Quand les humanités numériques rencontrent l'IA : une enquête sur la propriété à Paris

Aaron Parmentelat
École nationale des Chartes - EHESS

aaron.parmentelat@chartes.psl.eu

Résumé

Dans le cadre d'une étude en humanités numériques de la propriété à Paris pendant la première moitié du XX^{e} siècle, nous cherchons à extraire les informations contenues dans un annuaire de propriétaires pour en faire une base de données interrogeable. Pour ce faire, nous utilisons trois modèles d'intelligence artificielle et nous évaluons leur qualité en mobilisant des métriques combinant mesures quantitatives et analyses qualitatives. L'étude met en lumière l'importance d'adapter ces modèles et métriques aux enjeux spécifiques de la source, tout en conservant une démarche reproductible. Elle illustre également comment les humanités numériques peuvent tirer parti des modèles d'IA tout en restant attentives à leurs limites et à la nécessité d'un contrôle qualité.

Mots-clés

Humanités numériques, Segmentation d'image, Reconnaissance optique de caractères, Reconnaissance d'entités nommées, Annuaire de propriétaires, Intelligence Artificielle, Métriques.

Abstract

As part of a digital humanities study on property in Paris during the first half of the 20th century, we aim to extract information from a property directory and turn it into a searchable database. To do so, we use three artificial intelligence models and assess their performance using metrics that combine quantitative measures and qualitative analysis. The study highlights the importance of adapting these models and metrics to the specific challenges of the source material, while maintaining a reproducible approach. It also illustrates how digital humanities can benefit from AI models, while remaining mindful of their limitations and the need for quality control.

Keywords

Digital Humanities, Image segmentation, Optical Character Recognition, Named Entity Recognition, Property Directory, Artificial Intelligence, Metrics.

1 Introduction

L'émergence des techniques d'intelligence artificielle suscite un vif intérêt, comme en témoigne le Sommet pour l'action sur l'Intelligence Artificielle qui s'est tenu en février 2025. Cet ensemble d'outils ouvre des perspectives jusque-là inaccessibles, y compris dans le domaine des humanités numériques. Ils peuvent en particulier faciliter le traitement et l'étude de corpus massifs, comme celui dont nous allons parler dans cet article et qui concerne les régimes de propriété dans la ville de Paris.

Cette technologie soulève aussi des enjeux de vérification et de reproductibilité. Il est en effet indispensable de s'assurer que la tâche que nous essayons d'automatiser est bien définie (s'il est aisé de déterminer si la prédiction du modèle est vraie ou fausse) éet que le modèle utilisé soit efficace. Ceci est d'autant plus important quand les corpus étudiés sont volumineux et qu'il est trop chronophage de les relire manuellement.

Mon mémoire de recherche en humanités numériques porte sur la propriété parisienne à Paris en 1898 et 1951. Il a été encadré par Carmen Brando (EHESS), Frédérique Mélanie-Becquet (CNRS/LATTICE) et Gilles Postel-Vinay (PSE) dans le cadre du consortium Paris Time Machine [1]. J'étudie plus précisément les régimes de propriété, notamment la multi-propriété (une personne détient plusieurs biens) et la copropriété (plusieurs personnes possèdent un bien). Mes sources sont des annuaires de propriétaires annuels qui contiennent pour toutes les adresses de la capitale le nom de la personne qui possède l'immeuble (personne morale ou physique) et l'adresse de son domicile. Ces *Annuaires des propriétaires et des propriétés de Paris et du département de la Seine*, publiés de 1894 à 1951 pour la ville de Paris sont détenus par la BnF [2]. Mon but est d'en extraire les informations pour les structurer en une base de données sous la forme d'un fichier CSV. Pour l'année 1898, ce travail a déjà été effectué par l'équipe citée plus haut ainsi que Gabriela Elgarrista (Plateforme Géomatique EHESS, ENC). Je souhaite comparer l'année 1898 avec 1951, qui est le dernier volume de ces annuaires car une émergence de la copropriété rend la masse de propriétaires trop importante pour être publiée annuellement.

