Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Application de Transformers multimodaux à l'extraction d'informations des documents de sondage des sols

Stanislas Bagnol[1, 2] Killian Barrere[2] Véronique Eglin [2]
Elöd Egyed-Zsigmond [2] Jean-Marie Côme[1] David Pitaval[1]
(1) Ginger BURGEAP, 69003 Lyon, France
(2) INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne, France
s.bagnol@groupeginger.com, (2) {prenom}.{nom}@insa-lyon.fr

RÉSUMÉ

L'extraction d'information de documents complexes est un domaine de recherche qui bénéficie d'une très grande attention tant dans la littérature, que dans l'industrie dans le cadre de la digitalisation des données. Les Transformers et leurs adaptations ont très largement contribué à faire progresser cette recherche en s'appuyant sur des modèles de langue qui ont introduit une compréhension sémantique de l'organisation de la structure des documents. Les coupes de sondage sont des documents industriels complexes et riches en informations, pour lesquels aucune solution d'extraction d'informations n'avait été proposée. Nous montrons les limites des approches de bout-en-bout par des expérimentations avec le modèle DONUT. Comme alternative, nous proposons une chaîne de traitement hybride reposant sur le *fine-tuning* de Transformers multimodaux et des algorithmes heuristiques. Nous comparons deux architectures de Transformers multimodaux pré-entrainés : BROS et LayoutLMv3.

ABSTRACT

Multimodal Transformers Application to Soil Probing Log Documents Information Extraction.

Complex document Information Extraction is a reasearch field that received a lot of attention in scientific literature as well as in the industry, as part of document digitalization. Transformers and their adaptations has shown great progresses relying on langage models, introducing documents layout semantic understanding. Soil probing log documents are complex and information rich, and no information extraction method has yet been proposed. We show the limits of end-to-end methods throught experiments with DONUT model. As an alternative, we propose a hybrid pipeline based on Multimodal Transformer fine-tuning and heuristic algorithm. We compare two pre-trained Multimodal Transformer architectures : BROS and LayoutLMv3.

MOTS-CLÉS : Extraction d'informations de documents, Transformers multimodaux, Algorithme d'annotation, Méthodes de bout-en-bout.

KEYWORDS: Document Information Extraction, Multimodal Transformers, Labeling Algorithm, End-To-End Methods.

1 Introduction

La compréhension et l'extraction d'informations de documents est un domaine de recherche d'actualité, tant dans les milieux académiques que dans les milieux industriels. Récemment, les modèles

de Transformers multimodaux pré-entraînés ont obtenu de bons résultats sur des tâches applicatives spécifiques. Cependant, ces modèles peinent à généraliser leurs apprentissages sur de nouvelles tâches ou des collections de documents qu'ils n'ont jamais vues. L'adaptation de ces modèles sur des collections spécialisées, par *fine-tuning*, est souvent nécessaire.

Les coupes de sondage sont des documents largement utilisées dans l'industrie géotechnique et environnementale pour caractériser la composition des sols en profondeur. Malgré leur utilisation répandue, aucune solution efficace d'extraction automatique d'informations adaptée spécifiquement à ces documents industriels n'existe à ce jour. Ces documents (cf. Figure 1) sont par nature complexes en raison de leur grande diversité de mise en page, de leur contenu riche et hétérogène, ainsi que des variations de qualité liées à leur date de création et à leur origine. La compréhension de ces documents repose sur une analyse détaillée de leur structure, impliquant une combinaison d'informations à la fois sémantiques et visuelles. L'extraction d'informations (lithologies associées à leur profondeur, titre, identifiant, position GPS (x, y), altitude, ...) nécessite des modèles capables de traiter simultanément texte et mise en page. Contrairement aux jeux de données classiques en analyse et reconnaissance de documents, les coupes de sondage présentent des spécificités fortes, comme par exemple la présence d'échelles de profondeur proportionnelles, qui rendent l'adaptation des méthodes existantes peu triviale. L'absence de jeu de données annotées constitue une problématique majeure.

