in quantifiable metadata attributes how reductions of uncertainty in the velocity model approach the point where reliable intra-plume depth conversion can be performed across monitoring vintages. A goal based on value of information (VOI) methodologies (Lino et al., 2025) will drive acquisition and processing approaches including effective curation of AI-ready metadata toward cost-effective repeatable 4D workflows and measurable enhanced decision-quality surveillance per dollar spent.

Observations

Once large volumes of continuous reservoir monitoring data have been ingested in a system of record, they are available for data and metadata enrichment (Carr et al., 2023). Quantitative measures of Findable, Accessible, Interoperable, and Reusable (FAIR) can be determined by determining compliance with implementation profiles (Hutchison et al., 2024). Public domain evaluations are available for repositories including research and monitoring metadata catalogs, data needed to support embedded and connected dataflows for low-carbon energy research (Schwanitz et al., 2022) and geospatial resource activity data (Minghini et al., 2020). Public profile compliance measurements provide a baseline to evaluate the FAIR compliance ratings of the aggregated and anonymized metadata from a 100+ petabyte system of record used by resource operators for AI augmented workflows (Fig. 4).

The highest FAIR compliance scores were from the United States Geological Survey geoscience data catalog, comprised mostly of peer reviewed publications and academic reports, which require a fixed set of bibliographic metadata to be submitted (Hutchison et al., 2024). This matches well with our observation that successful AI-augmented queries have been documented working against a corpus of structured reports and support documents (Blake et al., 2025). The next highest scores came from the multi-petabyte curated dataset, reflecting a rigorous and documented application of business rules, governance policies, mandatory metadata attributes, and QA/QC workflows. Using some of the lower scoring public repositories could mean that cumulatively over the FAIR workflows being used to populate, validate and train AI-models, data could be as much as 97% non-compliant with industry accepted best practices for AI-ready data curation.



Figure 4. Comparison of FAIR compliance ratings for collections of data that could be used to support AI-augmented workflows for continuous monitoring data.

The requirements for FAIR compliance align with facets identified in risk management frameworks used by energy and resource organizations for data used in AI-augmented workflows. Examples of AI data-related risk include imprecision of large and diverse subsurface datasets (Jayabal, 2025) and the fidelity of their reproduction of the physics for a subsurface system. This risk can be mitigated by establishing business rules for carrying uncertainty metadata through the data lifecycle in accordance with FAIR data access standards. Metadata attributes can capture key data-related uncertainties, including sparse or spatial clustered sampling, sub-resolution physical heterogeneity, and measurement tool calibration drift (Akinnurun, 2025). In addition, quantitative differences between prior measurement distributions and posterior distributions following data assimilation can be explicitly encoded in metadata attributes. Addition of these values can be a

low cost and low effort way to improve the metadata on large volumes of monitoring data being used by AI-augmented workflows.

Advances in the availability of quantum sensors, storage and computing may also lead to both more complex time-series data volumes from monitoring, and new categories of metadata that will enrich existing FAIR criteria (Castagna, 2026). AI-augmented modeling workflows should therefore be deliberately engineered to accommodate information theory-based properties such as channel capacity, mutual information, cross entropy, divergence, variation of information, and spectral entropy, which can carry important weighting information for future AI-enhanced workflows.

Inconsistency and variability of large volume datasets over order of magnitude temporal and spatial scales can provide challenges to using data in AI-augmented integration workflows (Gabry, 2024). Data accuracy tags with detailed provenance as described in FAIR interoperability standards provide a means for AI workflows to weight data properly in AI-augmented workflows. One quantitative measure of improved metadata management is reduced data decision latency (Hackathorn, 2004); the elapsed time between recognition of a decision trigger and having interoperable, analysis-ready data available within a system of insight. In subsurface monitoring workflows, such systems of insight are usually best of breed interpretation applications with access to AI-augmented analytics. The empirical evidence shows that increased conformance with FAIR implementation profiles in systems of record can decrease data decision latency by up to 34% for large volume subsurface monitoring datasets (Fig. 5), with an additional 3% attributable to project time required for preparing decision support records for reuse.