Les annuaires se structurent en deux parties. La première

1. Mohamed KHEMAKHEM et al. "Fueling Time Machine : Information Extraction from Retro-Digitised Address Directories". In : *JADH2018 "Leveraging Open Data"*. Tokyo, Japan, sept. 2018. URL : https://hal.science/hal-01814189 (visité le 30/05/2025).

2. Lien de la notice : https://catalogue.bnf.fr/ark:/12148/cb32697229h

liste les immeubles de Paris par ordre alphabétique des rues et indique pour chaque adresse le nom et le domicile de son propriétaire. La deuxième liste les propriétaires par ordre alphabétique et indique pour chacun son domicile et les adresses du ou des immeubles qu'il possède. J'ai choisi de me concentrer sur cette deuxième partie qui me permet de traiter la multi-propriété de manière relativement peu chronophage, mais le volume de 1898 a été traité avec la première liste puis un travail manuel pour désambiguïser les multi-propriétaires.

J'ai essayé d'appliquer la chaîne de traitement élaborée pour le volume de 1898 à celui de 1951, mais j'ai constaté de très mauvaises performances. J'ai donc décidé de mettre au point une nouvelle chaîne de traitement adaptée à cet annuaire. Sa version finale est en langage python et utilise trois modèles d'intelligence artificielle qui traitent de tâches différentes et objectivement définies, et qui ont diverses méthodes d'entraînement (modèle prêt-a-l'emploi, modèle existant fine-tuné sur les données et modèle entièrement entraîné à partir d'annotations). Elle est composée de trois étapes : segmentation des zones de texte, extraction du texte (OCR) et structuration du texte dans les catégories correspondantes.

Dans cet article nous présentons les différentes étapes de ce processus et les modèles envisagés, puis nous étudions les performances de ces modèles dans une démarche hybride qualitative et quantitative.

2 Les étapes de la chaîne de traitement et les modèles correspondants

Dans le champ de l'histoire économique de la propriété foncière, les annuaires de propriétaires sont des sources particulièrement riches. Elles sont également précieuses car le résultat d'un travail de grande ampleur qui n'existe que pour des localités et périodes spécifiques, elles n'existent par exemple pas à Lyon sur une période similaire [3]. Ces annuaires permettent notamment d'étudier la répartition spatiale des immeubles et des domiciles des propriétaires et de comparer ces deux dimensions, en prenant en compte les régimes de propriété (ici, multi-propriété et copropriété).

Nous étudions la liste par nom des propriétaires, dont une page est reproduite ci-contre. Notre objectif est d'obtenir un tableau contenant : les noms des propriétaires, les adresses des domiciles, et les informations des immeubles possédés (quartiers, adresses, nombre de propriétés). Pour cela nous devons convertir l'annuaire en une base de données interrogeable.

Quelle architecture logicielle pouvons nous créer pour répondre à notre objectif ? La principale problématique à laquelle je me suis heurté est d'appliquer une lecture prenant en compte l'organisation du document en trois colonnes durant la reconnaissance automatique du texte. Le travail

FIGURE 1 – Page d-2 de l'annuaire de propriétaires parisien de 1951 (ark :/12148/cb32697229h)

mené sur le volume de 1898 a résulté en un modèle *Transkribus* reconnaissant à la fois les zones de texte et les caractères. J'ai essayé de le réutiliser pour le volume de 1951, mais cet essai s'est avéré vain à cause de différences notables de mise en page : en 1898 les pages sont constituées de deux colonnes séparées par un trait, contre trois colonnes non délimitées en 1951. J'ai dû adapter mon approche pour concevoir une nouvelle chaîne de traitement, reposant sur l'introduction de nouvelles méthodes d'intelligence artificielle efficaces sur les données, notamment en distinguant les étapes de segmentation des zones de texte et d'océrisation.

Cela définit trois grandes étapes dans notre chaîne de traitement :

- le découpage des colonnes pour conserver un ordre de lecture cohérent ;
- la transcription du texte ;
- la structuration du texte dans les catégories pertinentes, qui correspondent aux colonnes du fichier CSV de sortie.