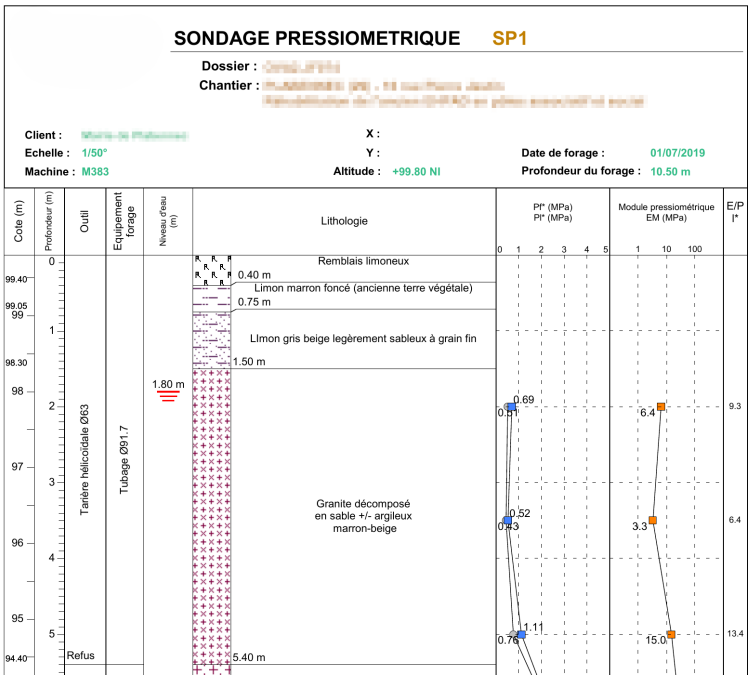


FIGURE 1 – Coupe de sondage tronquée, montrant 4 couches de lithologie (e.g. "*remblais limoneux*" entre 0 et 0,4 mètres de profondeur), une échelle de profondeur sur la gauche, et un identifiant (*SP1*).

Dans cette étude, nous nous focalisons sur les couples lithologie-profondeur qui constituent l'information centrale des coupes de sondage. Ils peuvent être difficiles à extraire car la profondeur d'une lithologie n'est pas toujours explicitement mentionnée dans le texte (e.g. Figure 3a) ; il convient alors de la déduire à l'aide de l'échelle de profondeur. D'autres informations, telles que la profondeur du sondage, la position GPS (x, y) ou l'altitude, peuvent être extraites efficacement grâce à leur structure clé-valeur. Nous avons choisi de ne pas nous attarder sur celles-ci car des solutions efficaces existent déjà dans la littérature.

Nos contributions peuvent être résumées de la manière suivante :

— Nous proposons une chaîne de traitement complète partant de l'image d'un sondage jusqu'à l'extraction d'informations structurées. Elle est composée de 3 parties principales :
    1. un OCR (Optical Character Recognition) pour extraire texte et boîte englobante associée ;
    2. un Transformer multimodal *fine-tuné* (basé sur les modèles LayoutLMv3 et BROS, que nous comparons dans cet article) afin de classifier les mots reconnus par l'OCR ;
    3. et d'un algorithme heuristique pour associer une profondeur à chaque lithologie.
— Nous proposons aussi une méthode d'annotation automatique des documents de sondage.

2 Etat de l'art

L'extraction d'informations de documents a bénéficié d'avancées significatives grâce à l'utilisation de mécanismes d'attention et d'architectures de Transformers. On distingue deux types d'approches : les méthodes reposant sur un OCR externe, et les méthodes de bout-en-bout intégrant la tâche d'OCR.

