Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture

Marie-Laurence Bonhomme

Répertoire des Notaires parisiens
Segmentation automatique et reconnaissance d'écriture
Rapport exploratoire

Marie-Laurence Bonhomme
Équipe ALMAnaCH
Inria

24 octobre 2018

TABLE DES MATIÈRES

Ce rapport est le produit de la phase 1 du projet LECTAUREP, réalisé dans le cadre de la convention de rechercherche particulière entre le ministère de la Culture, les Archives nationales et Inria (ALMAnaCH).
Pour l'équipe ALMAnaCH, les participants sont :
- Marc Bui (DRC EPHE-Paris 8),
- Laurent Romary (DR Inria),
- Daniel Stöckl Ben Ezra (DR EPHE),
- Éric Villemonte de la Clergerie (CR Inria),
- Marie Puren (IR Inria),
- Charles Riondet (IR Inria).

Répertoire des Notaires parisiens Segmentation automatique et reconnaissance d'écriture

Rapport exploratoire

***Résumé*—Les répertoires des notaires de Paris conservés aux Archives nationales sont parmi les fonds les plus consultés par le public, mais s'ils sont numérisés et disponibles sur la Salle des Inventaires Virtuelle, pour les exploiter les lecteurs doivent toujours en passer par un dépouillement méthodique car ces répertoires ne sont pas transcrits et on ne peut donc pas y effectuer de recherche en plein texte. Afin de les rendre plus aisément utilisables comme inventaires des minutes des notaires, et d'en permettre des exploitations nouvelles, appliquer les techniques de reconnaissance automatique d'écriture à ce volumineux corpus semble particulièrement opportun. La structure régulière des documents, et une certaine prévisibilité de leurs contenus constituent des atouts, tandis que la multiplicité des écritures rencontrées dans les répertoires est une difficulté qui ne peut pas être ignorée. Une phase d'expérimentation a produit des résultats encourageants quant aux performances de la reconnaissance automatique d'écriture sur ces documents, et offert des pistes quant aux moyens de les améliorer au cours d'un projet plus long et plus ambitieux.**

I. Introduction

Le présent rapport concerne la phase d'exploration préparatoire à la mise en place du projet LECTAUREP (LECTure AUtomatique des REPertoires), portant sur la réalisation d'une reconnaissance automatique d'écriture sur les répertoires des notaires de Paris et la mise à disposition de ces documents et de leur transcription sur une plateforme en ligne, offrant des possibilités de recherche avancée. Ce projet durera un an à compter de la constitution définitive de l'équipe *ad hoc*, et est un partenariat entre les Archives nationales (et plus particulièrement le département du Minutier central et le département de la maîtrise d'ouvrage des systèmes d'information), le ministère de la Culture et Inria (Institut national de recherche en informatique et automatique.)

II. Le corpus : Répertoires d'actes tenus par les notaires de Paris entre 1803 et 1940

Les répertoires des notaires parisiens sont conservés aux Archives Nationales, site de Paris, par le département du Minutier central. Un répertoire de notaire est un registre qui liste, pour chaque jour, les actes passés dans une étude donnée. Déjà présent dans la pratique professionnelle notariale sous l'Ancien Régime, il est cependant formalisé en 1803 lorsque le Premier consul Bonaparte réorganise le notariat par la loi du 25 Ventôse an XI, articles 29 et 301. Le nombre de colonnes du répertoire devient fixe, de même que les informations qui sont obligatoires. Les colonnes des répertoires ainsi que l'en-tête des ces tableaux sont désormais pré-imprimées.

Les actes sont divisés en deux catégories : les actes en minutes, et les actes en brevets. Les actes en minute sont ceux dont le notaire conserve l'original, ce qui n'est pas le cas pour les actes en brevet dont l'original, dépourvu de force exécutoire, est remis aux parties.

Plus de 917 notaires différents ont exercé entre 1803 et 1940 à Paris, et même si nous n'avons pas tous leurs répertoires (pertes accidentelles, etc.), on peut évaluer entre 1800 et 2000 le nombre de registres de répertoires laissés par le notariat parisien pour cette tranche chronologique, chacun de ces répertoires comptant 300 à 500 pages.

Le répertoire d'une étude notariale contient pour chaque acte les informations suivantes : la date à laquelle l'acte a été passé, et enregistré, un numéro d'ordre dans le répertoire, la nature de l'acte (le type d'acte juridique : procuration, contrat, mandat ...), le montant des frais perçus par le notaire pour la passation de cet acte et le nom des parties, ainsi que d'autres précisions le cas échéant comme l'adresse de ces parties, les biens concernés par un acte ...

Ces répertoires peuvent d'abord servir comme instruments de recherche pour les archives des études notariales conservées par les Archives nationales ; mais leur numérisation et l'utilisation des procédés de reconnaissance optique de caractères permettront d'autres usages, comme l'exploitation statistique des informations qu'ils contiennent.

Hétérogénéité des graphies

D'après quelques carottages, il apparaît qu'on rencontre systématiquement plus d'une écriture dans un registre, au moins de façon marginale (par exemple pour les visas trimestriels, vérifications périodiques de la bonne tenue du registre par les receveurs de l'enregistrement [1].) De ce fait, une approche de la reconnaissance d'écriture basée sur une documentation exhaustive des différentes mains semble impossible.

Description des tableaux

Chaque page d'un registre comporte un tableau à 7 colonnes : le numéro d'ordre de l'acte dans le répertoire, la date à laquelle il a été passé, la nature de l'acte (une

colonne si l'acte est en brevet, une autre s'il est en minute), les noms des parties et autres détails sur l'acte, la date de son enregistrement et le montant des droits acquittés par les parties. L'en-tête est imprimé et ne change pas.
La majorité des registres ne font pas apparaître explicitement les lignes (même si certaines études, ou certains clercs, ont pris soin de tracer des séparation entre chaque acte, comme pour certains registres du notaire Jean Panhard de l'étude XIV (MC/RE/XIV/24 et 25), aussi, pour pouvoir exploiter ces répertoires de manière intéressante, il faut pouvoir identifier les actes, qui constituent les lignes du tableau, pour avoir ensemble toutes les informations concernant un acte donné (on pourrait ainsi par exemple mettre en relation le type d'acte avec les frais perçus par le notaire.)
Ainsi, avant de pouvoir passer à la reconnaissance de caractères elle-même, il importe d'identifier la structure des documents et les informations qu'on souhaite en extraire.

