

---

# Un modèle d'intelligence artificielle au service de l'extraction des descriptions géologiques dans les documents numériques décrivant les forages

Laure Capar<sup>\*†1</sup>, Musaab Khalid<sup>‡2</sup>, Sébastien Hameau<sup>§1</sup>, Nicolas Gilardi<sup>¶1</sup>, and François Prognon<sup>||1</sup>

<sup>1</sup>Bureau de Recherches Géologiques et Minières – Bureau de Recherches Géologiques et Minières (BRGM) – France

<sup>2</sup>SOTATECH – SOTATECH – France

## Résumé

Le BRGM, service géologique national français, a pour une de ces missions de mettre à disposition sous une forme appropriée les informations couvrant le territoire national ainsi que le plateau continental, parmi lesquelles celles concernant les fouilles, forages et leviers géologiques recueillis en application ducode minier, depuis le début des années 60. Le BRGM gère la base de données du sous-sol (BSS), qui contient un certain nombre d'ouvrages souterrains, dont environ 150 000 sont des forages avec des documents numériques et dont la donnée est sous format non structuré. Ces documents peuvent contenir les descriptions géologiques avec les profondeurs des toits et murs des formations géologiques traversées par le forage. Ils peuvent être manuscrits ou non manuscrits (écrits à l'aide de normographes, tapés à la machine à écrire, ou issus de logiciels). Le dépouillement et la mise au format structuré de l'ensemble de ces documents nécessiteraient l'intervention d'une personne sur une quinzaine d'années. Avec le développement de l'intelligence artificielle, le BRGM a entrepris avec le soutien du projet ANR-22-EXSS-0005A - PEPR sous-sol - PC4 - Digital Earth, d'effectuer un projet de recherche sur l'extraction des descriptions géologiques avec les profondeurs des toits et murs des formations géologiques traversées par le forage, par intelligence artificielle. Le projet se décompose en trois étapes principales avec la production de modèles d'intelligence artificielle. La première étape a consisté à identifier les documents manuscrits de ceux qui étaient non manuscrits, avec un taux de succès de 85%. La seconde étape a permis d'identifier sur les documents non manuscrits, les pages sur lesquelles se trouvaient une description géologique, avec un taux de succès de 98%. La troisième étape est celle de l'extraction des informations souhaitées sous format structuré. Pour cela le modèle Qwen 2.5 VL donne de bons résultats avec des extractions quasiment correctes à 100% sur les documents donc les limites des couches géologiques sont explicitement mentionnées. Pour les autres documents donc les limites des couches géologiques ne sont pas explicitement mentionnées, le modèle Qwen 2.5 VL permet l'extraction des informations avec un taux de confiance moindre. Concernant les documents manuscrits, l'extraction des informations restent difficiles.

---

\*Intervenant

†Auteur correspondant: l.capar@brgm.fr

‡Auteur correspondant: contact@sotatech.fr

§Auteur correspondant: s.hameau@brgm.fr

¶Auteur correspondant: n.gilardi@brgm.fr

||Auteur correspondant: f.prognon@brgm.fr

**Mots-Clés:** Intelligence artificielle, forages, Extraction des descriptions géologiques