Applied workflows for fiber optic sensing data from low carbon energy projects

Jess Kozman and Waris Warsi, Katalyst Data Management*

Summary

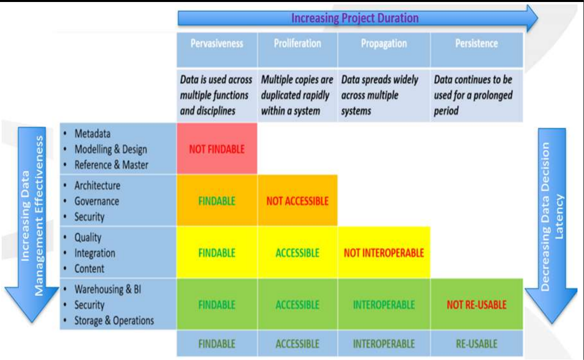
Data management practitioners in the low carbon energy sector are designing and deploying industry accepted optimum practices for large volumes of continuous and real time fiber optic sensing data. Volumes at petabyte scale can overwhelm even systems designed for what was previously considered Big Data from hydrocarbon exploration. Because of the long project lifecycles of low-carbon projects, the new practices will need to be future-proofed to support management, curation, and delivery of data types that have not been previously considered, for users that may not have been born yet, using technologies yet to be invented. Successful deployments are using edge computing, advanced data governance on the cloud, and optimized tiered storage capabilities to reduce cycle times in supporting business decisions.

Introduction

Government regulations for low carbon energy projects in various jurisdictions have started to catch up with available technologies for continuous real-time monitoring of geotechnical properties for geologic energy extraction and storage. Part of the attention to data submission requirements has been driven by the growth of distributed fiber optic sensing (DFOS) data for measurement, monitoring, and verification (MMV) activities (Mondanos et al., 2024). Modern fiber optic sensor deployments have the capacity to record multiple physical parameters including acoustic energy, temperature, and strain. Measurements can be made continuously in real time at frequencies up to 2kHz and at densely sampled surface and subsurface locations, with sub-meter resolutions down to 25cm. Low risk of failure and minimized maintenance requirements make DFOS an attractive option for instrumenting locations with expected project lifetimes of decades, providing lower operating costs over time compared with multiple repeated conventional survey. In many cases, DFOS systems provide improved signal-noise ratios over geophones, hydrophones, or permanently installed downhole gauges. Advances in engineered cables have shown further noise floor improvement over standard distributed acoustic sensing (DAS) cables (Diaz-Meza et al., 2023). Installation of fiber optic cables behind subsurface casing in wells can reduce the costs and risks of well intervention work, and DFOS measurements can be collected from previously installed fiber used for communications on surface trenched cables or seabed umbilicals, providing unique geometries. Longevity and long-life performance of optical fibers in helically wound downhole cables indicates they can be used throughout the asset lifecycle, from site assessment, through injection and storage, and into post-injection site monitoring (Li et al., 2018).

Theory and Method

Recent examples of datasets evaluated or delivered for storage and management, and public domain open file datasets available from low-carbon energy projects, have been examined to determine the applicability of existing industry accepted optimum data management practices. We focused on embedded data workflow methodologies and data fit organization implementation profiles (Li et at., 2023) optimized for cloud native, open source, technology agnostic and standards-based enterprise scale data platforms. These data ecosystems serve as systems of engagement for emerging data technologies, and support ingestion, enrichment, and consumption of data that is findable, accessible, interoperable and reusable (FAIR) to support data fit and driven organizations delivering democratized technical data (Sidik, 2023).



	Pervasiveness	Proliferation	Propagation	Persistence
	Data is used across multiple functions and disciplines	*Multiple copies are duplicated rapidly within a system*	*Data spreads widely across multiple systems*	*Data continues to be used for a prolonged period*
• Metadata • Modelling & Design • Reference & Master	NOT FINDABLE			
• Architecture • Governance • Security	FINDABLE	NOT ACCESSIBLE		
• Quality • Integration • Content	FINDABLE	ACCESSIBLE	NOT INTEROPERABLE	
• Warehousing & BI • Security • Storage & Operations	FINDABLE	ACCESSIBLE	INTEROPERABLE	NOT RE-USABLE
	FINDABLE	ACCESSIBLE	INTEROPERABLE	RE-USABLE

Figure 1: The relationship between data management effectivness and data decision latency for DFOS data exhibiting the four behaviors of large data on the cloud.

We considered how industry standard knowledge areas and environmental elements from the recognized data management body of knowledge (Dama International, 2017) will contribute to success of data management implementations as data behaviors evolve over time (Figure 1). For this analysis we considered data behavior dimensions that are most critical to DFOS data curation requirements, over the time duration of low energy projects. While existing workflows have been developed around the defining characteristics of Big Data, usually considered to be volume, variety, and velocity, we focused on how data behaves over time on cloud native platforms, and how it supports the use of emerging data technologies such as generative algorithms, large language models, augmented reality, and machine knowledge (GLAM). In this context, the dimensions of pervasiveness, proliferation, propagation, and persistence evolve over time for large data on the cloud, from the intersection of the primary dimensions of Big Data (Figure 2).