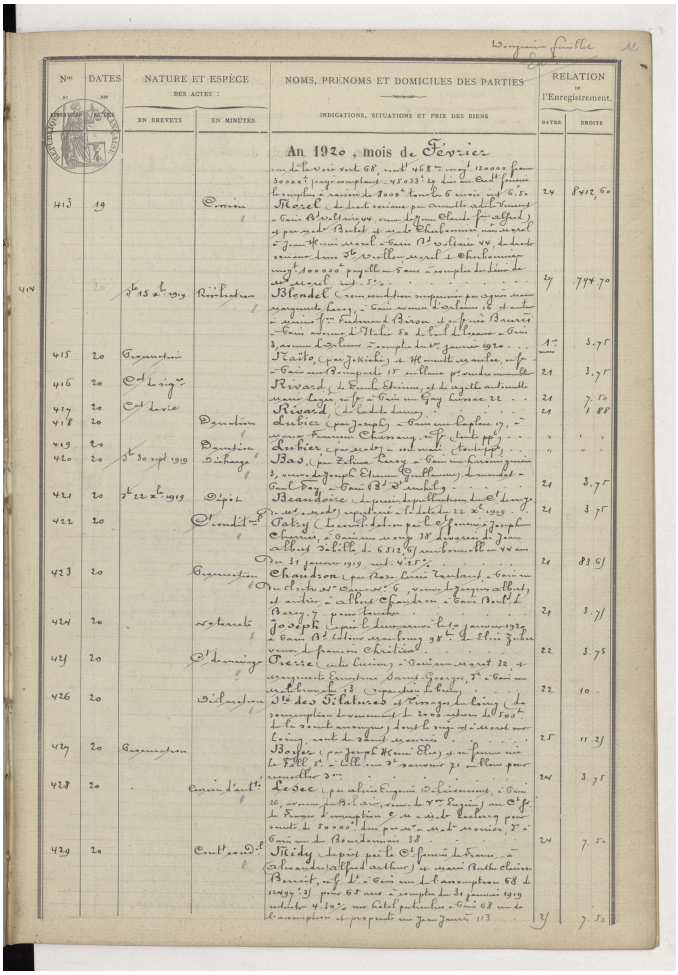

FIGURE 1. Registre de l'étude Marotte (XLIII), février 1920

III. LA SEGMENTATION

La segmentation des tableaux est le premier problème à résoudre. Pour y parvenir, il convient de comprendre tant la structure des documents physiques que l'organisation sémantique des différents éléments.

A. Structure des tableaux

Toutes les pages sont divisées en 7 colonnes, auxquelles nous avons ajouté 2 colonnes à droite et à gauche du tableau dessiné pour prendre en charge les cas où le texte ne se trouve pas dans les colonnes extrêmes du tableau mais dans les marges, du fait d'un obstacle dans la colonne ou pour une autre raison. Nous découpons donc toutes les pages en 9 colonnes, en suivant celles qui sont effectivement présentes sur les images.
Chaque page comporte également un en-tête imprimé indiquant le contenu des colonnes. Puisqu'il est identique sur chaque page nous avons choisi de l'ignorer dans la détermination des zones d'information qui nous intéressent ; il pourra être réintégré au moment de la sortie si on le souhaite.
Sous cet en-tête imprimé se trouve toujours une mention de date, en partie imprimée également : le mois et l'année en cours, sous la forme "An 19__, mois de _____".

Structure et contenu des colonnes.

Colonne 1.

— *Structure.* Cette colonne est normalement vide, mais elle contient parfois une ligne de texte alignée en haut de la cellule.

— *Contenu.* Cette colonne est normalement vide, mais elle peut contenir des chiffres : le numéro d'ordre de l'acte, habituellement indiqué en colonne 2, lorsqu'un obstacle se trouve dans cette colonne, ou pour d'autres raisons.

Colonne 2.

— *Structure.* Cette colonne contient normalement une ligne de texte, alignée en haut de la cellule, mais peut parfois être vide. En présence d'un obstacle (par exemple un tampon) cette ligne de texte peut être décalée vers le bas.

— *Contenu.* Cette colonne contient normalement des chiffres : le numéro d'ordre de l'acte enregistré. Cette numérotation est réinitialisée à chaque 1er janvier ; sur les documents numérisés dont on dispose, on a entre 1000 et 3000 actes par an, avec une tendance générale à la hausse d'année en année.
Mais cette colonne peut être vide lorsqu'en raison d'un obstacle (par exemple un tampon), ou pour une autre raison, la numérotation est inscrite dans la colonne 1. Il est également possible que les deux colonnes contiennent du texte.

Colonne 3.

— *Structure* Cette colonne contient normalement une ligne de texte, alignée en haut de la cellule. En présence d'un obstacle (par exemple un tampon) cette ligne de texte peut être décalée vers le bas.

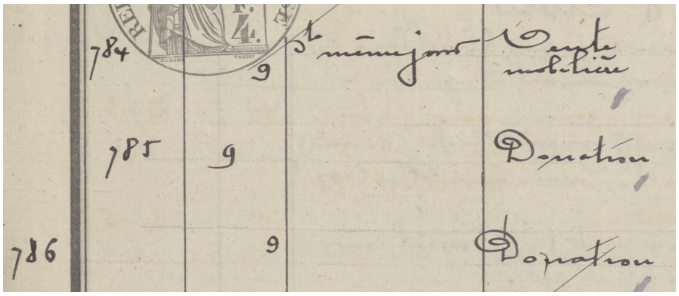

FIGURE 2. Exemple d'anomalie dans les colonnes 1 et 2.

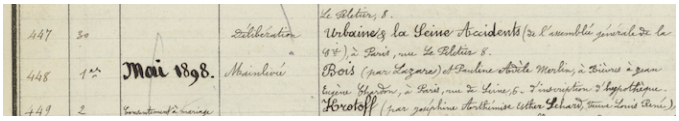

FIGURE 3. Exemple de mention de date dans la colonne 4.

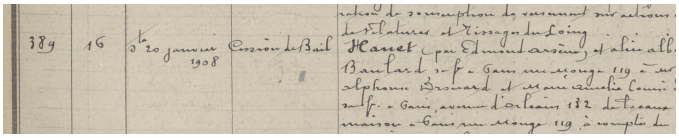

FIGURE 4. Exemple de mention "suite" dans la colonne 4.

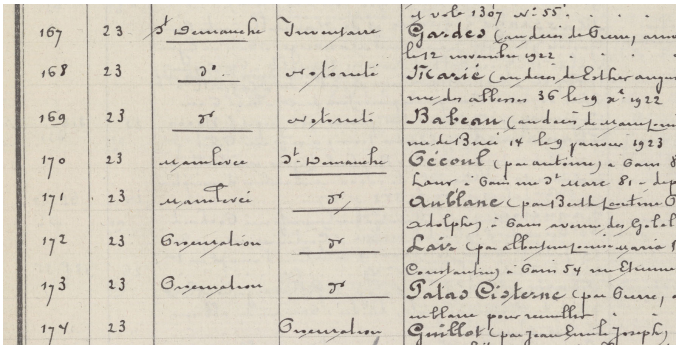

FIGURE 5. Exemple de mentions de notaire remplaçant et idem dans les colonnes 4 et 5.

— *Contenu* Cette colonne contient normalement des chiffres, de 1 à 31 : la date à laquelle l'acte enregistré a été passé. Pour les premiers jours du mois, on rencontre aussi "1°" ou "1e". On peut aussi trouver des objets du type "14 (10 et)" ou "27 (26 avril &).

Colonnes 4 & 5.

— *Structure* Ces colonnes contiennent normalement une à cinq lignes de texte, alignées en haut de la cellule. En principe, s'il y a du texte dans la colonne 4 la colonne 5 est vide, et vice-versa, mais il peut y avoir du texte dans les deux colonnes à la fois.

— *Contenu* Ces colonnes contiennent normalement le type de l'acte enregistré (contrat, vente, donation, procuration etc.) Cette information est contenue dans la colonne 4 lorsqu'il s'agit d'un acte en brevet (dont l'original est remis aux parties) et dans la colonne 5 pour les actes en minute (dont le notaire conserve l'original.)