Les Transformers multimodaux, ont entraîné des améliorations remarquables des techniques d'extraction d'informations de documents complexes. Ils reposent sur un OCR externe qui fournit au modèle les mots et leurs positions. LayoutLM (Xu *et al.*, 2020) intègre des informations de mise en page (i.e. positions 2D des mots dans la page) dans une architecture de Transformer. Ensuite, LayoutLMv2 (Xu *et al.*, 2021) et LayoutLMv3 (Huang *et al.*, 2022) intègrent une représentation visuelle de l'image du document permettant d'encoder la structure de documents complexes comme des tableaux. LayoutLMv3 utilise une projection linéaire de patchs de l'image (i.e. sous-images) qu'il transmet directement au Transformer multimodal. BROS (Hong *et al.*, 2022) utilise à la fois les positions absolues et relatives entre les mots, ce qui permet de mieux capturer les représentations de structures de types clés-valeurs. LayoutLMv3 et BROS sont pré-entraînés sur des tâches prétextes auto-supervisées et ont l'avantage de distribuer leurs poids.

De leur côté, les Visual Transformers (VT), qui ne reposent sur aucun système d'OCR, offrent des perspectives prometteuses. Ce sont des modèles de vision par ordinateur adaptant l'architecture des Transformers aux tâches de vision. Ils utilisent des mécanismes attentionnels pour combiner les représentations textuelles et visuelles du document. DONUT (Kim *et al.*, 2022) utilise une architecture de VT comme encodeur. Pré-entraîné sur une tâche de pseudo-OCR, DONUT utilise un décodeur génératif (Lewis *et al.*, 2019) pour produire des réponses structurées et peut être *fine-tuné* sur des tâches de classification, d'analyse de documents ou de réponse aux questions. Ces approches de bout-en-bout limitent la propagation des erreurs d'OCR dans les prédictions du modèle.

Les jeux de données standards pour l'extraction d'informations de documents sont généralement des factures, formulaires ou tickets de caisse comme FUNSD (Jaume *et al.*, 2019), CORD (Park *et al.*, 2019) ou SROIE (Huang *et al.*, 2019). On retrouve aussi des jeux de données pour la réponse aux questions tels que DocVQA (Mathew *et al.*, 2021). Ces documents se distinguent des coupes de sondage par leur mise en page et leur contenu, ce qui limite leur exploitation pour notre étude. Les connaissances acquises par les modèles pré-entraînés ne sont pas directement transférables à notre problématique sans tâche d'apprentissage supplémentaire.

Bien que les Transformers multimodaux puissent être *fine-tunés* sur des tâches d'extraction de relations (e.g. avec l'ajout d'un decodeur SPADE (Hwang *et al.*, 2021) à BROS) pour lier les lithologies à une profondeur, il ne peuvent déduire la profondeur depuis l'échelle (cf. Figure 3a). Les VT génératifs (e.g. DONUT) peuvent être utilisés pour générer des séquences de textes contenant à la fois la

lithologie et sa profondeur associée (cf. alternative 2. Figure 2). En revanche, ils sont incapables de déterminer la profondeur d'une lithologie en se reportant à l'échelle. Lorsque la profondeur n'est pas explicitement mentionnée, DONUT essaie de générer une valeur de profondeur sans raisonnement logique, probablement en se basant sur les données d'apprentissage.

Les limites des méthodes actuelles nous ont amenés à concevoir une chaîne de traitement hybride en 3 étapes, combinant modèles d'apprentissage automatique et algorithmes heuristiques.