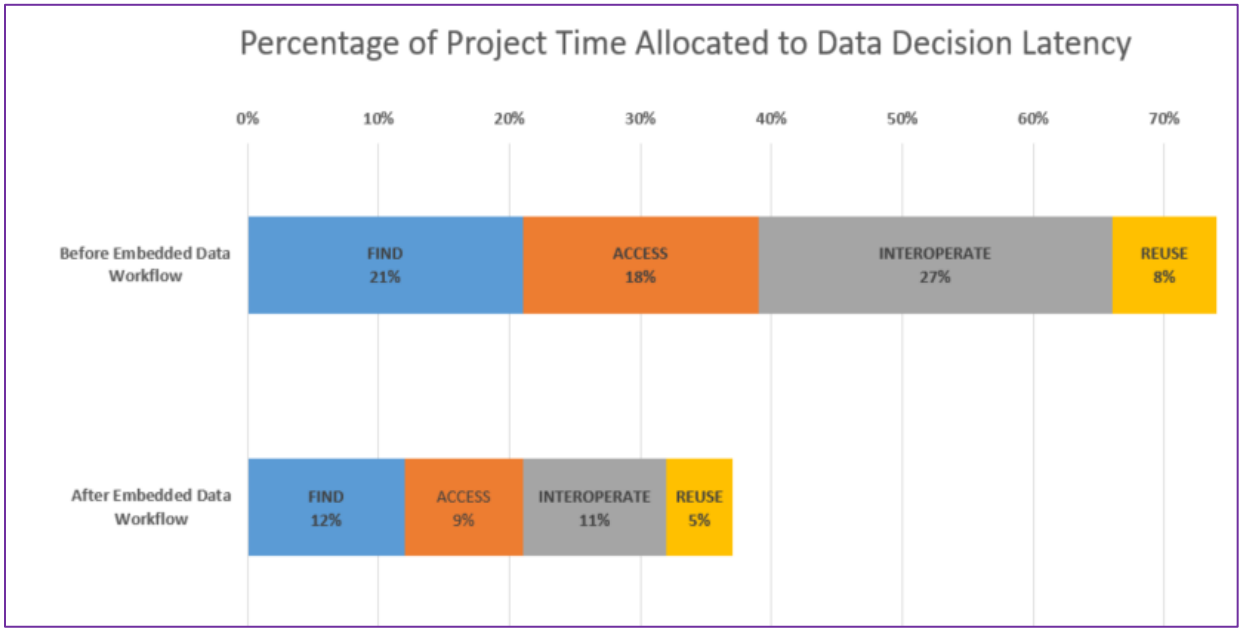


Figure 5. Reduction in total data decision latency before and after the implementation of embedded data workflow methodologies including population of critical metadata for large volume reservoir monitoring data. Measurements were collected from multiple case studies and forced-ranking style surveys. After Kozman et al. (2024).

Existing workflows demonstrate the relationship between metadata completeness and decision latency. Raw data acquired to monitor micro-seismic events during CO2 injection can be delivered in sets of thousands of individual files (Jacques, 2019) organized by time-date stamp with a relationship to the triggering event (Fig. 6). Because of operational changes over the course of a CO2 injection project, the geospatial metadata, size of file, equipment used and threshold for the triggering event can all change. Using data with the resulting biases to model and optimize subsurface processes using AI is both time consuming and problematic, especially when critical metadata must be gleaned from disparate reports in multiple locations.