Il y a du texte dans les colonnes 4 et 5 à la fois lorsqu'est indiquée une autre information que la nature de l'acte enregistré ; il s'agit le plus fréquemment de mentions du type "ste même jour" ou "ste 6 fevrier 1920" qui signalent que cette entrée dans le répertoire s'inscrit dans la continuité d'une entrée précédente.

Mais on trouve aussi d'autres indications : les mentions du type "substituant Me Demanche" qui signifie que l'acte a été enregistré par un notaire en remplaçant un autre ; lors d'un changement de mois en cours de page, il arrive que le mois et l'année soient inscrits dans la colonne 4 ou 5 ; enfin on rencontre aussi l'abbréviation "d°" qui signifie *idem*, ainsi que des guillemets " qui de même correspondent à la répétition de l'information à la ligne au-dessus.

Dans certains registres, on rencontre des pages où, en bas de page, dans la colonne 4 ou 5 est inscrit au crayon de papier le nombre total d'actes en minutes et en brevets pour la page en cours. On peut trouver ce décompte également en cours de page, lors d'un changement de mois.

Colonne 6.

— *Structure* Cette colonne contient en haut de page, sous l'en-tête imprimée, une ligne de texte de taille supérieure au reste, en partie imprimée. Sur le reste de la page elle contient un nombre variable de lignes de texte, d'une seule à parfois une vingtaine.

Dans chaque cellule, qui correspond à un acte, le premier mot ou groupe de mot est inscrit en caractères plus grands (et sur une partie des registres plus épais) que le reste du texte. Lorsque la dernière ligne de texte de la cellule n'est pas remplie par son contenu, elle est dans certains registres complétée par des points espacés.

— *Contenu* Cette colonne contient le nom et l'adresse des parties à l'acte, et peut également contenir d'autres informations comme le prix de vente d'un bien ou la date d'un décès.

Lorsqu'on change de mois ou d'année en cours de page, on peut trouver dans cette colonne une mention de date, centrée, en caractères plus épais et plus grands, parfois soulignée.

On rencontre également quelques cas particuliers où deux actes passés par les mêmes parties, et liés l'un à l'autre, ont été enregistrés ensemble : on a alors dans cette colonne 6 deux fois le nom de la partie à l'acte (C'est-à-dire la personne concernée par l'acte) en caractères épais et plus grands, comme deux actes différents, mais le contenu des autres colonnes est comme si un seul acte était enregistré sur cette ligne du tableau.

Colonne 7.

— *Structure* Cette colonne contient normalement une ligne de texte, alignée en bas de la cellule ; on

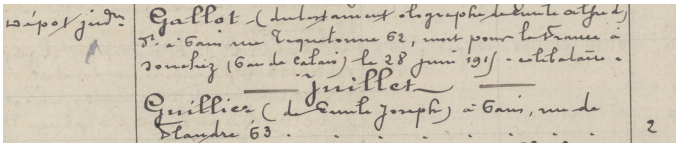

FIGURE 6. Exemple de mention de mois dans la colonne 6.

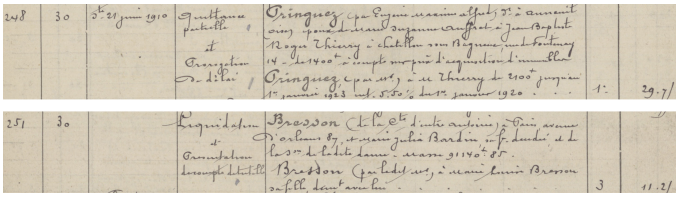

FIGURE 7. Exemples de deux actes rassemblés comme un seul dans la colonne 6.

peut cependant rencontrer deux lignes de texte, très rapprochées verticalement. Cette colonne est parfois vide ; on rencontre parfois plusieurs pages sans rien d'inscrit dans cette colonne.

— *Contenu* Cette colonne contient normalement des chiffres, entre 1 et 31 : la date à laquelle l'acte a été enregistré. On peut également y trouve des guillemets " qui signifient la répétition de l'information inscrite à la ligne précédente.

Lorsque la date d'enregistrement est un premier du mois, on peut rencontrer au lieu d'un "1" seul les formes 1e ou 1er, mais également une date sur deux lignes, comme "1er / mars" ou "juin / 2".

Lorsque la date d'enregistrement n'appartient pas au mois de la page en cours, le mois peut être indiqué en exposant ; par exemple : "4 xbr" (4 décembre).

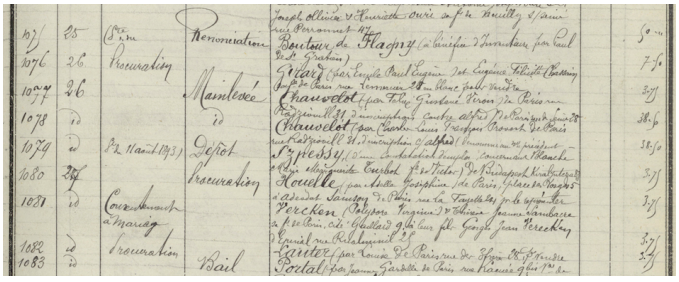

FIGURE 8. Exemple d'un cas où la colonne 7 est laissée vide.

Colonne 8.

— *Structure* Cette colonne contient normalement une ligne de texte, alignée en bas de la cellule. Elle est parfois laissée vide.

— *Contenu* Cette colonne contient normalement des chiffres : le montant des droits acquittés par les parties à l'acte auprès des services fiscaux pour la passation de l'acte enregistré chez le notaire. Elle peut également contenir les mentions "gratis" ou "assistance judiciaire", ainsi que des guillemets " signifiant la répétion de l'information inscrite à la ligne précédente.

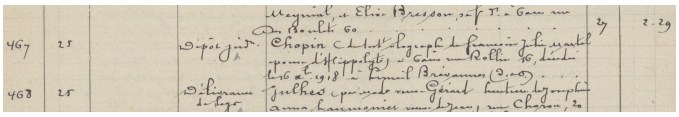

FIGURE 9. Exemple d'un cas où les colonnes 7 et 8 sont laissées vides pour un acte.

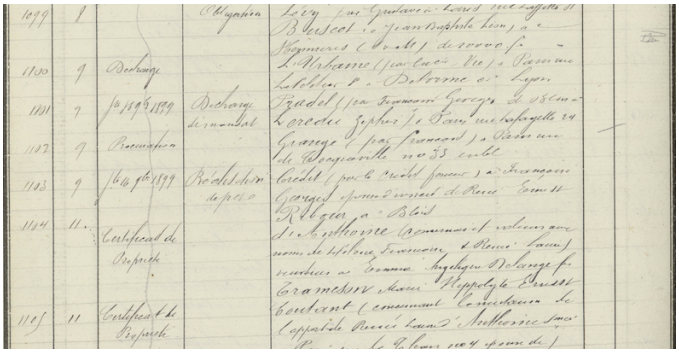

FIGURE 10. Exemple d'un cas où les colonnes 7 et 8 sont laissées vides.

Colonne 9.

— *Structure* Cette colonne est normalement vide. Nous l'avons tracée lors de notre segmentation au cas où se rencontreraient sur la colonne 8 le même type d'anomalies que sur la colonne 2, conduisant à ce que le texte se trouvant normalement dans la colonne 8 se trouve en dehors du tableau tracé sur les documents, et donc dans cette colonne 9.

— *Contenu* Cette colonne est normalement vide.