3 Pipeline d'extraction des couples lithologie-profondeur

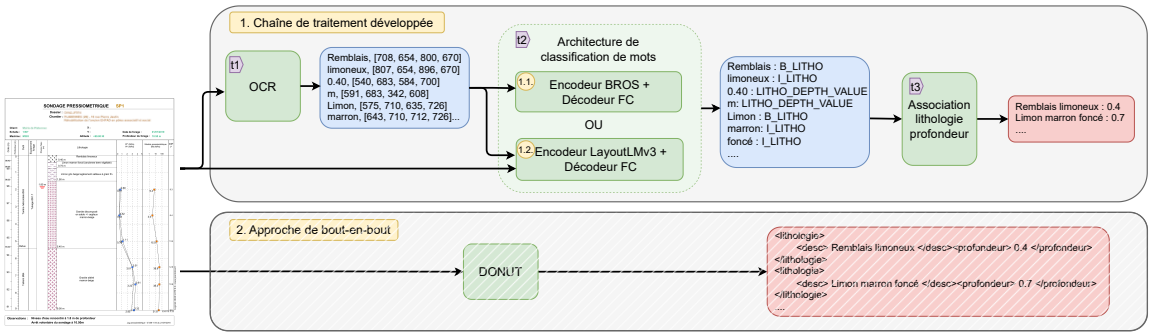


FIGURE 2 – Comparaison de deux méthodes d'extraction des couples lithologies profondeur. (1) L'approche en 3 étapes que nous proposons : OCR (t1), classification de mots en utilisant les Transformers multimodaux (2 alternatives : BROS et LayoutLMv3) (t2) et un algorithme d'association des lithologies et profondeurs (t3). (2) Une approche de bout-en-bout via le modèle DONUT.

Les approches de bout-en-bout utilisant des architectures de Visual Transformers (alternative 2. Figure 2) offrent des perspectives intéressantes pour notre problématique. DONUT peut générer en une seule étape des séquences de texte structurées contenant l'information sur les lithologies et leur profondeur à partir d'une image. Cette méthode limite la propagation des erreurs, d'OCR notamment, tout au long de la chaîne de traitement et simplifie l'annotation de documents.

Pourtant, nos expériences préliminaires indiquent que DONUT *fine-tuné* n'est pas fiable pour l'extraction des couples lithologie-profondeur. Lorsque les profondeurs ne sont pas explicitement indiquées, le modèle tente alors d'inventer une profondeur. Ces hallucinations entraînent un écart moyen de 15.55 mètres, ce qui est disproportionné comparé aux sondages qui ont généralement des profondeurs comprises entre 2 et 4 mètres.

Ces résultats nous ont amenés à privilégier une approche hybride (cf. alternative 1. Figure 2) en 3 étapes : (t1) un système d'OCR qui extrait le texte et sa position, (t2) une architecture de classification de mots basé sur un Transformer multimodal et (t3) un algorithme heuristique d'association des lithologies et profondeurs.

3.1 OCR

L'extraction initiale du texte et de sa localisation spatiale (étape t1) est réalisée via la librairie DocTR (Mindee, 2021) avec les modèles db_resnet50 (pour la détection) et crnn_vgg16_bn (pour la

reconnaissance des mots), sélectionnés pour leur efficacité sur des textes courts et isolés typiques de notre corpus. Après comparaison, ces modèles se sont révélés supérieurs à PaddleOCR, Tesseract ou TrOCR en précision et rapidité. Bien que nous n'ayons pas pu mesurer directement la performance isolée de l'étape OCR faute d'un jeu de référence annoté spécifique, son efficacité est indirectement démontrée par les performances globales obtenues lors de l'étape suivante de classification des mots (expérimentée à la Section 5.2), puisque celle-ci repose également sur les résultats d'OCR.

3.2 Architecture de classification de mots

L'architecture de classification de mots (étape t2) proposée comprend un encodeur, fondé sur un Transformer multimodal pré-entraîné, ainsi qu'un décodeur entièrement connecté. Dans ce travail, nous avons retenu LayoutLMv3 pour sa capacité à intégrer efficacement le texte et les informations visuelles (notamment via ses représentations basées sur des patchs d'images), ainsi que BROS pour son aptitude à capturer les relations spatiales clés-valeurs spécifiques aux coupes de sondage.

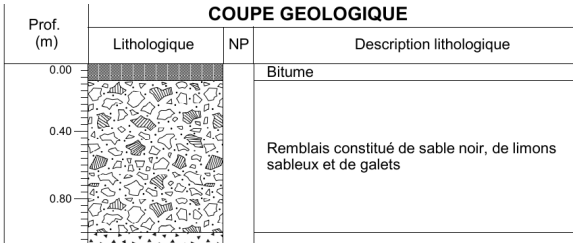
Nous avons effectué un *fine-tuning* de notre architecture sur une tâche de classification de mots, qui consiste à assigner un label à chaque mot. Nous avons 10 labels différents : `Other`, `Title`, `Key_Word`, `End_Depth`, `Depth_Scale`, `XYZ` (i.e.position GPS (x, y) + altitude), `Litho_Depth_Value`, `B_Litho`, `I_Litho`. Pour l'identification des lithologies, nous avons adopté une approche inspirée du BIO Tagging (Hwang *et al.*, 2019; Hong *et al.*, 2022), très utilisée dans les tâches d'extraction d'entités nommées (NER). Nous avons ainsi défini deux catégories spécifiques : `B_Litho` (*Beginning*) pour marquer le début de chaque description lithologique, et `I_Litho` (*Inside*) pour identifier les mots qui constituent la suite de cette description.