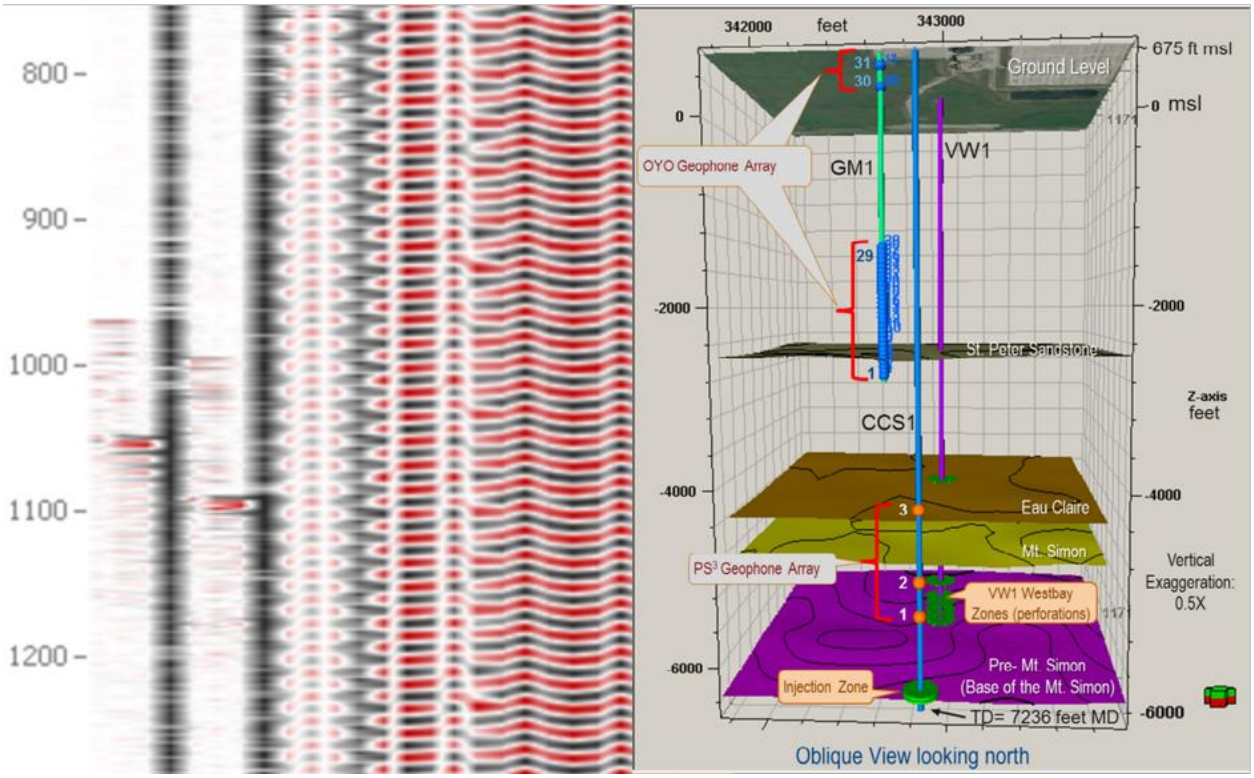


Figure 6. Left: Visualization of raw seismic event data showing a triggering event recorded at two wells during CO2 injection. Right: In-ground equipment configuration diagram extracted from support documents showing how the data was recorded. Successfully linking these two data items is required to meet FAIR data compliance requirements. (Data from Jacques, 2019, Illinois Basin-Decatur Project, visualization provided by Dynamic Graphics, Inc. - used by permission.)

Additive Information

There is a need to develop methodologies for ingesting and storing continuous monitoring data using industry standard taxonomies that will support compliance with FAIR data implementation profiles. Monitoring data will be used to support AI-augmented workflows. Multiple operators and service providers will need to collaborate to develop data curation standards that allow FAIR compliant links to metadata source documents using shared taxonomies. This work is novel in explicitly linking continuous subsurface monitoring data ingestion with FAIR-compliant metadata curation as a prerequisite for reliable AI-augmented workflows, rather than treating metadata standardization and AI modeling as independent problems.

Acknowledgements

The author wishes to acknowledge the work of multiple technical teams at Katalyst Data Management and Dynamic Graphics, Inc. for support of this work.

References

Abubakar, A., H. Di, A. Kaul, C. Li, Z. Li, V. Simoes, L. Truelove and T. Zhao. (2022). Deep learning for end-to-end subsurface modeling and interpretation: An example from the Groningen gas field, Society of Exploration Geophysicists, The Leading Edge, 41. 259-267, https://doi.org/10.1190/tle41040259.1

Akinnurun, Y. (2025). Uncertainty Quantification and Risk Assessment in Reservoir Modeling Using Bayesian Machine Learning, Medium, https://medium.com/@yomibest/uncertainty-quantification-and-risk-assessment-in-reservoir-modeling-using-bayesian-machine-4b294a899949

Badawi, D. and Gildin, E. (2025). Neural operator-based proxy for reservoir simulations considering varying well settings, locations, and permeability fields. Computers & Geosciences, 196, 105826. https://doi.org/10.1016/j.cageo.2024.105826.