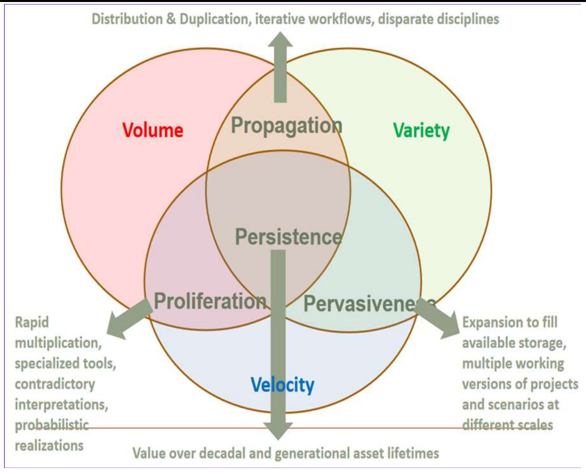


Figure 2: How the traditional dimenstions of Big Data intersect to describe present and future behaviors of DFOS digital data on emerging data curation platforms.

Our case studies have shown that for project lifetimes that could eventually exceed 100 years (Romanak and Dixon, 2022), measurable reductions in data decision latency can be achieved only by effective application of all key knowledge areas of the data management body of knowledge, applied in a logical sequence as the behaviors of the data evolve.

Examples

Applications driving acquisition of fiber optic data include continuous passive micro-seismic monitoring for baseline seismicity, active monitoring during injection for induced seismicity and reactivation of existing faults and fractures, site specific characterization of strain and temperature baselines, real-time monitoring of injection activity and well integrity, and subsurface gas and fluid plume distribution using vertical seismic profiling, cross well tomography (Wuestefeld and Weinzierl, 2020), full 3D and 4D seismic acquisition, and in-situ strain, pressure, and gas venting detection and analysis (Su et al., 2021). Data has been collected from a variety of onshore and offshore environments, in conjunction with temporary, moveable, and permanent sources, and alongside conventional surface and downhole receiver arrays.

Low carbon projects utilizing distributed fiber optic sensing include exploration for geothermal, period 1 gases and critical minerals, extraction of geologic energy carriers and materials for low energy projects, carbon capture, utilization and storage (CCUS), geologic energy storage, monitoring of disposal sites for low-carbon energy byproducts, and optimization of renewable energy equipment and infrastructure.

Early volumes submitted to data management and curation platforms from DFOS tended to be micro-seismic detection data collected from fluid injection or temperature data used to identify and locate well integrity issues. This allowed data management practitioners to develop workflows and methodologies for selecting and prioritizing meta-data attributes that allow for search and access of this data associated with both the point location of wells and polygon project areas on map-based user interfaces.

Indexing of large volume data sets (more than 100 TB per delivery) becomes more challenging when data is collected from otherwise “dark” fiber optic cables installed on surface equipment, in underground infrastructure, or on a subsea umbilical. These fibers can be illuminated to collect strain data to monitor the structural health of marine structures associated with renewable power installations (Silva et al., 2023). In these cases, the data may not be strictly associated with a single previously indexed well or field location, necessitating the creation of project definitions that utilize faceted taxonomies to include facility and geo-contextual information (Michiorri et al., 2022). The embedded data workflow methodology provides auditable and consistent processes for adding, storing, and retrieving this metadata.

Our most robust embedded workflows have been developed for the volumes of data available from strain-rate measurements in geothermal fields (University of Wisconsin, 2017), passive micro-seismicity monitoring from commercial scale geologic storage of carbon dioxide (Kavin, 2021), and continuous carbon dioxide plume migration path monitoring from deviated deep wells (Isaenkov et al., 2021). Raw acquired data volumes from these projects can range up to 100’s of TB per day from continuously operated permanently installed surface orbital vibrators. Key methodologies employed to manage this data on cloud platforms have included high-capacity edge processing at the project site, to provide down-sampling in frequency and spatial domains at the interrogation unit. Typical re-sampling can reduce sampling frequency from tens of kHz to single kHz, and spatial sampling from sub-meter to single meter. Further software decimation can reduce data volumes by another order of magnitude to 2ms sample rates compatible with industry standard geophysical formats, and up to multiple meter channel spacings. Source deconvolution can reduce record lengths from 100’s of seconds of listening time to less than 10 seconds and stacking of multiple source signatures at a single location, and source-receiver pairs can finally yield a GB range dataset that can be transmitted over an internet connection or high-speed satellite link. This entire sequence can generate data volume reduction factors of up to 3 million ($3 \cdot 10^6$), to provide actionable data to a remote operation center in real time, while preserving original and intermediate data volumes at the project site as needed for government submission and regulatory requirements. In some cases, the original volumes are collected onsite on solid-state drives or cloud storage devices and shipped to nearby cloud access points-of-presence or data availability regions and then ingested for storage on intelligently tiered, low-cost and higher latency storage levels for future retrieval when required by government, joint venture partner, or regulatory audits (Figure 3). Note that final migrated volumes for a single vintage of acquisition can be on the order of tens of megabytes in industry standard processed formats. The embedded workflow is designed to make these volumes available as quickly as possible, with auditable lineage and technical assurance meta-data associated through the data lifecycle.

