

Automatic Geo-AI Workflow for Stress Information Extraction: Detecting, Digitizing, Harmonizing and Validating Measurements from Legacy Documents for the Stress Database Netherlands

M.D. Demmers¹, Anniek Theuwissen¹, Hamed Ali Diab-Montero¹, Loes Buijze¹, and Jurrien Dijk¹

¹ TNO – Geological Survey of the Netherlands, Utrecht, The Netherlands

Large volumes of subsurface data from the oil and gas domain remain locked in unstructured and semi-structured document formats, including legacy reports, scanned archives, and heterogeneous digital files. Within national subsurface data infrastructures, many of these documents originate from historical operator submissions and paper archives that were never systematically inventoried or digitized, an issue that becomes more pressing as operators exit mature basins and custody of legacy archives changes. The absence of structured inventories and standardized ingestion workflows creates a substantial risk of irreversible data loss and severely limits reuse potential for emerging subsurface applications such as geothermal energy and carbon capture and storage (CCS).

A particularly valuable subset of these materials concerns subsurface stress data, which is essential for safe and optimal drilling and for subsurface safety and integrity. Formation Integrity Tests (FIT) and Leak-Off Tests (LOT) are a key data stream, but their pressure–flow information is typically embedded in tables, plots, and figures across a wide range of technical documents. While this information has traditionally been interpreted manually, the scale and heterogeneity of the document landscape make such approaches increasingly infeasible. Moreover, national-scale, data-driven analysis requires that FIT/LOT measurements are not only extracted, but also harmonized, validated, and stored consistently before any quantitative interpretation can take place.

The objective of this work is to develop and demonstrate an end-to-end automated framework that transforms unstructured subsurface pressure and stress information from heterogeneous documents into a unified, database-ready form, so that results can be appended to the Stress Database Netherlands and used in a consistent way by stakeholders. The focus is on scalable extraction, standardization, and validation of tabular and graphical stress data, thereby creating the necessary foundation for subsequent data-science-driven analysis.

We implement these capabilities in a Human-in-the-Loop (HITL) platform in which data management plays a central role and computer vision (CV) accelerates the process. Relevant objects in unstructured documents are detected using deep-learning-based object detection, and data managers work closely with AI systems: accreditations of detected attributes and corrections are captured as valuable training data to continuously improve the models. The workflow finds relevant objects, digitizes the values within them, standardizes extracted values into defined schemas, validates records against schema constraints and expert bounds, and subsequently supports (automated) interpretation to derive secondary metrics that are commonly used by stakeholders. Object detection is performed with two CV models for locating tables containing pressure–flow measurements and one model for locating figures where pressure–flow curves are plotted. Digitization leverages established approaches: OCR and out-of-the-box tools are used to digitize tables and extract values from graphics, after which standardization fits extracted values per object to the defined schema and automated validation is applied, with anomalies flagged for expert review.

The workflow produces a structured, validated dataset of subsurface pressure measurements extracted from unstructured documents. Each record includes standardized attributes such as well identifier, test type (e.g., FIT or LOT), measured pressures, depths, and provenance metadata (document source, page, and figure/table reference). Tables and figures are digitized into machine-readable formats, extracted values are harmonized according to the schema, and automated validation ensures that values are consistent with expected physical ranges and formats. Overall, the framework provides an end-to-end approach for transforming heterogeneous subsurface documents into a unified, validated dataset of stress measurements, reducing reliance on manual interpretation, mitigating the risk of data loss from legacy archives, and enabling consistent, repeatable workflows that support integration into the Stress Database Netherlands and scalable downstream analyses of wellbore integrity and drilling operations.