SPE · AAPG · SEG
CCUS
Carbon Capture, Utilization, and Storage

CCUS: 4427990

Digital Solutions for Large Volume CO_2 Characterization, Testing and Monitoring Data

Matthew Fry*[1], William Hanson[2], Jess Kozman[1], Jane Wheelwright[2] 1. Katalyst Data Management, 2. Dynamic Graphics

Abstract

Digital solutions for managing large volumes of data are a key component of workflows to identify potential CO_2 storage sites and to demonstrate conformance with testing and monitoring requirements. Interest in these data sets has grown with heightened awareness of the need for well curated datasets to train and validate Artificial Intelligence (AI) and machine learning (ML) models. Real-time data sets contribute to analysis of those models, and these datasets grow to hundreds of terabytes with distributed fiber optic sensor deployments collecting continuous passive monitoring data. The resulting emerging data driven workflows require high-quality datasets to support important geotechnical decisions about hydraulic flow units and plume monitoring simulations. Existing optimum industry accepted data management practices can be applied to improve AI/ML augmented results.

Introduction

Finding and accessing the most appropriate site-specific CCUS data for modeling workflows can be assisted by AI/ML tooling such as text-to-SQL systems. Subsurface metadata configurations that include schema information with curated table and column descriptions can improve both query accuracy and performance (Price, 2025). Workflows with accurately discovered data including uncertainties and technical assurance metadata can support business decisions related to confining zones, storage capacity, injectivity, and leakage pathway risk. Existing industry standards for data definitions, formats and schemas can be adopted and adapted for CCUS datasets. Open-source artifacts supporting these standards, such as Well-Known Schemas (WKS) and Work Product Components (WPC), are collaboratively maintained by industry forums to improve enterprise level trust (Moradi and Hepsø, 2024). The workflows are designed to be fit-for-purpose for subsurface screening methodologies, building on data management strategies already in use for mature oil and gas exploration and production data

workflows (Blake et al., 2025). Proven best practices for data management and curation are re-purposed and future proofed from other data driven, capital intensive, and highly regulated industries. Targeted industry sectors using large volumes of monitoring data include health, transport, gaming, high-energy physics, computational fluid dynamics, and astrophysics. Identification of event-triggered time-series data streams with spatial metadata, for example, can benefit from best practices being implemented by large scale radio telescope installations (Thavasimani, 2021). Some of these data handling methods can be applied to data submission guidelines for regulatory submission of large passive monitoring data sets such as ocean bottom node seismic (NSTA, 2025).

Existing petabyte-scale aggregated datasets are available for study that contain datasets at larger temporal and spatial scale than reservoir characterization studies from conventional oil and gas field operations (Ringrose et al., 2013). Datasets used to establish best practices contain, for example, metadata from thousands of time-series passive monitoring records indexed to individual micro-seismic events, and over 3000 interpretation projects targeted at CCUS operators ingested since 2019. Anonymized metadata trends from these data sets can be used to test for reductions in data decision latency times across embedded search and order workflows for key screening, testing and monitoring data types. A value case for reductions in project cycle time can be based on persistent usage of large volume continuous reservoir monitoring data over the decadal project lifecycle of a CO_2 injection project. Analysis demonstrates how metadata components such as column descriptions in structured databases can increase efficacy in extracting meaning for AI/ML augmented workflows (Price, 2025). Highly contextualized business domain expertise is required to mitigate the risk of AI-related errors in these applications, especially for long-form responses to open-ended prompts (NIST, 2024).

Methods

Spatial and temporal biases of data lead to increased risk when using large volume monitoring data to support AI augmented workflows (Brodaric, 2021). Levels of AI/ML data readiness to support CO_2 projects can be measured with existing data management best practices, developed and field validated for large volumes of hydrocarbon exploration data, and adapted specifically for large volumes of CO_2 monitoring data (Geary, 2023). Government agencies and other repositories of public domain data supporting CO_2 injection projects provide examples of both the effort level required and the value to AI readiness of extensive curation and data labeling (Morkner et al., 2022). This value is recognized in site selection, determination of storage capacity suitability, and in hazard modeling and mitigation. Forced-ranking style surveys and case studies demonstrate a relationship between data that is Findable, Accessible, Interoperable, and Reusable (FAIR) as measured by industry specific implementation profiles (Schultes, 2020), and levels of enterprise level data trust (Hiniduma, Byna and Bez, 2025) that reduce the data-related risk in AI augmented workflows.