Blake, R., J.B. Kozman, J. Lamb, L. Pelegrin. (2025). Augmented data management for subsurface CCUS data sets, Society of Exploration Geophysicists, Geoscience World, Proceedings of the Carbon Capture, Utilization, and Storage Conference and Exhibition, Paper No: ccus-2025-4186419, pp. 239-243
https://doi.org/10.15530/ccus-2025-4186419

Carr, S., J. Kozman and C. Lopes (2023) Managing large volume time-lapse datasets for sustainable carbon storage in subsurface reservoirs, Sociedade Brasileira de Geofísica, 18th International Congress, Brazilian Geophysical Society https://sbgf.org.br/mysbgf/eventos/expanded_abstracts/18th_CISBGf/d41d8cd98f00b204e9800998ecf8427eSBGF%20Abstract%20-%20KDM%20-%20Carbon%20Storage%20-%20%20JK%2012%20May%202023%20v3.3.pdf

Castagna, J.P. (2026) Application of Information Theory in the Geosciences. 32nd Annual Milton B. Dobrin Lecture, University of Houston, Department of Earth and Atmospheric Sciences, College of Natural Sciences and Mathematics, https://www.uh.edu/nsm/earth-atmospheric/news-events/dobrin-lecture/

Furre, A.-K., M. J. Warchoł, H. Alnes and A. S. M. Pontén (2024) Sleipner 26 years: how well-established subsurface monitoring work processes have contributed to successful offshore CO2 injection, open access research article, Geoenergy, Volume 2, https://doi.org/10.1144/geoenergy2024-015

Gabry, M.A. (2025) Transforming Subsurface Energy Landscape Through AI and Electrification, Society of Petroleum Engineers, The Way Ahead, University of Houston, Department of Petroleum Engineering, Cullen College of Engineering, https://jpt.spe.org/twa/university-of-houston-transforming-subsurface-energy-landscape-through-ai-and-electrification

Gahlot, A.P., H. Li, Z. Yin, R. Orozco and F. J. Herrmann. (2024). A Digital Twin for Geological Carbon Storage with Controlled Injectivity, arXiv:2403.19819v1 [physics.geo-ph] 28 Mar 2024. Downloaded 19-Jan-2026 from: https://arxiv.org/pdf/2403.19819

Hewitt, H. and. Gadrbouh R. (2023). Futureproofing Rich Metadata File Ingestion with OSDU. Proceedings of Third EAGE Digitalization Conference and Exhibition, London, UK, 20-22 March. ISBN: 978-1-7138-7495-9, 162-165.

Hutchison, V. B., Norkin, T., Zolly, L.S. and Hsu, L. (2024). State of the Data: Assessing the FAIRness of US Geological Survey Data. *Data Science Journal*, **23**(22), https://datascience.codata.org/articles/10.5334/dsj-2024-022

Jacques, P., R.A. Bauer, and W. Malkewicz, (2019) The Illinois Basin-Decatur Project (IBDP) Microseismic Monitoring: Geophones, Arrays, Procedures, Velocity Models, Chronology, Catalogue and Notes on Data. CO2 DataShare, Illinois State Geological Survey, downloaded from: https://co2datashare.org/dataset/illinois-basin-decatur-project-dataset/resource/619bfa9f-6093-4692-830c-443f55a1c289

Karniadakis, G. E., Kevrekidis, I.O., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nat Rev Phys* **3**, 422–440 (2021). https://doi.org/10.1038/s42254-021-00314-5.
Kozman, J. Lowery, S. and Lamb, J. (2024). End-to-end workflow for managing large volume data from CCUS=. Proceedings of the Carbon Capture, Utilization, and Storage Conference and Exhibition, 11–13 March 2024, pp. 251-260. Paper No: ccus-2024-4014747, https://doi.org/10.15530/ccus-2024-4014747

Li, B., & Y. E. Li, (2021). Neural network-based CO2 interpretation from 4D Sleipner seismic images. Journal of Geophysical Research: Solid Earth, 126, e2021JB022524. https://doi.org/10.1029/2021JB022524