Il arrive que du texte soit à cheval sur plusieurs colonnes du tableau : il s'agit en général des mentions de vérification des registres. On rencontre aussi des mentions introductives, signalant notamment que le registre a été coté et paraphé par le président du Tribunal Civil de la Seine, qui occupent plusieurs colonnes.

En outre, les différents scribes de ces registres notariaux sont plus ou moins respectueux du tracé des colonnes imprimés, certains dépassant volontiers ces bordures.

B. Segmentation manuelle avec Transkribus

Pour pouvoir entraîner un modèle d'HTR, il a d'abord fallu produire des données d'entraînement, c'est-à-dire des transcriptions réalisées manuellement, qui supposaient en premier lieu la segmentation des pages choisies. Transkribus est un logiciel développé dans le cadre et grâce au financement du projet européen READ (*Recognition and Enrichment of Archival Documents*), permettant à la fois l'entraînement de modèles de reconnaissance automatique d'écriture, la transcription de documents et la recherche par mot-clef dans ces documents. Son développement a commencé en 2013, à l'époque sur un financement du

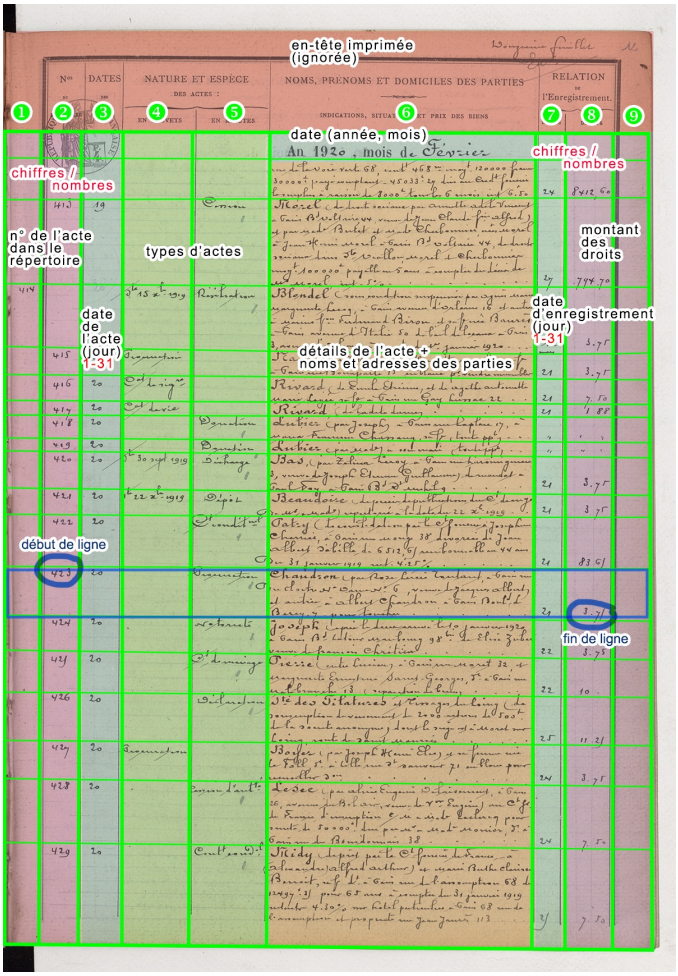

FIGURE 11. Structure d'un tableau.

projet européen tranScriptorium et désormais et jusqu'en 2020 grâce au projet européen READ. La plupart de ses fonctionnalités sont en *open source*, mais ce n'est pas le cas des modèles d'HTR sur lesquels l'utilisateur n'a pas directement la main. À l'heure actuelle, si on souhaite utiliser Transkribus pour HTRiser un ensemble de documents, on fournit aux personnes responsables de la maintenance de la plateforme les transcriptions, et ce sont eux qui entraînent un modèle et nous le transmettent ensuite.

En plus du logiciel que chacun peut installer sur son ordinateur, Transkribus propose également une interface en ligne qui permet de transcrire des documents déjà segmentés, sans pouvoir modifier cette segmentation.

Transkribus propose une reconnaissance automatique des zones de texte, mais celle-ci n'est pas très performante sur les documents à structure tabulaire, et particulièrement sur des documents comme les nôtres ; les colonnes pouvaient être reconnues [2], mais les lignes des actes ne l'étaient pas du tout.

Nous avons donc opéré un découpage manuel, en traçant d'abord des *table regions* puis des colonnes et des lignes, avant de lancer la reconnaissance des lignes de texte (avec l'outil *Find Lines in Text Regions*.) Contrairement à l'analyse de la structure tabulaire, cette détection des lignes de texte est assez performante et n'a nécessité qu'assez peu de corrections.

Si ce travail de segmentation manuelle n'est pas difficile, il est néanmoins chronophage et il ne serait pas envisageable de le réaliser sur l'ensemble des documents ; c'est pourquoi nous avons développé une solution de segmentation automatique des images.

C. Segmentation automatique par traitement d'images

Identification des colonnes. La segmentation automatique des images se fait en plusieurs étapes d'une procédure simple adaptée à ce type de documents :

- Au niveau vertical, on détecte d'abord la ligne qui limite l'entête du tableau ;
- L'orientation de l'image est ensuite corrigée (car les images scannées ne sont généralement pas parfaitement droites) ;
- On détecte ensuite les coins de l'entête du tableau ;
- Au niveau horizontal, on détecte d'abord la colonne centrale, où est concentrée la plupart du texte ;
- On déduit de la position de la colonne centrale celle des autres colonnes, puisque leur largeur est constante ;
- Ce sont ensuite les lignes de texte qui sont détectées, et c'est à partir de ces lignes et des caractéristiques de nos documents qu'il est possible de tracer les lignes du tableau.

La segmentation se fait grâce à l'analyse du profil de projection verticale et horizontale des images binarisées, adjointe à notre connaissance de la structure des documents qui est très stable. La binarisation est un traitement de l'image afin qu'elle ne comporte plus que deux classes de pixels : des pixels noirs et des pixels blancs. La projection, verticale ou horizontale, consiste en la production d'un histogramme représentant la répartition des pixels noirs sur l'image. Cette projection est orientée, c'est-à-dire que l'image est lue dans un sens ou dans un autre — ici, verticalement ou horizontalement. L'analyse du profil de projection verticale permet de corriger l'orientation de l'image. La projection horizontale permet de déterminer où se trouve la colonne 6, c'est-à-dire celle qui contient le plus de texte. Sur les registres numérisés dont nous disposons, la distance entre les différentes colonnes est relativement stable, ce qui permet de tracer les autres colonnes à partir de cette colonne centrale.