Results

A comparison of different sources of digital data that can support evaluation and optimization of CO_2 geologic storage projects illustrates their relative conformance with FAIR data management principles. The discovered gaps in requirements are compared to criteria for reducing data-related AI risk to discover correlations. For instance, the access criteria for FAIR data include metadata completeness (Wilkinson, 2016), and generative AI risk reduction requirements for data governance include fully maintained and auditable change records that reflect provenance, such as source and timestamp (NIST, 2024). This informs efforts to add more complete metadata to monitoring data delivered for CCUS projects (Fig. 1).

Generative AI risk frameworks also highlight the need to enhance provenance content with organizational efforts that track the history and origin of data used as input to training, validation, and modeling workflows. Large volumes of CO_2 monitoring data are often transmitted in grouped files organized by time-date stamp (Jacques, 2019) and are subsequently submitted to regulators in collections identified by

a common spatial reference (NSTA, 2025). Maintaining version control and provenance for use in AI augmented workflows requires populating and enriching mandatory metadata attributes with an auditable workflow on a data platform system of record, and when delivering the data for analysis and interpretation. Existing metadata models, schemas, and workflows from other high volume data applications across disparate scientific disciplines can be adopted and adapted for larger volumes of continuous CO_2 monitoring data (Patterson, 2019).

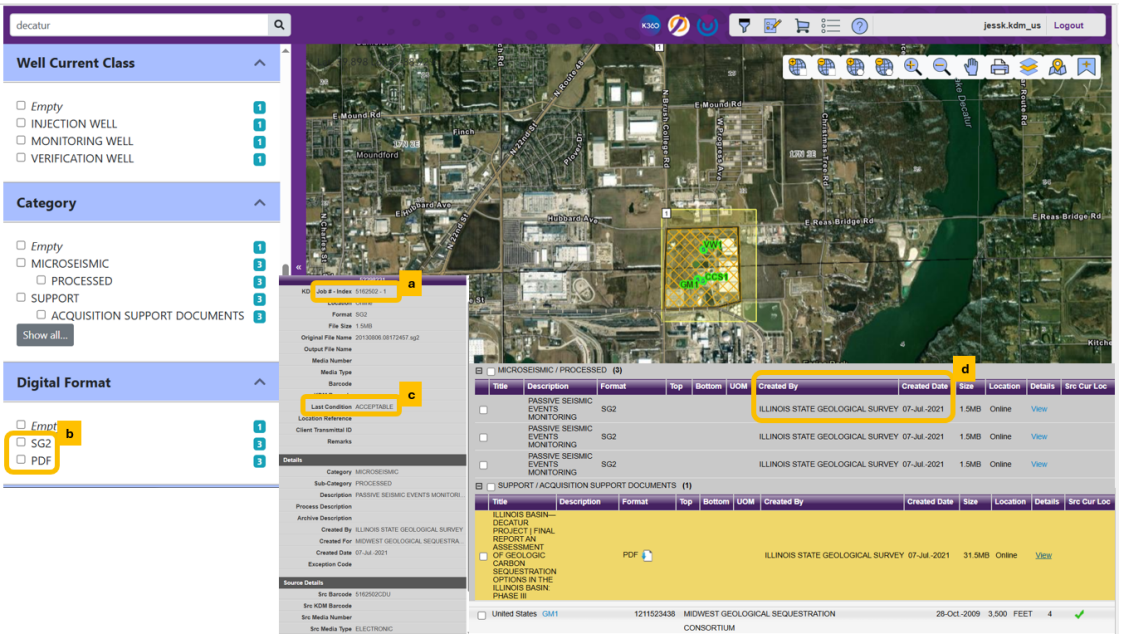

Figure 1. Example of map-based data management portal for data from CO_2 injection, monitoring, and verification wells, showing applications of FAIR data management principles and practices, including; a) unique and persistent identifiers (index numbers), b) industry standard formats (*.SG2 and *.pdf), c) data quality (last condition), and d) provenance - lineage tags (created by, created for), supporting data for AI workflows. Data from Illinois State Geological Survey, 2021., Used by permission under DOE Cooperative Agreement No. DE-FC26-05NT42588, 2021.