Pour la segmentation des colonnes (zones de texte), beaucoup de modèles existent mais nous avons identifié le modèle *YOLOv8 Segmentation* d'Ultralytics [4]. Il s'agit d'un modèle de détection et de segmentation d'objets, couramment utilisé pour l'analyse d'images. Par exemple cet article de Chahan Vida-Gorène et Aliénor Decours-Perez qui

3. Loïc BONNEVAL, François ROBERT et Jean-Luc Préfacier PINOL. *L'immeuble de rapport : l'immobilier entre gestion et spéculation (Lyon 1860-1990)*. Rennes, France : Presses universitaires de Rennes, 2013. 242 p. ISBN : 978-2-7535-2170-4.

4. Plus d'informations sur leur site web : https://docs.ultralytics.com/tasks/segment/

utilise ce modèle pour détecter la localisation et le contenu d'inscriptions et de graffitis arméniens endommagés [5].

Afin d'optimiser ses performances sur notre corpus spécifique, nous avons choisi de procéder à un *fine-tuning*, c'est-à-dire un ré-entraînement du modèle sur un sous-ensemble des données annoté manuellement. Pour cela, 75 pages de l'annuaire ont été sélectionnées et annotées : le modèle devient alors un modèle d'apprentissage supervisé, entraîné sur des exemples issus de notre propre jeu de données. L'un des avantages de la bibliothèque Ultralytics est sa grande accessibilité : elle fournit des scripts prêts à l'emploi accompagnés de tutoriels vidéo ce qui facilite son utilisation – en particulier pour des personnes moins familières avec l'entraînement de modèles d'intelligence artificielle.

Pour la reconnaissance optique de caractères (OCR) nous avons comparé deux modèles. Tout d'abord, **Tesseract** [6], un outil *open source* largement utilisé dans la communauté des humanités numériques pour sa simplicité de prise en main et sa large documentation. Ensuite, **PERO-OCR** [7] un moteur plus récent et moins simple d'utilisation, développé par l'Université technique de Prague et qui a notamment été utilisé dans le cadre du projet ANR *SoDUCo* [8] consacré à l'analyse de sources d'annuaires commerciaux anciens similaires à notre corpus. Ces modèles sont tous les deux utilisables tels quels et nous n'allons pas procéder à un entraînement supplémentaire.

Après océrisation nous obtenons le texte brut de la source, il nous reste donc à identifier précisément les différentes informations qu'elle contient (noms, adresses, etc.). Cette opération relève de la **reconnaissance d'entités nommées** (*Named Entity Recognition* ou NER), une technologie d'apprentissage supervisé permettant de classer des portions de texte dans des catégories pré-définies. Contrairement à la segmentation ou à l'OCR nous ne réutilisons pas un modèle existant mais nous entraînons un modèle de zéro à partir d'un sous-ensemble annoté du corpus.

Pour l'annotation j'utilise l'outil **Prodigy**, un environnement interactif et intuitif. Sur cinq pages – soit 1160 lignes – j'annote chaque entrée avec les étiquettes correspondantes. Ce jeu de données sert ensuite de base à l'entraînement de modèles de NER avec la bibliothèque **spaCy**. Notons que les étiquettes sont faciles à définir car toutes les entrées ont la même structure, par exemple le cas limite de la « rue de Rouen, à Dijon » est découpé comme « rue » (type_voie) « de Rouen » (nom_voie), « Dijon » (ville), « Côte d'Or » (loc) car ici Rouen correspond à un nom de voie. Notons également qu'il existe d'autres outils pour l'entraînement de modèles de NER, mais nous nous sommes limités à entraîner plusieurs modèles avec spaCy par contrainte de temps.

Les étiquettes définies pour notre modèle sont :

- « PER » pour le nom du propriétaire ;
- « PRENOM » pour son prénom ;
- « STATUT » pour sa civilité ;
- « ORG » pour le nom de l'organisation (si le propriétaire est une personne morale) ;
- « NUM », « TYPE_VOIE » et « NOM_VOIE » pour l'adresse du domicile du propriétaire ;
- « ARR » pour l'arrondissement du domicile (pour les adresses parisiennes) ;
- « VILLE » pour la ville du domicile (pour les adresses françaises et non parisiennes) ;
- « LOC » est une information supplémentaire sur le domicile qui renvoie à son département (en France) ou son pays (à l'étranger) ;
- « PART » correspond à son régime de propriété (copropriété, indivision, etc.) ;
- « CODE » renvoie aux codes des immeubles qu'il possède ;
- « ENTETES » renvoie aux entêtes (comme les numéros de pages).