Identification des lignes. L'identification des lignes de texte quant à elle se fait grâce à une version de l'algorithme de *seam carving* [3] particulièrment adaptée aux manuscrits réguliers [4]. Cet algorithme de *seam carving* commence par calculer l'énergie de chaque pixel, une valeur qui

dépend du contraste entre ce pixel et les pixels voisins. Une *energy map* (une "carte" des énergies de l'image, similaire à une projection horizontale) est ensuite établie, et sur celle-ci l'agorithme recherche récursivement le chemin le moins "coûteux", c'est-à-dire celui qui passe par le moins de pixels noirs (qui ont la valeur d'énergie la plus forte), d'un bout à l'autre de la page, de gauche à droite. Ces chemins à basse énergie correspondent aux espaces blancs entre les lignes de texte, et définissent donc ces lignes en négatif. [5]
Cet algorithme *seamcarve heartbeat* n'est appliqué qu'à la colonne 6, celle dans laquelle du texte est présent sur toute la hauteur de la cellule, contrairement aux autres colonnes qui ne contiennent que quelques mots ou chiffres, normalement en haut ou en bas de la cellule.
On opère différemment pour les lignes de texte ou de chiffres dans les colonnes 1-5 et 7-9 : on applique d'abord un filtre gaussien (c'est-à-dire qu'on rend l'image floue, ce qui permet d'obtenir des formes moins précises et donc où les formes des différents caractères, par exemple deux chiffres, se touchent et forment ainsi des composantes connectées) avant de les binariser. On obtient ainsi des composantes connectées, à peu près un polygone par chiffre, nombre ou mot, qu'on peut aligner à une ligne de la colonne 6 grâce à la position verticale de son centroïde (le centre de gravité d'un polygone, l'intersection de ses diagonales) [6]. Il arrive que des objets appartenant à des lignes différentes soient faussement régroupés dans un même polygone ; on peut résoudre ce problème par un découpage en fonction de maxima déterminés par rapport à la distance moyenne entre les lignes. L'avantage de cette méthode est qu'elle permet d'avoir les hampes et les descendantes dans un même polygone.

Reconstruire les actes. Une fois les lignes de texte obtenues, il faut reconstituer les lignes du tableau, c'est-à-dire rassembler ensemble les lignes de texte qui se rapportent à un même acte dans les différentes colonnes. Cette reconstitution est possible grâce à ce que nous avons établi quant à la structure des documents.
La situation la plus classique est la suivante : la ligne du tableau correspondant à un acte commence par un numéro d'ordre dans la colonne 2 (ou éventuellement dans la colonne 1) suivi au même niveau par la date à laquelle l'acte a été passé dans la colonne 3, puis la nature de l'acte dans la colonne 4 ou 5. On a ensuite la partie principale dans la colonne 6, le nom des parties et autres informations sur l'acte, qui peut contenir un nombre variable de lignes de texte, suivie de la date d'enregistrement (col. 7) de l'acte et des droits perçus par les services fiscaux (col. 8 ou 9) lors de sa passation.
Ainsi, le texte contenu dans les colonnes précédant la colonne 6 est aligné, avec plus ou moins de précision, avec la première ligne du texte de la colonne principale, et le texte des colonnes qui la suivent est aligné avec la dernière ligne de texte de cette colonne 6. Par ailleurs, le premier mot ou groupe de mot du texte de la colonne 6 pour chaque acte, c'est-à-dire le nom de la partie principale à l'acte enregistré, est écrit en caractères "gras" et de taille supérieure par rapport au reste de la cellule, ce qui constitue un autre indice pour identifier le début d'un acte. Ainsi on peut tracer les lignes du tableau à partir des lignes de texte des différentes colonnes : en haut, en suivant les lignes de texte des colonnes 0/1, 2 et 3/4 qu'on aligne avec une ligne de la colonne 6 (celle qui commence par des caractères plus grands et plus épais) ; en bas en suivant les lignes de texte des colonnes 7 et 8/9, qu'on aligne avec une ligne de la colonne 6.
Il peut arriver tout d'abord qu'un acte commence sur une page et se termine sur la suivante : dans ce cas sur la première page la ligne est délimitée par la fin du tableau, et sur la page suivante l'acte est délimité en haut par l'indication du mois et de l'année.

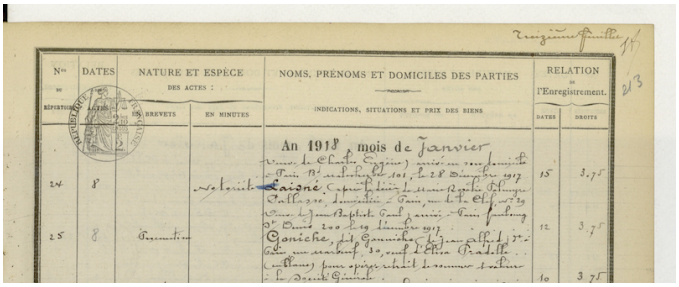

FIGURE 12. Exemple d'un acte se terminant en haut d'une page

IV. CLASSIFICATIONS

Pour certaines colonnes, dont on sait qu'elles ne peuvent contenir qu'un ensemble fini d'éléments, il n'est pas nécessaire de recourir à l'HTR, que l'on peut réserver à la colonne 6. Il est en effet possible de procéder par classification, c'est-à-dire en ne reconnaissant pas des caractères individuels pour former des mots, mais des formes : on fournit à un *classifier* des classes, c'est-à-dire des exemples de forme et les mots ou nombres (dans notre cas) qui leur correspondent, et ce *classifier* peut ensuite associer les formes qu'il détecte à ces classes.

Colonnes de chiffres. Pour les colonnes contenant des nombres, les composantes connexes (l'ensemble des chiffres composant ces nombres) sont détectées grâce à un histogramme de gradient orienté (*histogram of oriented gradient*), une technique d'abord développée pour détecter les piétons dans des images et qui présente donc l'intérêt de pouvoir identifier des objets dont la forme n'est pas constante [7].
Les composantes connexes ainsi obtenues sont associées à des classes, et cette classification est corrigée manuellement avant que soit entraîné sur la base de données résultante un réseau de neurones convolutif (*Convolutional Neural Network* ou *CNN*) qui est alors capable de reconnaître y compris des images de nombres absents de ses données d'entraînement.

Les types d'actes. Pour les colonnes 4 et 5, qui contiennent le type de l'acte enregistré, on procède également par classification. Sur la base des pages transcrites, on fournit à un *classifier* un certain nombre d'exemples de formes correspondant à des classes, avec leurs variations (par exemple, on trouve à la fois "cont^{t} cond^{nel}" et "cont^{t} condit^{nel}" pour contrat conditionnel) ; après entraînement et correction, le CNN peut répartir les formes détectées sur les images entre les différentes classes.