Open-source, technology-agnostic and standards-based cloud native data platforms can be used to ensure interoperability of CO_2 monitoring data across functional and discipline silos (Wheelwright, 2024). This enables secure, timely and reliable delivery of complex time-series data through a set of core services that enable FAIR data delivery. Consistent metadata capture enables key workflows supporting time-series data consumption, visualization and analysis in four dimensions. Participation in these long-term collaborative initiatives can simplify data governance and eliminate the need for duplication or transformation of data (Fig. 2).

Discussion

Completeness of metadata is measured in multiple facets of FAIR implementation profiles. FAIR data compliance assists in providing metadata used by AI-augmented workflows to reduce risk. Using AI-augmented workflows with large volume CO_2 monitoring data sets will require metadata that identifies and describes potential spatial and temporal biases in the data used to train and validate models. These biases can be introduced by non-random sampling or by edge processing strategies designed to reduce on-site data volumes (Isaenkov et al., 2021). Proper metadata can help reduce errors such as false positive identification of micro-seismic events when attempting to overcome the limitations of human processing workflows (Stork et al., 2020).

A fully curated digital solution for management and delivery of large volume CO_2 monitoring data will enable low-risk usage of data for AI-enabled workflows. The requirements of FAIR data management principles map closely to facets of AI risk management frameworks being adopted by the energy and

resource industry (Wierling et al., 2021). There is an opportunity to collaboratively build end-to-end workflows for datasets that will enable use of AI for data analysis and interpretation.

Uncertainty measures in curated metadata become more important as characterization and monitoring data is used to train, validate and calibrate models of CO_2 subsurface assets. Identifying and mitigating inherent biases in datasets can avoid modeling outcomes outside of possible physics-based results. Using curated uncertainty metadata can improve workflows using newly acquired time-series data to correct errors in simple machine learning models (Seabra et al., 2024), or normalizing for biases introduced as a CO_2 plume growth introduces increasing variability of relative physical measurements (Gahlot, 2025).

Proper indexing and auditable workflows of business context and lineage can also assist in recovering original raw and field data sets to meet regulatory requirements as they change over time. Regulatory thresholds for reporting micro-seismic events can change depending on proximity to newly built infrastructure (Schultz et al., 2021) or population density growth (BCER, 2025), or AI algorithms can advance to provide more finely tuned detection (Mandler, Karimi and Hutton, 2024), leading operators to retrieve datasets supporting events that may not have been included in previous visualizations.