Nous avons donc opté pour un entraînement personnalisé de certains modèles afin d'augmenter leurs performances sur cette source. Il s'agit maintenant d'étudier la pertinence de ces modèles sur nos données. Ce contrôle de qualité est essentiel pour s'assurer que notre chaîne de traitement est efficace et permette ensuite une analyse des régimes de propriété à Paris dans la première moitié du XX^{e} siècle.

3 Comment évaluer l'efficacité des modèles d'intelligence artificielle de la chaîne de traitement ?

Nous avons fait évoluer notre problématique : comment évaluer l'efficacité et la résilience des modèles dans leurs missions ? Nous souhaitons avoir des modèles efficaces et un processus reproductible dans le cadre d'une démarche de recherche.

Quand nous fournissons des annotations aux modèles ils produisent des métriques de performance comme la précision, le rappel ou le F-score que nous détaillerons plus tard. Pour les autres outils, une annotation manuelle du corpus – une « vérité de terrain »– est demandée pour y confronter les prédictions du modèle. Pour chacune des trois tâches de notre pipeline – segmentation, océrisation et reconnaissance d'entités nommées – nous avons mis en place une stratégie de contrôle qualité adaptée, associant des démarches qualitatives et quantitatives.

5. Chahan VIDAL-GORÈNE et Aliénor DECOURS-PEREZ. "Detecting and Deciphering Damaged Medieval Armenian Inscriptions Using YOLO and Vision Transformers". In : t. 14936. Springer Nature Switzerland, 11 sept. 2024, p. 22. DOI : 10.1007/978-3-031-70642-4_2. URL : https://enc.hal.science/hal-04722397 (visité le 30/05/2025).

6. Plus d'information sur leur dépôt GitHub : https://github.com/tesseract-ocr/tesseract

7. Plus d'information sur leur dépôt GitHub : https://github.com/DCGM/pero-ocr

8. N. ABADIE et al. "A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories". In : *Document Analysis Systems. DAS 2022*. Sous la dir. de S. UCHIDA, E. BARNEY et V. EGLIN. Document Analysis Systems. DAS 2022. Issue : 13237. La Rochelle, France : Springer, Cham, mai 2022. DOI : 10.1007/978-3-031-06555-2_30. URL : https://hal.science/hal-03698609 (visité le 06/06/2024).

Le modèle de segmentation YOLOv8 a été fine-tuné sur l'annuaire à partir d'un corpus d'entraînement de 75 pages réparties comme suit :

- **53 pages** pour l'entraînement (le modèle apprend à partir de ces données) ;
- **15 pages** pour la validation (le modèle ajuste ses paramètres en fonction de nouvelles données) ;
- **7 pages** pour le jeu de test (pour évaluer les performances du modèle sur des données qu'il n'a jamais vues et donc vérifier son caractère généralisable).

Trois classes sont distinguées : *col_1*, *col_2* et *col_3* correspondant aux trois colonnes d'une page. Le modèle nous fournit une matrice de confusion normalisée :

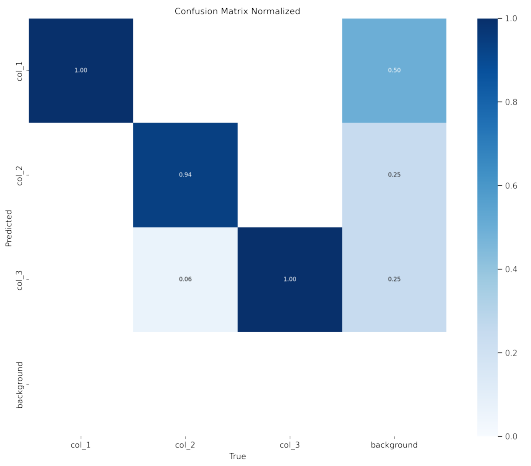

FIGURE 2 – Matrice de confusion normalisée du modèle de segmentation *(YOLOv8 fine-tuné)*

La matrice montre que les colonnes identifiées comme *col_2* sont correctes dans 94% des cas ; les 6% d'erreurs sont classées à tort comme des *col_3*. Pour les deux autres classes, la reconnaissance est parfaite (100%).
Nous obtenons également les métriques de précision et de rappel :

- **Précision** : parmi toutes les entités détectées par le modèle, combien sont correctes ?
- **Rappel** : parmi toutes les entités correctes, combien le modèle en a-t-il détecté ?