V. HTR

A. La reconnaissance automatique d'écriture manuscrite

Les premières méthodes de reconnaissance automatique d'écriture adaptées aux écritures manuscrites (*Handwritten Text Recognition*, ou HTR) sont les modèles statistiques, apparus dans les années 1950, pour prendre en compte la variabilité de la forme des caractères et des mots dans l'écriture manuscrite, qui rendait peu efficace l'application de masques fixes qui pouvaient fonctionner pour des documents imprimés. Ces méthodes s'apparentent à de la classification, et procèdent en segmentant les images en lettres ou parties de lettres. On utilise par exemple la méthode des *k plus proches voisins* : après binarisation, une représentation vectorielle des images est produite et la reconnaissance des caractères se fait en comparant la forme détectée à celle des k (un nombre de voisins à définir) formes connues qui en sont les plus proches (où cette "distance" est également un paramètre à définir.)
Un des principaux désavantages de ces méthodes est qu'elles nécessitent une segmentation en caractères ou fragments de caractères, qui dans le cas de l'écriture manuscrite ne peut pas être univoque.
À partir des années 1980, on applique à la reconnaissance d'écriture les *modèles de Markov cachés* (MMC), des modèles statistiques ayant fait leurs preuves pour la reconnaissance de la parole. Ils permettent d'obtenir une meilleure segmentation, et l'incorporation de modèles linguistiques. Ces MMC sont combinés à des *classifiers* statistiques (réseaux de neurones) dans la continuité de la période précédente, et ces techniques hybrides sont appliquées avec succès dans deux domaines principaux : la reconnaissance des adresses pour le tri automatisé du courrier, et la lecture des chèques.
Aujourd'hui, les outils de reconnaissance d'écriture manuscrite les plus performants fonctionnent sur la base de *réseaux de neurones récurrents* (*Recurrent Neural Networks* ou RNN), des réseaux de neurones non-linéaires, qui comportent des boucles les dotant d'une "mémoire" allant au-delà de l'événement immédiatement antérieur, et par conséquence des systèmes où les décisions sont prises non pas uniquement en fonction de la décision précédente, mais d'une chaîne plus longue de précédents ; ce fonctionnement dynamique permet une meilleure appréhension du langage, où pour prédire le mot ou le caractère suivant il est nécessaire de connaître plus que celui qui précède immédiatement.
Plus précisément, les réseaux de neurones utilisés pour la reconnaissance d'écriture manuscrite sont des réseaux *Long Short Term Memory* (LSTM) qui ont la capacité de garder en mémoire des événements "lointains" sans perte d'information, ce qui n'est pas le cas des RNN classiques [8].
Ces technologies de reconnaissance d'écriture manuscrite reposent sur un modèle optique (qui reconnait des formes de mots, morceaux de mots ou caractères) et un modèle linguistique (probabiliste, où des n-grammes définissent les enchainements de mots ou de caractères possibles, à une distance n du mot ou caractère en question.) On peut également leur associer des dictionnaires ou des lexiques.
Un modèle d'HTR est entraîné sur des données d'apprentissage, c'est-à-dire des transcriptions manuelles vérifiées qui servent de *ground truth* ou *vérité terrain*, auxquelles on compare les résultats de la reconnaissance automatique et sur la base desquelles ont effectue des corrections.
Les performances d'un modèle sont évaluées par la comparaison entre les résultats de l'HTR et une *ground truth* : on calcule le *Character Error Rate* (CER) et le *Word Error Rate* (WER), c'est-à-dire le taux d'erreurs concernant les caractères et celui concernant les mots, en prenant en compte les trois types d'erreurs : les caractères mal reconnus, ceux qui sont absents et ceux qui sont introduits par l'HTR sans être effectivement présents. [9] [10] [11]

B. Obtenir des données d'entraînement : Travail collaboratif avec Transkribus

En l'absence de trancriptions préexistantes, nous avons choisi de produire d'abord une cinquantaine de pages transcrites pour entraîner un modèle de reconnaissance d'écriture. Dans cette optique, il importe donc d'avoir des transcriptions qui respectent ce qui est effectivement dans les documents (on ne développe pas les abbréviations, on ne normalise pas l'usage des majuscules ou des accents . . .) et d'établir des normes qui seront suivies par les différents transcripteurs. Il est également important, même si les transcriptions sont relues par la suite, de ne pas utiliser par exemple des points d'interrogation pour signaler une incertitude quant à un mot ou un caractère ; si un mot est illisible, il est préférable de ne pas transcrire la ligne du tout : elle ne sera alors pas prise en compte comme donnée d'entraînement.
Pour réaliser ces transcriptions et obtenir un premier modèle d'HTR à tester, nous nous sommes tournés vers Transkribus, du fait de son interface utilisateur relativement facile à prendre en main et qui permet des transcriptions collaboratives et conserve les versions antérieures d'un document à chaque sauvegarde.

a) Importation des images numérisées dans Transkribus. Les images numérisées originales issues de la campagne de numérisation des registres notariaux sont des images de doubles pages ; pour simplifier les étapes suivantes, il nous a semblé opportun de diviser ces images en deux, afin d'avoir une page de registre par image.
On verse les images découpées dans Transkribus en créant une *collection* ; on peut ensuite inviter d'autres utilisateurs dans celle-ci. Le créateur de la collection a le statut

d'*Owner* : il a tous les droits sur cette collection. Les autres types d'utilisateurs sont les *Editors* et les *Transcribers* ; les premiers peuvent transcrire les documents qu'ils ont été invités à transcrire et ajouter d'autres *transcribers*, mais ne peuvent pas ajouter ou supprimer des documents ; les seconds ne peuvent que transcrire les documents sur lesquels ils ont été invités [12].

b) Segmentation. On trace d'abord sur chaque page, grâce à l'outil *Table*, une *Table Region* qui exclut l'en-tête imprimée mais inclut les marges à gauche et à droite de la première et de la dernière colonne, puisque qu'on ajoute deux colonnes dans notre table virtuelle par rapport aux six colonnes du formulaires imprimé. En sélectionnant la *Table Region* qu'on vient de tracer, on utilise l'outil intitulé *Splits a shape with a vertical line* pour tracer les colonnes, puis l'outil *Splits a shape with a horizontal line* pour tracer les lignes du tableau (qui délimitent les actes.)
En sélectionnant à nouveau toute la table ainsi tracée (en passant par l'onglet *Layout*) on utilise ensuite l'outil de détection automatique de lignes de texte de Transkribus : dans l'onglet *Tools*, en choisissant l'option *CITlab Advanced*, on décoche *Find Text Regions* pour ne détecter que les lignes et on lance l'analyse.
On peut avoir besoin de corriger des lignes mal détectées, ajouter des lignes qui n'ont pas été tracées, ou de déplacer des lignes qui ne seraient pas dans la bonne cellule de la table virtuelle, ce qui est possible dans l'onglet *Layout*. Pour tracer une ligne, le meilleur outil est l'outil *baseline*, puisque que la *line* qui va avec est créée en même temps.

c) Transcription. Pour transcrire, il faut d'abord sélectionner une cellule ; on entre ensuite sa transcription dans l'éditeur de texte situé en-dessous de l'image, aux lignes correspondantes. L'éditeur de texte permet également d'ajouter des annotations.

Note : les états / labels des documents. Transkribus permet d'attribuer 5 labels différents aux images : *New*, *In Progress*, *Done*, *Final* et *Ground Truth*. Lorsqu'on importe un document dans Transkribus, les images ont toutes par défaut le statut *New*. Pour faciliter la transcription et la relecture, on peut par exemple décider de marquer comme *In Progress* les pages qui ont été segmentées et où il n'y a pas encore de texte transcrit, ou où la transcription n'est pas terminée ; *Done* pour les pages entièrement transcrites ; *Final* ou *Ground Truth* pour les pages transcrites et relues.

Les développeurs de Transkribus recommandent de commencer à entraîner un modèle d'HTR avec entre 5000 et 15000 mots, c'est-à-dire entre 25 et 75 pages. Il peut également être opportun de transcrire quelques pages supplémentaires qui ne serviront pas de données d'entraînement mais de données de test, pour évaluer les résultats du modèle entraîné sur des pages qu'il ne "connait" pas.

d) Entraînement. Dans le cas de Transkribus, nous ne procédons pas par défaut nous-mêmes à l'entraînement d'un modèle de reconnaissance d'écriture (il est néanmoins possible de demander l'accès à cette fonctionnalité [13]) ; une fois qu'un nombre suffisant de pages a été transcrit, on envoie un e-mail à email@transkribus.eu (ou à l'un des membres de l'équipe Transkribus avec lequel on est en contact) en indiquant la collection et le document où se trouvent les pages transcrites, et l'équipe Transkribus se charge alors d'entraîner un modèle d'HTR.
Une fois ce modèle entraîné, un membre de l'équipe de Transkribus nous répond pour nous donner les résultats (le taux d'erreur sur les caractères et le taux d'erreur sur les mots) obtenus par le modèle sur les données d'entraînement et sur des données réservées comme données de test, ainsi que l'accès à ce modèle dans Transkribus. On peut ensuite utiliser le modèle entraîné pour reconnaitre le texte de n'importe quelle image importée dans Transkribus.

e) Tests. Après avoir fait entraîner un premier modèle sur une quarantaine de pages du répertoire coté MC/RE/XLIII/43 de l'étude Marotte, dont le CER était autour de 13,5%, nous avons obtenu un deuxième modèle, entraîné sur une dizaine de pages supplémentaires, avec de meilleurs résultats : un CER autour de 10%.