Lino, A., Y. Kobayashi and Y. Horiuchi (2025) An Efficient Workflow to Quantify the Value of Information and Uncertainty Reduction from Time-Lapse Seismic Data for Monitoring Optimization and Reservoir Characterization in CCS Projects, Paper presented at the SPE Annual Technical Conference and Exhibition, Houston, Texas, USA, October 2025. Paper Number: SPE-227993-MS https://doi.org/10.2118/227993-MS

Minghini, M., V. Cetl, L. Ziemba, R. Tomas, D. Francioli, D. Artasensi, E. Epure, and F. Vinci.(2020). Establishing a new baseline for monitoring the status of EU Spatial Data Infrastructure: European Commission Joint Research Centre Technical Report, JRC122351, DOI:10.2760/296219. https://op.europa.eu/en/publication-detail/-/publication/dfa18cf5-3dba-11eb-b27b-01aa75ed71a1/language-en

Mondanos, M., P. Marchesini, A. Stork, C. Maldaner, G. Naldrett, and T. Coleman. (2024). A distributed fibre-optic sensing monitoring platform for CCUS, European Association of Geoscientists & Engineers, First Break, Volume 42, Issue 3, Mar 2024, p. 77 – 83, DOI: https://doi.org/10.3997/1365-2397.fb2024026

Price, G. (2025) The Impact of Metadata Configurations on Text-to-SQL Performance: A Comprehensive Analysis, CorralData, Research paper, downloaded from https://github.com/CorralData/research-metadata-text-sql-2025/blob/main/PriceCorralData2025_MetadataTextSQL.pdf

Prochnow, S. J., N. S. Raterman, M. Swenberg, L. Reddy, I. Smith, M. Romanyuk and T. Fernandez (2022).A subsurface machine learning approach at hydrocarbon production recovery & resource estimates for unconventional reservoir systems: Making subsurface predictions from multidimensional data analysis, Journal of Petroleum Science and Engineering, Volume 215, Part A, 110598, ISSN 0920-4105, https://doi.org/10.1016/j.petrol.2022.110598.

Raissi, M., Perdikaris, P., and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J of Comp. Phy.*, 378, 686-707. https://doi.org/10.1016/j.jcp.2018.10.045.

Ringrose, P. S., Mathieson, A. S., Wright, I. W., Selama, F., Hansen, O., Bissell, R., Saoula, N. & Midgley, J. (2013). The In Salah CO2 storage project: lessons learned and knowledge transfer. *Energy Procedia*, 37, 6226– 6236 (GHGT11)

Rutqvist, J. (2012). The Geomechanics of CO2 Storage in Deep Sedimentary Formations. *Geotechnical and Geological Engineering*, June 2012, 30 (3), 525-551.

Schwanitz, V. J., A. Wierling, M. E. Biresselioglu, M. Celino, M. H. Demir, M. Bałazińska, M. Kruczek, M. Paier and D. Suna (2022). Current state and call for action to accomplish findability, accessibility, interoperability, and reusability of low carbon energy data. Scientific Reports, *Springer Nature Portfolia*, 12, 5208. https://doi.org/10.1038/s41598-022-08774-0

Sheng, H., X. Wu, X. Sun and L. Wu (2023) Deep learning for characterizing CO2 migration in time-lapse seismic images, Fuel, Volume 336, 126806, ISSN 0016-2361, https://doi.org/10.1016/j.fuel.2022.126806.

Wang, Y. and Singh, L. (2021). Analyzing the impact of missing values and selection bias on fairness. *Int J of Data Sci and Anal.*, 12, 101-119. https://doi.org/10.1007/s41060-021-00259-z

Wretblad, N., Holmstrom, O., Larsson, E., Wiksater, A., Soderlund, O. et al. (2024). Improving Relational Database Interactions with Large Language Models: Column Descriptions and Their Impact on Text-to-SQL Performance. arXiv:2408.04691. https://doi.org/10.48550/arXiv.2408.04691

Yan, M. (2025) Carbon Capture and Storage Types, Applications, Principles, Best Practices in the World, in B. Dindoruk, C. Temizel, and R. Ratnakar, ed. Carbon Capture and Storage in the Oil and Gas Industry: Solutions for the Energy Transition. 1st Edition, Elsevier Science, ISBN: 9780443157615, eBook ISBN: 9780443157622