L'article cité plus haut obtient à partir du même modèle en moyenne 0,91 de précision et 0,88 de rappel [9]. La précision moyenne de notre modèle de 0,77 et le rappel moyen de 0,96, ce qui est légèrement inférieur mais montre que le modèle est globalement efficace. Il reconnaît très souvent les colonnes mais a tendance à également détecter des morceaux d'images en dehors de ses fonctions.
Nous décidons de procéder à un contrôle qualitatif des résultats pour toutes les pages de l'annuaire, en vérifiant à la main si trois colonnes de dimensions cohérentes sont détectées, et corriger les quelques erreurs. Ce contrôle nous permet également de gérer des cas limites comme les pages comportant des noms de propriétaires commençant par deux lettres différentes, qui sont scindées en deux comme dans un dictionnaire.
Ce contrôle – bien qu'impraticable à grande échelle – s'avère très efficace ici car il maximise la fiabilité du corpus pour un investissement de temps raisonnable.

Concernant l'OCR, nous avons testé deux modèles candidats, Tesseract et PERO-OCR, que nous appliquons tels quels. Aucune métrique d'évaluation n'est intégrée à ces outils donc nous devons trouver une solution de contrôle qualité.

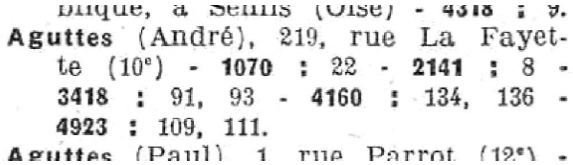
Aguttes (André), 219, rue La Fayette (10e) - 1070 : 22 - 2141 : 8 - 3418 : 91, 93 - 4160 : 134, 136 - 4923 : 109, 111.

FIGURE 3 – Entrée de Monsieur André Aguttes, page d-3 (ark :/12148/cb32697229h)

Nous observons dans cette entrée que André Aguttes résidant au 219, rue La Fayette dans le dixième arrondissement possède plusieurs immeubles : le numéro 22 de la voie de code 1070, le numéro 8 de la voie de code 2141, et ainsi de suite. Les codes correspondent au rang de la voie dans l'ordre alphabétique : de la « rue de l'Abbaye » de code 1 à la « rue de la Zone » de code 4963. Nous constatons que les chiffres – codes des voies, numéros d'immeubles, arrondissements – sont cruciaux dans cette source : une erreur d'un chiffre entre 1647 et 1648 change de la rue des Éperons (sixième arrondissement) en l'impasse des Épinettes (dix-septième arrondissement).
Nous avons tiré la page 75 au hasard et classé les erreurs détectées en trois catégories selon un indice de gravité. Les erreurs graves concernent au moins deux caractères d'affilée ou une erreur dans la reconnaissance des lignes. Les erreurs moyennement graves sont des erreurs d'un seul caractère sur des informations importantes pour la suite du traitement : un chiffre, un nom de voie ou un signe de ponctuation (qui pourrait nuire à l'efficacité du NER). Les erreurs peu graves concernent des erreurs d'accents ou d'un caractère dans le nom de la personne.