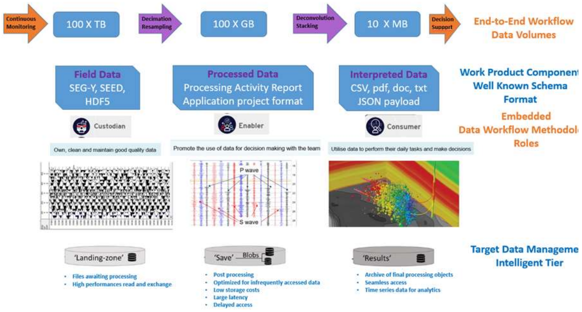


Figure 3: Example of an end-to-end workflow supporting large volumes of DFOS digital data on data platforms

For continuous injection plume monitoring applications including carbon dioxide, compressed air, or even by-products from low-carbon nuclear power sources, detection of events can be used as the trigger for producing intermediate processing volumes. This may require edge or remote processing to employ bandpass filtering, and standard algorithms for phase arrival estimation and event detection based on energy density ratios at coincident channel triggers. The resulting wavelet snippets used for stacking, beamforming, and/or geolocation are treated as processing activity work product components in the embedded data workflow methodology, and assigned to an intelligent storage tier that reflects their behavior on binary large object (BLOB) based data ecosystems.

Finally, the actionable reports as required by internal operating asset teams, joint venture audits, or government regulators or environmental protection agencies, can be stored as structured data sets on a high availability tier where they can be accessed by data science and analytic tools. By utilizing open source, industry standards based, and cloud provider agnostic application programming interfaces (API's), the embedded data workflow methodology aims to reduce the amount of future work and retooling required when new machine and compute enhanced technologies become available. It also provides entitlements and obligations metadata attached at a work product component level to adapt to future cybersecurity and penetration threats that will come with future storage and delivery methodologies, utilizing the three capitals of the data fit organization implementation profile (Emmanuel and Klaus, 2024).

The data governance frameworks required to successfully support a sustainable data workflow methodology for large volumes of digital DFOS data on future data ecosystems, platforms, and systems of engagement have proved difficult in the past to implement using top-down proscriptive policies and procedures (Jian et al., 2022). The embedded data workflow methodology instead uses self-identified data roles and responsibilities to identify candidate participants in the enterprise level data workflows and map the appropriate level of data governance to a simplified three level framework that reflects an emphasis on FAIR data implementation profiles and collection of quantifiable metrics for reductions in data decision latency. Where this methodology has been deployed for large volumes of geotechnical data on cloud platforms, organizations have previously reported reductions in cycle times, decreased data related risk, and improved quality of business decisions. These results correlate well with early experiences using similar data workflow methodologies for DFOS data volumes.

Conclusions

Effective curation of DFOS data in sustainable embedded data workflow methodologies will be a requirement for future proofing data ecosystem systems of engagement as low carbon projects enter decades-long site care periods. The application of existing industry accepted optimum practices for data management will provide the flexibility required to accommodate new data types, new data analysis and interpretation technologies, and new digital data storage platforms. Recent successes in adopting and adapting existing open-source work product components, ingestion schema, and data definitions from workflows for hydrocarbon extraction have demonstrated the efficacy of this strategy. Our development teams documented a 64% reduction in time and effort required to implement an end-to-end data ingestion workflow for critical minerals, by re-using tools and methodologies from energy wellbore and geophysics implementations. These projects in turn are contributing to a 37% overall reduction in data decision latency over the course of a project lifecycle for other critical mineral operators deploying the same technologies (Figure 4).

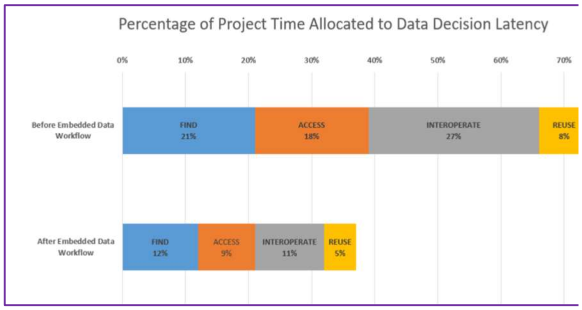


Figure 4: Aggregated results of data decision latency metrics from low-carbon projects deploying embedded data workflow methodologies. Applying time value of information based on fully burdened costs for geotechical full-time equivalent personnel yielded an average time to positive return on investment of 5.1 months.

Acknowledgements

The authors wish to acknowledge the global subject matter experts, consultants, and operational staff of Katalyst Data Management who contributed to development of the practices described in this paper. These groups operate out of offices in Canada, the United States, Malaysia, the United Kingdom, Brazil, and Australia.