Page	Modèle Marotte_M1		Modèle Marotte_M2	
	Word Error Rate (WER)	Character Error Rate (CER)	WER	CER
Marotte 43, image 5, gauche	35,34	14,30	25,86	10,25
Marotte 43, image 5, droite	44,08	17,09	39,02	14,35
Marotte 43, image 25, gauche	41,04	13,61	37,22	12,17
Marotte 43, image 25, droite	41,48	12,62	37,21	11,40
Average Rate	0.920	0.882	0.477	0.539

TABLE I. TRANSKRIBUS, ÉVALUATION DES RÉSULTATS, RÉPERTOIRE MC/RE/XLIII/43 ÉTUDE MAROTTE

Résultats. Nous avons également effectué des tests du modèle Marotte_M2 sur quelques pages d'autres registres, et donc d'autres écritures que celle sur laquelle ce modèle a été entraîné ; les résultats ne sont pas bons. Il semble probable que pour réaliser une HTR sur l'ensemble des registres, il faudra au moins fournir au modèle utilisé quelques pages de chaque écriture pour obtenir des résultats corrects.

Page	Modèle Marotte M2	
	Word Error Rate (WER)	Character Error Rate (CER)
MC/RE/CXIII/20, image 100, page de droite	85.89	45.58
MC/RE/CXIII/22, image 65, page de droite	89.46*	52*
MC/RE/CXIII/19, image 77, page de gauche	84.69*	39.11*

* calculés avec wer++ [14]

TABLE II. ÉVALUATION DU MODÈLE ENTRAÎNÉ SUR 50 PAGES DU REGISTRE MAROTTE MC/RE/XLIII/43 SUR DES PAGES D'AUTRES REGISTRES

C. Le passage en production avec Kraken

Kraken est un logiciel d'OCR et HTR open-source, développé à partir d'OCRopus par Benjamin Kiessling,

membre de l'équipe ALMAnaCH. Contairement à OCRopus, qui rassemble différents outils d'analyse de documents, Kraken est un logiciel "clés-en-main" qui prend en charge la binarisation des images, la segmentation, l'entrainement d'un modèle d'OCR/HTR et la reconnaissance d'écriture avec ce modèle [15]. D'abord développé pour les documents imprimés (en caractères latins ou autres, par exemple en arabe [16]), Kraken a produit de bons résultats sur des manuscrits, latins et hébreux médiévaux notamment.
Comme Transkribus, Kraken fonctionne grâce à des réseaux de neurones récurrents, et contrairement à Transkribus il est entièrement *open-source*, y compris ses modèles d'OCR/HTR. En revanche, Kraken (pour l'instant) ne peut être utilisé que grâce à des lignes de commandes entrées dans un terminal : il n'y a pas d'interface utilisateur ; par ailleurs, il ne peut pas être installé sous Windows.
Une rapide expérimentation menée avec les données d'entraînement du registre Marotte (MC/RE/XLIII/43), qui a mis en évidence à la fois des promesses et des difficultés potentielles :

— *Des difficultés liées à la segmentation.* Les différents systèmes d'HTR n'utilisent pas tous la segmentation de la même manière pour reconnaître l'écriture. Si Transkribus utilise la *baseline* pour reconnaître la ligne de texte, Kraken, quant à lui, découpe chaque ligne en une image rectangulaire (une *box*) en prenant les plus grandes valeurs parmi les coordonnées de la ligne. Ainsi, l'interopérabilité de la segmentation entre les différents systèmes testés n'est pas optimale. En revanche, il est évident que la solution manuelle utilisée jusqu'à présent, via Transkribus, ne pourra être utilisée en production.
— *Un modèle potentiellement très efficace.* Malgré cette importante difficulté, ce premier test a montré que la reconnaissance d'écriture manuscrite était plutôt correcte (25% de taux d'erreur sur les caractères pour un premier essai.) Ceci laisse présager d'une précision supérieure à celle des modèles de Transkribus une fois l'obstacle de la segmentation levé. La seconde phase du projet devrait permettre d'obtenir des résultats plus complets et plus précis.

D. Partenariat potentiel avec le projet Tikkoun Sofrim

Le projet *Tikkoun Sofrim — Digitization of HTR Hebrew Manuscripts by Adaptive Crowdsourcing* a été lancé en juillet 2018 sous la direction de Daniel Stökl Ben Ezra, directeur d'études à l'EPHE et également membre de l'équipe ALMAnaCH d'Inria. Ce projet porte entre autres sur l'application des technologies d'HTR aux mauscrits hébreux.
Dans le cadre de ce projet sera développée une interface permettant à la fois de corriger la segmentation en lignes et zones de texte des documents et de réaliser et corriger la trasncription de ceux-ci, ainsi qu'un outil de création de vérité terrain (*ground truth*) ou données d'entraînement. Ces outils seront développés sur la base de Kraken pour la reconnaissance d'écriture, ainsi qu'Archetype et LAREX. Archetype [1] est un *framework* rassemblant des outils pour l'annotation et la description d'image, d'abord développé pour l'analyse paléographique et écritures anciennes. LAREX [2] est un outil de segmentation semi-automatique de documents, spécialisé dans la segmentation des documents anciens, manuscrits ou imprimés.
Pour la réalisation de ces outils, le projet Tikkoun Sofrim recrutera deux ingénieurs ; les responsables de ce projet estiment que si un troisième ingénieur était recruté pour y travailler, l'interface pourrait être développée d'ici février 2019. Le projet LECTAUREP pourrait, pour bénéficier de cette interface, s'associer à Tikkoun Sofrim et recruter ce troisième ingénieur ; le projet disposerait ainsi d'une interface opérationnelle tout en utilisant la reconnaissance d'écriture de Kraken.