Gravité des erreurs	Tesseract	PERO
Erreurs peu importantes	130	87
Erreurs moyennement graves	91	110
Erreurs graves	11	3
Total erreurs	232	200

TABLE 1 – Les erreurs classées par gravité des deux moteurs d'OCR sur la page d-75

Malgré un nombre total d'erreurs légèrement plus élevé dans les catégories faibles et moyennes, PERO se distingue par un taux d'erreurs graves bien plus faible. Il se montre également robuste dans la reconnaissance des chiffres en

9. VIDAL-GORÈNE et DECOURS-PEREZ, "Detecting and Deciphering Damaged Medieval Armenian Inscriptions Using YOLO and Vision Transformers".

faisant une seule erreur de chiffre contre 67 pour Tesseract. Nous choisissons donc PERO comme moteur d'OCR sur l'ensemble du corpus.
Une autre manière d'évaluer un moteur OCR serait le **Character Error Rate (CER)** : il mesure le nombre de modifications nécessaires pour transformer la sortie du modèle en la vérité de terrain (distance de Levenshtein), retranché au nombre total de caractères. Lors d'un autre projet sur une source similaire (imprimé datant de la première moitié XX^{e} siècle) nous avons obtenu un CER de 1,1% pour PERO contre 5,9% pour Tesseract – un écart significatif et cohérent avec nos conclusions.
Dans ce cas précis, nous avons toutefois préféré une classification qualitative des erreurs, qui permet une lecture plus fine et mieux adaptée à nos objectifs d'extraction des données.

Le modèle de reconnaissance d'entités nommées a été entraîné avec *Spacy*, exclusivement à partir d'annotations manuelles. Nous disposons de métriques pour chaque étiquette : précision, rappel et F-score. Nous avons déjà vu ces métriques pour le fine-tuning du modèle de segmentation, à l'exception du **F-score**. Ce dernier correspond à la moyenne entre la précision et le rappel et permet un aperçu synthétique de la performance du modèle. Nous entraînons un premier modèle sur l'ensemble des étiquettes.

Étiquette	Précision	Rappel	F-score
NUM	97.75	97.75	97.75
ARR	96.43	97.59	97.01
NOM_VOIE	98.35	92.97	95.58
PER	91.49	97.73	94.51
TYPE_VOIE	93.88	94.85	94.36
VILLE	91.30	84.00	87.50
LOC	91.30	84.00	87.50
CODE	86.93	86.36	86.64
STATUT	65.00	68.42	66.67
PRENOM	60.00	61.90	60.94
ORG	23.08	27.27	25.00
PART	14.29	16.67	15.38
ENTETE	0.00	0.00	0.00

TABLE 2 – Scores du premier modèle de NER ordonnés par F-score

L'étiquette *ENTETE* est totalement absente des prédictions à cause de sa faible représentation dans les données. Elle est retirée du jeu d'entraînement, et les en-têtes sont supprimées du texte océrisé à l'aide d'expressions régulières. Nous entraînons un deuxième modèle ne prenant pas en compte les entêtes.
Les étiquettes 'Statut', 'Prénom' et 'Org' ont des scores relativement plus faibles que les autres en raison de leur faible représentation dans les données. Les F-score de ce deuxième modèle sont tous compris entre 64 et 100, alors que pour le premier modèle cet intervalle est de 0 à 97,75. De plus l'étiquette 'Code' est stratégique car c'est celle que nous allons utiliser pour délimiter les entrées dans la suite

Étiquette	Précision	Rappel	F-score
PART	100.00	100.00	100.00
NUM	96.67	96.67	96.67
ARR	95.45	97.67	96.55
TYPE_VOIE	93.94	93.94	93.94
NOM_VOIE	91.79	92.48	92.13
CODE	91.56	90.97	91.26
PER	86.52	95.06	90.59
VILLE	95.83	79.31	86.79
LOC	95.83	76.67	85.19
STATUT	75.00	78.95	76.92
PRENOM	69.49	70.69	70.09
ORG	72.73	57.14	64.00

TABLE 3 – Scores du second modèle de NER ordonnés par F-score

de la chaîne de traitement et elle gagne environ 5 points dans toutes les métriques entre le premier et le second modèle. L'étiquette 'Part' est également importante car elle contient le régime de propriété qui est sur ce quoi porte notre analyse, et elle passe d'un score très bas dans le premier modèle à un score parfait dans le second. Nous choisissons donc d'utiliser le deuxième modèle sur l'ensemble des pages de l'annuaire.
Une approche quantitative des métriques, alliée avec une connaissance qualitative des données les plus importantes (pour la suite du traitement et pour l'analyse) permet de choisir le modèle le plus adapté à notre démarche.