3.3 Algorithme heuristique d'attribution de profondeurs aux lithologies

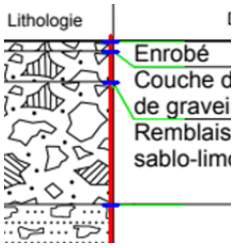
Les prédictions de l'architecture de classification de mots permettent de récupérer les descriptions lithologiques ainsi que les mots de profondeur lithologiques (label `Litho_Depth_Value`). L'objectif de cette étape est d'attribuer une valeur de profondeur à chaque description lithologique (étape t3). Bien que cette tâche d'association lithologie-profondeur puisse s'apparenter à un problème d'Entity Linking ou d'extraction de relations, nous avons opté pour une approche heuristique afin d'exploiter pleinement les informations visuelles et spatiales implicites présentes dans les documents. À partir des positions des lithologies, nous localisons leur colonne par analyse de l'histogramme horizontal, puis nous détectons les traits horizontaux séparateurs via dilatation morphologique et seuillage adaptatif. Chaque lithologie est ensuite associée au trait de séparation le plus proche situé en-dessous, ce qui permet de lui attribuer une valeur de profondeur explicite si celle-ci se trouve entre ces deux éléments. De plus, nous reportons ces traits de séparation sur l'échelle, afin de déterminer les valeurs de profondeur par interpolation.

Dans certains cas le trait de séparation peut être décalé pour permettre l'intégration du texte (cf. Figure 3b). Par conséquent, nous avons développé une fonctionnalité de suivi de trait qui permet de suivre le trait jusqu'à la bordure de la colonne (i.e. trait rouge) sur l'image binarisée en faisant des pas de pixel noir en pixel noir vers la gauche, en haut ou en diagonale (haut gauche). On calcule alors la profondeur du point de contact (trait bleu) par rapport à l'échelle.

Une fois les profondeurs explicites et interpolées (à l'aide de l'échelle) obtenues, nous comparons ces valeurs : en cas de divergence entre les deux, nous privilégions la valeur calculée, jugée plus fiable.



(a) Profondeur implicite à déduire de l'échelle.



(b) Module de suivi de trait.

FIGURE 3 – Exemple et cas particulier de profondeur lithologique à déduire.

4 Constitution d'un jeu de données d'entraînement

4.1 Algorithme d'annotation automatique de documents de sondage

Nous avons développé un algorithme heuristique d'annotation qui a permis d'annoter 200 coupes de sondage de notre collection de documents internes. Cet ensemble de données initial est important pour entraîner une première version de l'architecture de classification de mots. L'algorithme est basé sur un ensemble de règles et de caractéristiques communes aux différents formats de sondage.

Afin d'identifier les zones du document contenant les descriptions lithologiques, les mots reconnus par l'OCR (présenté dans la Section 3.1) sont, après application de *stemming*, comparés avec une liste de termes lithologiques courants (e.g. limon, remblais, argile, ...). La case du tableau de sondage englobant les mots faisant référence à une lithologie (cf. Figure 1) est ensuite déterminée en appliquant un algorithme de détection de segments (Grompone Von Gioi *et al.*, 2012) qui vise à identifier ses bordures dans toutes les directions. L'ensemble du texte contenu dans une case et ensuite récupéré (par exemple dans la Figure 1 : "*Limon marron foncé (ancienne terre végétale) 0.75m*"), et séparé en nom de lithologie et profondeur. L'échelle de profondeur représentant une information importante pour identifier la profondeur d'une lithologie par interpolation, nous la localisons de manière automatique grâce à la présence de valeurs numériques ordonnées et alignées verticalement sur la gauche du document. Nous repérons également des données de l'en-tête de type clé-valeur (e.g profondeur du sondage, date, position (x, y),...) grâce à des listes de mots clés (e.g. "*Sondage pressiométrique*") ou à une expression régulière pour l'identifiant (e.g. "*SP1*").