VI. Keyword Spotting

L'un des principaux usages attendu des résultats de la reconnaissance d'écriture est la recherche par mot-clé dans les répertoires d'actes de notaires. Dans ce cas, la piste du *Key Word Spotting* semble assez pertinente, puisque cette méthode permet de chercher des motifs textuels avec précision sans besoin d'une transcription parfaite. Le *Keyword Spotting* (KWS) fonctionne soit en recherchant une chaîne de caractères (*query by string*, comme un moteur de recherche classique) soit en recherchant une image de forme (*query by example.*)
Un système de KWS peut fonctionner sur la base d'une modèle optique seul, ou grâce à un modèle optique associé à un lexique (*lexicon-based KWS.*) Les modèles avec lexique nécessitent des transcriptions pour être entraînés, mais un *Keyword Spotting* sans lexique peut fonctionner sur des images qui ne sont pas encore transcrites. L'intérêt du KWS par rapport à la recherche approximative (*fuzzy search*) est donc de rendre possible la recherche dans les documents avant qu'ils ne soient transcrits — en effet la recherche approximative nécessite une transcription, même si elle n'a pas besoin d'être de très bonne qualité.
Cependant, les modèles de KWS qui associent au modèle optique un lexique produisent évidemment de meilleurs résultats. Dans le cadre de ce projet d'HTR sur les répertoires des notaires de Paris, un outil de KWS pourrait permettre des recherches, avec des résultats plus ou moins précis, à la fois dans les documents transcrits, avec des transcriptions de plus ou moins bonne qualité, et dans ceux qui ne le sont pas encore.
Transkribus propose un outil de *Keyword Spotting*, mais celui-ci n'est accessible qu'à partir du logiciel et ne pourrait pas être transféré à une plateforme en ligne de consultation des répertoires numérisés.
Un exemple intéressant d'utilisation du *Keyword Spotting* est le moteur de recherche du projet HIMANIS [3], qui permet aux utilisateurs d'évaluer les résultats d'une recherche, en signalant si ils correspondent ou non aux termes recherchés

1. http ://archetype.ink/
2. https ://github.com/chreul/LAREX
3. https ://www.himanis.org/

[17] ; les utilisateurs contribuent ainsi à l'amélioration de l'outil.

VII. Conclusion

Les répertoires des notaires de Paris conservés par le département du Minutier central sont tous numérisés, mais n'ont pas tous fait l'objet de la même campagne de numérisation. En effet, les répertoires numérisés avant 2012 l'ont été à partir de microfilms, tandis que les répertoires numérisés après 2012 ont été numérisés à partir des documents originaux et en couleur. Il faudra effectuer des tests sur des images de répertoires numérisés à partir de microfilms afin de déterminer s'il sera ou non nécessaire de procéder à une nouvelle campagne de numérisation pour ceux-ci ; cela dépendra également de la qualité de ces numérisations en noir et blanc dans l'optique de leur mise à disposition sur la plateforme en ligne.

La diversité des écritures dans les répertoires représente une difficulté, mais différentes solutions peuvent être envisagées, dont il faut évaluer le coût ; effectuer en interne les transcriptions nécessaires pour qu'un modèle d'HTR soit entraîné sur toutes les écritures rencontrées n'apparaît pas comme une option viable, mais il pourrait être possible de ne fournir au modèle que des échantillons d'une partie de ces écritures et d'évaluer si les résultats ainsi obtenus sont corrects. On pourrait essayer de recenser les écritures et de déterminer des ensembles de mains similaires, pour rationnaliser ce processus, mais une telle démarche prendrait nécessairement un certain temps. On peut aussi envisager la mise en place d'une plateforme de transcription collaborative, comme celle mise en place pour les testaments de poilus[4], où des transcriptions réalisées par des lecteurs permettraient d'améliorer les performances du modèles d'HTR — mais il faut alors considérer le coût de la mise en place d'une telle plateforme, et de la relecture et correction éventuelle par des transcripteurs "accrédités", ainsi que la possibilité de constituer une communauté active pour l'alimenter.

Une piste pour l'amélioration des résultats de l'HTR est aussi la fourniture de référentiels de noms propres et de noms de lieux, qui permettront d'améliorer le modèle linguistique. Quoiqu'il en soit, de nouvelles transcriptions seront nécessaires.

Malgré ces difficultés, et d'autres questions qui restent en suspens comme celle de l'hébergement de la plateforme de consultation des répertoires (qui devra pouvoir héberger les images dans une résolution suffisamment haute pour qu'elles soient exploitables par les lecteurs), cette phase exploratoire a produit des résultats encourageants quant aux performances d'un modèle de reconnaissance automatique d'écriture sur ce corpus des répertoires des notaires parisiens. De plus la question de la segmentation des documents, qui avait d'abord posé problème, est désormais résolue. Un important travail humain sera encore nécessaire pour produire un modèle de reconnaissance d'écriture performant, mais il s'agit là toujours d'un préalable nécessaire dans ce genre de projet, surtout lorsque les documents ont des auteurs multiples, et ce travail exploratoire a permis de préciser les conditions nécessaires à l'amélioration des résultats obtenus.

4. https ://testaments-de-poilus.huma-num.fr/

Références

[1] *Loi de l'enregistrement du 22 frimaire an 7 (12 décembre 1798) (Deuxième édition...) / commentée... par M. Perry,...*, 1853. [Online]. Available : https://gallica.bnf.fr/ark:/12148/bpt6k65686849

[2] M. Diem, S. Fiel, and F. Kleber, "D6.5 basic layout analysis p2," p. 11.

[3] S. Avidan and A. Shamir, "Seam Carving for Content-Aware Image Resizing," p. 9.

[4] M. Seuret, D. S. B. Ezra, and M. Liwicki, "Robust heartbeat-based line segmentation methods for regular texts and paratextual elements," in *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, Kyoto, Japan, November 10-11, 2017*. ACM, 2017, pp. 71–76. [Online]. Available : http://doi.acm.org/10.1145/3151509.3151521

[5] R. Saabni and J. El-Sana, "Language-Independent Text Lines Extraction Using Seam Carving." IEEE, Sep. 2011, pp. 563–568. [Online]. Available : http://ieeexplore.ieee.org/document/6065374/

[6] C. Yi and Y. Tian, "Text String Detection from Natural Scenes by Structure-based Partition and Grouping," *Ieee Transactions on Image Processing*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011. [Online]. Available : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337634/

[7] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," vol. 1. IEEE, 2005, pp. 886–893. [Online]. Available : http://ieeexplore.ieee.org/document/1467360/

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Terme Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available : https://www.mitpressjournals.org/doi/pdf/10.1162/neco.1997.9.8.1735

[9] W. Swaileh, "Language Modelling for Handwriting Recognition," Ph.D. dissertation, Normandie Université, 2017.

[10] L. Mioulet, "Reconnaissance de l'écriture manuscrite avec des réseaux récurrents," Ph.D. dissertation, Université de Rouen, 2005.

[11] V. Romero, N. Serrano, A. H. Toselli, J. A. Sanchez, and E. Vidal, "Handwritten Text Recognition for Historical Documents," p. 7.

[12] Users guide - Transkribus Wiki. [Online]. Available : https://transkribus.eu/wiki/index.php/Users_guide

[13] How to train a HTR model in Transkribus. [Online]. Available : https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf

[14] N. S. Martínez-Santos, "WERpp : Calculates the Word Error Rate between two text files," Jul. 2018, original-date : 2012-02-27T17 :08 :26Z. [Online]. Available : https://github.com/nsmartinez/WERpp

[15] J. Hietbrink. Kraken, the unknown python OCR system. [Online]. Available : https://www.webuildinternet.com/2016/10/01/kraken-the-unknown-python-ocr-system/

[16] M. Romanov, S. Bowen Savant, M. T. Miller, and B. Kiessling, "Important New Developments in Arabographic Optical Character Recognition (OCR)," 2016. [Online]. Available : https://www.academia.edu/28923960/Important_New_Developments_in_Arabographic_Optical_Character_Recognition_OCR_

[17] T. Bluche, S. Hamel, C. Kermorvant, J. Puigcerver, D. Stutzmann, A. H. Toselli, and E. Vidal, "Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project." IEEE, Nov. 2017, pp. 311–316. [Online]. Available : http://ieeexplore.ieee.org/document/8269990/