Tout au long de la chaîne de traitement notre démarche repose sur des indicateurs quantitatifs, en restant ancrée dans une lecture qualitative des résultats ou une correction manuelle des erreurs. Cette approche hybride permet de conserver un regard critique sur les performances des modèles et d'ajuster le processus d'extraction aux spécificités de la source et aux objectifs de l'analyse. De plus, dans une logique d'*Open science*, mes modèles et jeux d'entraînement sont publiés sur GitHub [10].

4 Conclusion

Au fil de cette exploration, nous avons réfléchi aux conditions de mise en place d'une chaîne de traitement de données s'appuyant sur différents modèles d'IA pour répondre à nos tâches de segmentation d'images, reconnaissance de texte et extraction d'entités nommées.
L'un des principaux enseignements de cette démarche est que l'usage de l'IA, pour être pertinent dans le cadre des humanités numériques, doit toujours être guidé par une connaissance des données et une compréhension fine des objectifs de recherche. Le contrôle qualité est une étape essentielle de toute utilisation de méthodes computationnelles sur des corpus de SHS, y compris des modèles d'IA.
Le recours à des méthodes quantitatives — scores de précision, matrices de confusion, taux d'erreur — est essentiel

10. Voir mon dépôt GitHub : https://github.com/parm-a/memoire_M2

pour évaluer les performances des modèles, mais est toujours complété par un travail qualitatif :des contrôle manuel, définition de la gravité des erreurs, attention portée aux étiquettes les plus importantes. Ce double regard est indispensable et montre l'intérêt de l'interdisciplinarité des humanités numériques.
Ce type d'approche comporte aussi des limites. La constitution de jeux d'entraînement annotés manuellement est chronophage, et la vérification humaine est difficilement généralisable à des corpus massifs. Certains biais sont présents dès la source, et d'autres peuvent s'introduire lors de la définition de catégories ou dans la représentativité des données d'entraînement. Il faut donc garder à l'esprit que les résultats fournis par l'IA ne sont jamais neutres : ils reflètent toujours des choix, des simplifications, des biais. Pour autant, cela ne signifie pas qu'ils sont inutilisables – simplement qu'ils doivent être accompagnés, interrogés, contextualisés.

En ce sens, je pense que l'intelligence artificielle peut jouer un rôle précieux dans les humanités numériques à condition d'être mobilisée comme un outil critique et non comme une boîte noire. Loin d'automatiser la pensée, elle peut au contraire enrichir nos hypothèses et ouvrir de nouvelles pistes, à condition que nous gardions le contrôle sur les interprétations et les conclusions. Cette exigence de réflexivité – sur les sources, les modèles, mais aussi nos propres biais de chercheuses et chercheurs – est au cœur des humanités numériques, dans une démarche réellement hybride.

Références

ABADIE, N. et al. "A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories". In : *Document Analysis Systems. DAS 2022*. Sous la dir. de S. UCHIDA, E. BARNEY et V. EGLIN. Document Analysis Systems. DAS 2022. Issue : 13237. La Rochelle, France : Springer, Cham, mai 2022. DOI : 10.1007/978-3-031-06555-2_30. URL : https://hal.science/hal-03698609 (visité le 06/06/2024).

BONNEVAL, Loïc, François ROBERT et Jean-Luc Préfacier PINOL. *L'immeuble de rapport : l'immobilier entre gestion et spéculation (Lyon 1860-1990)*. Rennes, France : Presses universitaires de Rennes, 2013. 242 p. ISBN : 978-2-7535-2170-4.

KHEMAKHEM, Mohamed et al. "Fueling Time Machine : Information Extraction from Retro-Digitised Address Directories". In : *JADH2018 "Leveraging Open Data"*. Tokyo, Japan, sept. 2018. URL : https://hal.science/hal-01814189 (visité le 30/05/2025).

VIDAL-GORÈNE, Chahan et Aliénor DECOURS-PEREZ. "Detecting and Deciphering Damaged Medieval Armenian Inscriptions Using YOLO and Vision Transformers". In : t. 14936. Springer Nature Switzerland, 11 sept. 2024, p. 22. DOI : 10.1007/978-3-031-70642-4_2. URL : https://enc.hal.science/hal-04722397 (visité le 30/05/2025).