L'ensemble des annotions sur les 200 coupes de sondages ont été vérifiées par un humain, et évalué à une précision de 93% pour la tâche de classification des mots. Nous avons estimé que ces résultats étaient suffisants pour entraîner un modèle, avec l'objectif qu'il puisse généraliser ses apprentissages et corriger les erreurs d'annotation.

4.2 Augmentation du jeu de données par entraînements et prédictions itératifs

Nous avons employé une méthode itérative d'apprentissage et d'annotation de documents par prédiction, visant à enrichir notre jeu de données et améliorer les performances du modèle. Nous avons entraîné un première version du modèle de classification de mots grâce au premier jeu de données annotées algorithmiquement (Section 4.1). Ses prédictions sur de nouveaux lots de documents ont été vérifiées et incorporées au jeu de données afin d'entraîner une nouvelle version du modèle. Nous avons réitéré ce processus plusieurs (4) fois jusqu'à traiter les différents formats de documents connus.

Ainsi, notre jeu de données est constitué de 874 coupes de sondages de formats variés issues de différentes périodes et origines. Nous avons utilisé 699 documents pour l'entraînement du modèle final et 175 pour son évaluation.

5 Expérimentations et résultats

5.1 Configuration des modèles

Pour les modèles d'extraction d'informations, nous avons utilisé un Optimizer Adam avec un taux d'apprentissage de $5e^{-5}$. Nous entraînons ces modèles sur 5 itérations et une taille de batch de 1. Pour le modèle utilisant un encodeur LayoutLMv3, les trois couches linéaires complètement connectées sont de tailles 728, 512 et 256 avec une sortie de taille 10 (i.e. nombre de classes). Nous avons intégré 2 couches de dropout avec une probabilité de 0.1 pour limiter le sur-apprentissage. Nous avons utilisé les versions de base des modèles LayoutLMv3 (133M de paramètres) et BROS (110M de paramètres). Pour le modèle utilisant un encodeur BROS [1], nous avons utilisé une couche linéaire de taille 768 vers une sortie de taille 10.

La longueur maximale des séquences est de 512 tokens (i.e. mots ou sous-mots). Si le document est plus court, on complète avec le token spécial `<PAD>`. Si le document est plus long, nous scindons le document en blocs de 512 tokens. Pour les blocs suivants, on utilise une méthode de recouvrement (i.e. stride) de 64 tokens. Cette méthode consiste à concaténer en début de séquence les 64 derniers tokens du bloc précédent. Elle permet de réduire la taille du modèle tout en permettant de traiter n'importe quel document. La quasi-totalité des documents font moins de 512 tokens.

5.2 Résultats de la classification de mots

La méthode qui utilise un encodeur LayoutLMv3 obtient des meilleurs résultats que BROS sur l'ensemble des classes (Table 1). On constate que les modèles ont tendance à ajouter plus de tokens dans les classes que d'en oublier (Rappel > Précision) sauf pour la classe `Other`. Les erreurs de prédiction se font essentiellement entre la classe `Other` et l'ensemble des autres classes. Les performances de classification de mots sur les coupes de sondage sont comparables aux performances sur des datasets de références dans le domaine de l'extraction d'informations de documents (Table 1), ce qui démontre l'efficacité de notre méthode malgré la complexité spécifique de nos documents.

5.3 Extraction des couples lithologie-profondeur

Nous avons vérifié manuellement 2500 couples de lithologie-profondeur, provenant d'environ 500 coupes de sondage, sur lesquelles nous avons identifié 465 erreurs. Ces erreurs incluent des lithologies manquantes ($\approx 10\%$) et majoritairement des profondeurs erronées ou mal attribuées ($\approx 90\%$).