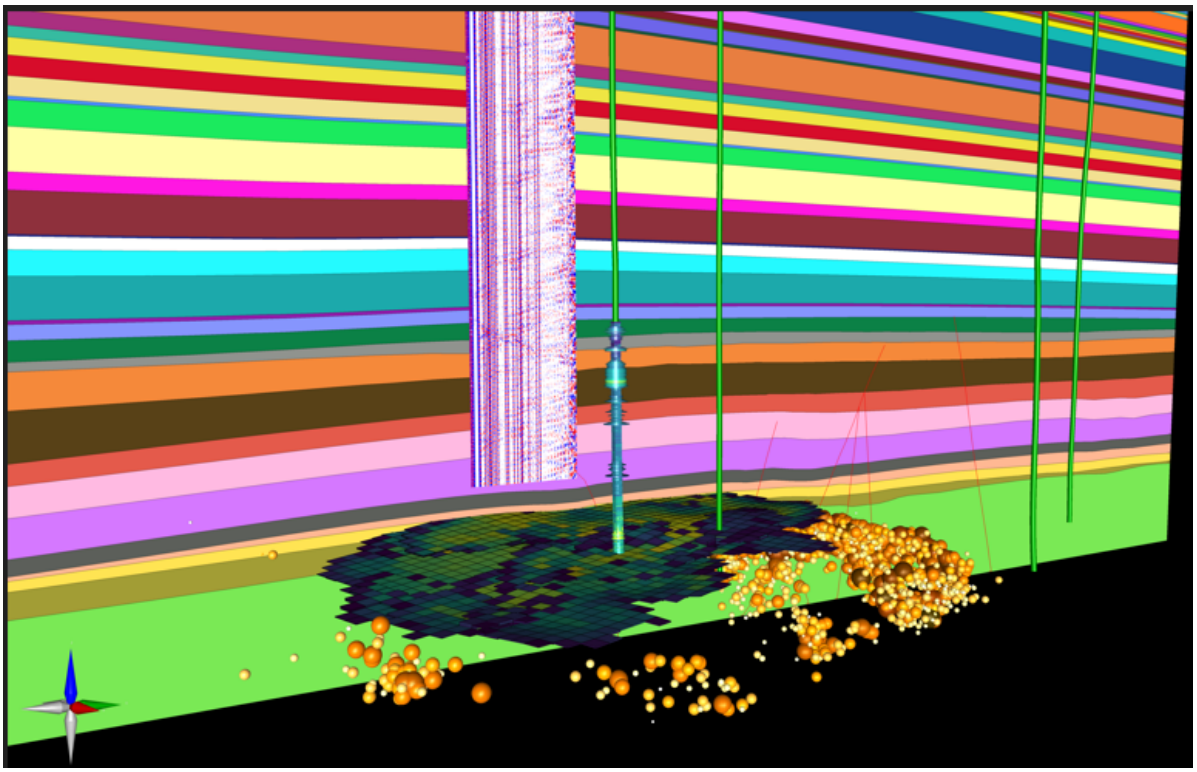

Figure 2. An analysis and interpretation application can benefit from search and order capabilities from curated data delivery platforms. The visualization shows micro-seismic events, CO_2 plume simulation, and continuous monitoring data from passive seismic monitoring at a CO_2 injection site. Collection and preservation of mandatory metadata attributes can enable an end user to maintain links between microseismic events recorded during CO_2 injection, and the original raw field data related to that event by both geospatial and temporal coordinates. Provenance, and lineage data can help identify and mitigate biases in such multidisciplinary and complex datasets. In this case a time-slider function allows a user to identify the seismic traces associated with an event. Graphic from Dynamic Graphics, Inc., used by permission.

Conclusions

Applying industry-accepted optimum practices for data management and curation is an important and often overlooked step in making subsurface screening and monitoring data available for interoperation and reuse in CO_2 storage projects. Guidelines and assessment techniques for determining if existing data sets are fit-for-purpose will assist with the demanding requirements of emerging AI and machine learning technologies.

References

Blake, R., Kozman, J., Lamb, J. and Pelegrin, L. 2025. Augmented data management for subsurface CCUS data sets. *Proceedings of the Carbon Capture, Utilization, and Storage Conference and Exhibition*, 3–5 March 2025: 239-243, ASME. https://doi.org/10.15530/ccus-2025-4186419

British Columbia Energy Regulator (BCER). 2025. Induced Seismicity Operational Manual, Version 1.0: February 2025. https://www.bc-er.ca/files/operations-documentation/Induced-Seismicity-Data-and-Submission/IS_OpsManual.pdf

Brodaric, B., Boisvert, E., and Smirnoff, E. 2021. Geoscience data quality and machine learning: Challenges and opportunities. *Geoscience Data Journal* **8** (2): 108–122. https://doi.org/10.1002/gdj3.121

Gahlot, A.P., Orozco, R., Yin, Z., Bruer, G., and Herrmann, F.J. 2025. An uncertainty-aware digital shadow for underground multimodal CO2 storage monitoring. *Geophysical Journal International* **242**, 1–31 (Advance Access publication Research paper) https://doi.org/10.1093/gji/ggaf176

Geary, A. 2023. Seismic Soundoff: Carbon storage data management done right. *The Leading Edge* **42** (12): 852. https://doi.org/10.1190/tle42120852.1

Hiniduma, K., Byna, S. and Bez, J. L. 2025. Data Readiness for AI: A 360-Degree Survey. 2024. *arXiv e-prints* 0.48550/arXiv.2404.05779. https://dlnext.acm.org/doi/pdf/10.1145/3722214

