# Large Scale Real Seismic Data Curation from North Sea for Generative-AI Models

Ipsita Bhar[1], Huseyin Tuna Erdinc[1], Thales Souza[1], Charles Jones[2] and Felix J. Herrmann[1]
[1]Georgia Institute of Technology, Atlanta, USA [2]Osokey Ltd, Henley-on-Thames, UK

March 27, 2026

## 1 Abstract

Understanding subsurface structures is essential for applications such as energy exploration, underground storage, and environmental monitoring. Seismic inversion seeks to estimate subsurface properties from seismic observations, but the problem is inherently ill-posed, leading to non-unique and uncertain solutions. Although machine learning has accelerated inversion workflows, its progress is constrained by the lack of realistic, publicly available datasets, as most high-quality seismic data remain proprietary. Existing synthetic datasets such as OpenFWI enable large-scale training but fail to capture the complexity and variability of real subsurface conditions. To address this limitation, we develop an automated data curation pipeline that transforms raw field data into machine learning–ready datasets for generative AI workflows in seismic inversion. The pipeline leverages publicly available North Sea surveys from the UK National Data Repository and integrates seismic volumes with well-log data. Since migrated seismic data are often available in the time domain while well logs are in depth, the pipeline performs time-to-depth conversion using checkshot measurements to estimate average velocities, followed by spatial interpolation to construct smooth velocity fields. Quality control is ensured by validating interpolated velocities against well measurements. The pipeline further generates depth-domain quasi-2D seismic sections from curated 3D volumes by extracting sections along trajectories passing through well locations. These sections are resampled using 2D Fast Fourier Transform (FFT) to preserve low-frequency content while suppressing noise, and well logs are processed to match seismic resolution. The final dataset consists of aligned seismic sections and well data stored in a standardized HDF5 format for training of generative models for uncertainty-aware seismic inversion.