Afin de limiter l'introduction de valeurs erronées dans la base de données finale, nous avons mis en place un mécanisme permettant d'identifier les erreurs d'attribution des profondeurs en détectant les cas où plusieurs lithologies partagent une même profondeur au sein d'un même sondage. Les

1. BROS for token classification (HuggingFace)

Dataset	Classe	BROS$_{BASE}$			LayoutLMv3$_{BASE}$		
		Précision	Rappel	F1-score	Précision	Rappel	F1-score
Sondage (notre)	Other	0.98	0.94	0.96	0.99	0.96	**0.97**
	Title	0.95	0.95	0.95	0.96	0.96	**0.98**
	Key_Word	0.96	0.99	0.98	1.00	1.00	**1.00**
	Key_Word_Answer	0.94	0.99	0.96	0.98	0.99	**0.99**
	End_Depth	0.92	0.96	0.94	0.99	0.99	**0.99**
	B_Litho	0.83	0.82	0.82	0.88	0.87	**0.87**
	I_Litho	0.91	0.95	0.93	0.91	0.97	**0.94**
	Depth_Scale	0.82	0.88	0.85	0.86	0.91	**0.88**
	Z_Start	0.90	0.99	0.94	0.94	0.99	**0.96**
	Litho_Depth_Value	0.77	0.80	0.79	0.73	0.90	**0.80**
	Moyenne pondérée	0.948	0.946	0.946	0.965	0.963	**0.963**
FUNSD		-	-	0.831	-	-	**0.903**
CORD		-	-	0.957	-	-	**0.966**

TABLE 1 – Comparaison des performances entre BROS et LayoutLMv3.

documents concernés sont temporairement exclus du corpus, en attendant une correction ultérieure. Le taux d'erreur est fortement lié à l'ancienneté du document. Par exemple, on détecte 8% d'erreurs sur les documents des années 2010 et 2020 tandis que ce taux est de 83% sur les documents des années 1990, notamment en raison de la moindre qualité graphique des coupes anciennes (faible contraste, typographies variées, dégradations matérielles liées au vieillissement).

Malgré ces difficultés, nous atteignons une précision finale élevée (98 %) sur les documents correctement traités, ce qui représente un résultat particulièrement satisfaisant compte tenu de la complexité des coupes de sondage que nous avons traitées.

6 Conclusion et perspectives

Dans cet article, nous avons évalué les performances de différents modèles pour l'extraction des couples lithologie-profondeur de coupes de sondage. Face aux limites intrinsèques du modèle DONUT, nous avons proposé une chaîne de traitement hybride en trois étapes : une reconnaissance de mots par un OCR externe ; une classification de mots utilisant un Transformer multimodal ; et un algorithme heuristique d'association des lithologies et profondeurs. Les modèles BROS et LayoutLMv3 ont été comparés *fine-tunés* sur des données annotées automatiquement. LayoutLMv3 obtient les meilleurs résultats avec un F1-Score de 96.3%. Notre algorithme d'association des lithologies et profondeurs a permis de traiter environ 80% des documents sur lesquels nous obtenons une précision de 98%.

Les documents anciens et de mauvaise qualité constituent la majorité des coupes de sondage où notre méthode échoue. Les erreurs d'OCR sont courantes et sont propagées tout au long de notre chaîne de traitement. Notre algorithme d'association de profondeurs aux lithologies s'appuie sur des règles peu robustes aux changements, particulièrement lorsqu'il s'agit de traiter des images de documents anciens et bruitées. Les approches de bout-en-bout apportent des perspectives intéressantes afin de limiter la propagation des erreurs et pourraient permettre de traiter les 20% de documents restants. Nous avons montré les limites de DONUT mais des variantes récentes pourraient limiter les phénomènes d'hallucinations et offrir plus de contrôle sur l'information. On peut citer CREPE (Okamoto *et al.*, 2024) qui génère la position des informations extraites ou LayoutLLM (Luo *et al.*, 2024) qui propose une chaîne de pensée guidant un raisonnement en plusieurs étapes.