]Isaenkov, R., Pevzner, R., Glubokovskikh, S., et al. 2021. An automated system for continuous monitoring of CO2 geosequestration using multi-well offset VSP with permanent seismic sources and receivers: Stage 3 of the CO2CRC Otway Project. *International Journal of Greenhouse Gas Control*, **108**: 103317. https://www.sciencedirect.com/science/article/abs/pii/S1750583621000694

Jacques, P., Bauer, R.A. and Malkewicz, W. 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. https://co2datashare.org/dataset/illinois-basin-decatur-project-dataset/resource/619bfa9f-6093-4692-830c-443f55a1c289

Mandler, H., Hutton, L., and Karimi, S. 2024. Passive Monitoring and Induced Seismicity Risk Management for CCS Projects, *Canadian Society of Exploration Geophysics*, Featured Focus Article, https://cseg.ca/passive-seismic-monitoring-and-induced-seismicity-risk-management-for-carbon-storage

Moradis, M. and Hepsø, V. 2024. Unlocking the Potential of Big Data: Establishing System Trust Through Open Innovation Ecosystems, *Wiley Online Library,* (Open Access Research Article). https://doi.org/10.1111/radm.12734

Morkner, P., Bauer, J., Creason, C. G. et al. 2022. Distilling Data to Drive Carbon Storage Insights. *Computers & Geosciences*, **158**,104945, ISSN 0098-3004. https://doi.org/10.1016/j.cageo.2021.104945.

National Institute of Standards and Technology (NIST). 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Open File Report, NIST, Trustworthy and Responsible AI (NIST AI 600-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

North Sea Transition Authority. 2025. Information Reporting, Form and Manner of NDR Information, date of publication, April 2025, downloaded April 2025, https://www.nstauthority.co.uk/media/tetbcbto/information-reporting-form-and-manner-for-ndr-april-2025.pdf

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

Ringrose, P.S., Mathieson, A.S., Wright, I.W., Selama, F., Hansen, O., Bissell, R., Saoula, N. and Midgley, J. 2013. The In Salah CO2 Storage Project: Lessons Learned and Knowledge Transfer. *Energy Procedia*, **37:** 6226-6236, ISSN 1876-6102. https://doi.org/10.1016/j.egypro.2013.06.551

Schultes, E., Magagna, B., Hettne, K.M., Pergl, R., Suchánek, M., Kuhn, T. 2020. Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. *In:* Grossmann, G., Ram, S. (eds) Advances in Conceptual Modeling. ER 2020. Lecture Notes in Computer Science, **12584**. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13

Schultz, R., Beroza, G.C., and Ellsworth, W.L. 2021. A risk-based approach for managing hydraulic fracturing–induced seismicity. *Science* **372**: 504–507. https://scits.stanford.edu/sites/g/files/sbiybj22081/files/media/file/504.full__0.pdf

Seabra, G.S., Mücke, N.T., Silva, V.L.S., Voskov, D., and Vossepoel, F.C. 2024. AI enhanced data assimilation and uncertainty quantification applied to Geological Carbon Storage, *International Journal of Greenhouse Gas Control*, **136**: 104190, ISSN 1750-5836. https://doi.org/10.1016/j.ijggc.2024.104190

Stork, A., Baird, A.F., Horne, S.A., et al. 2020. Application of machine learning to microseismic event detection in distributed acoustic sensing data. *Geophysics*, **85** (5) 10.1190/GEO2019-0774.1. https://www1.gly.bris.ac.uk/~gljpv/PDFS/Stork_etal_2020_Geophys.pdf

Thavasimani, P. 2021. RapidXfer - Data transfer framework for square kilometre array. *Journal of Data, Information and Management*. **3**: 251–260. https://doi.org/10.1007/s42488-021-00055-1

Wierling, A., Schwanitz, V.J., Altinci, S. et al., 2021. FAIR Metadata Standards for Low Carbon Energy Research—A Review of Practices and How to Advance. *Sustainable Energies* **14** (20): 6692. https://doi.org/10.3390/en14206692

Wheelwright, J. 2024. Harnessing Value from the Open Source OSDU Platform By Data Visualization and Analysis. Society of Petroleum Engineers, *Seismic Techbyte* Aberdeen, https://www.spe-aberdeen.org/uploads/1040_TECHBYTE_Seismic24Wheelwright_DynamicGraphicsInc.pdf