Références

GROMPONE VON GIOI R., JAKUBOWICZ J., MOREL J.-M. & RANDALL G. (2012). LSD : a Line Segment Detector. *Image Processing On Line*, **2**, 35–55. DOI : 10.5201/ipol.2012.gjmr-lsd.

HONG T., KIM D., JI M., HWANG W., NAM D. & PARK S. (2022). BROS : A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10), 10767–10775. Number : 10, DOI : 10.1609/aaai.v36i10.21322.

HUANG Y., LV T., CUI L., LU Y. & WEI F. (2022). LayoutLMv3 : Pre-training for Document AI with Unified Text and Image Masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, p. 4083–4091, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3503161.3548112.

HUANG Z., CHEN K., HE J., BAI X., KARATZAS D., LU S. & JAWAHAR C. V. (2019). ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, p. 1516–1520. ISSN : 2379-2140, DOI : 10.1109/ICDAR.2019.00244.

HWANG W., KIM S., SEO M., YIM J., PARK S., PARK S., LEE J., LEE B. & LEE H. (2019). Post-OCR parsing : building simple and robust parser via BIO tagging.

HWANG W., YIM J., PARK S., YANG S. & SEO M. (2021). Spatial Dependency Parsing for Semi-Structured Document Information Extraction. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 330–343, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.28.

JAUME G., KEMAL EKENEL H. & THIRAN J.-P. (2019). FUNSD : A Dataset for Form Understanding in Noisy Scanned Documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, p. 1–6. DOI : 10.1109/ICDARW.2019.10029.

KIM G., HONG T., YIM M., NAM J., PARK J., YIM J., HWANG W., YUN S., HAN D. & PARK S. (2022). OCR-Free Document Understanding Transformer. In S. AVIDAN, G. BROSTOW, M. CISSÉ, G. M. FARINELLA & T. HASSNER, Éds., *Computer Vision – ECCV 2022*, p. 498–517, Cham : Springer Nature Switzerland. DOI : 10.1007/978-3-031-19815-1_29.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv :1910.13461 [cs], DOI : 10.48550/arXiv.1910.13461.

LUO C., SHEN Y., ZHU Z., ZHENG Q., YU Z. & YAO C. (2024). LayoutLLM : Layout Instruction Tuning with Large Language Models for Document Understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 15630–15640, Seattle, WA, USA : IEEE. DOI : 10.1109/CVPR52733.2024.01480.

MATHEW M., KARATZAS D. & JAWAHAR C. V. (2021). DocVQA : A Dataset for VQA on Document Images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 2199–2208, Waikoloa, HI, USA : IEEE. DOI : 10.1109/WACV48630.2021.00225.

MINDEE (2021). doctr : Document text recognition. https://github.com/mindee/doctr.

OKAMOTO Y., BAEK Y., KIM G., NAKAO R., KIM D., YIM M. B., PARK S. & LEE B. (2024). CREPE : Coordinate-Aware End-to-End Document Parser. In E. H. BARNEY SMITH, M. LIWICKI & L. PENG, Éds., *Document Analysis and Recognition - ICDAR 2024*, p. 3–20, Cham : Springer Nature Switzerland. DOI : 10.1007/978-3-031-70546-5_1.

PARK S., SHIN S., LEE B., LEE J., SURH J., SEO M. & LEE H. (2019). CORD : A Consolidated Receipt Dataset for Post-OCR Parsing.

XU Y., LI M., CUI L., HUANG S., WEI F. & ZHOU M. (2020). LayoutLM : Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 1192–1200. arXiv :1912.13318 [cs], DOI : 10.1145/3394486.3403172.

XU Y., XU Y., LV T., CUI L., WEI F., WANG G., LU Y., FLORENCIO D., ZHANG C., CHE W., ZHANG M. & ZHOU L. (2021). LayoutLMv2 : Multi-modal Pre-training for Visually-rich Document Understanding. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 2579–2591, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.201.