Solenn Tual, Nathalie Abadie, Bertrand Duménieu,
Joseph Chazalon, Edwin Carlinet

Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du XIX^{e} siècle : application aux métiers de la photographie

Volume 6, n^{o} 1-2 (2025), p. 179-200.

https://doi.org/10.5802/roia.98

La Revue Ouverte d'Intelligence Artificielle est membre du Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 2967-9672

Revue Ouverte d'Intelligence Artificielle
Volume 6, nº 1-2, 2025, 179-200

Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du XIX^{e} siècle : application aux métiers de la photographie

Solenn Tual[a], Nathalie Abadie[a], Bertrand Duménieu[b], Joseph Chazalon[c], Edwin Carlinet[c]

[a] LASTIG, Université Gustave Eiffel, IGN-ENSG, 73 Avenue de Paris, 94165 Saint-Mandé Cedex (France)
E-mail : solenn.tual@ign.fr, nathalie-f.abadie@ign.fr
URL *:* https://www.umr-lastig.fr/solenn-tual/
URL *:* https://www.umr-lastig.fr/nathalie-abadie/

[b] Centre de Recherches Historiques, EHESS, 54 Boulevard Raspail, 75006 Paris (France)
E-mail : bertrand.dumenieu@ehess.fr
URL *:* http://crh.ehess.fr/index.php?5206

[c] LRE, EPITA, 14-16 rue Voltaire, 94270 Le Kremlin-Bicêtre (France)
E-mail : joseph.chazalon@epita.fr, edwin.carlinet@epita.fr
URL *:* https://www.lrde.epita.fr/wiki/User:Chazalon
URL *:* https://www.lrde.epita.fr/wiki/User:Carlinet.

Résumé. — Les annuaires professionnels anciens, édités à un rythme soutenu dans de nombreuses villes européennes tout au long des XIX^{e} et XX^{e} siècles, forment un corpus de sources unique par son volume et la possibilité qu'ils donnent de suivre les transformations urbaines à travers le prisme des activités professionnelles des habitants, de l'échelle individuelle jusqu'à celle de la ville entière. L'analyse spatio-temporelle d'un type de commerces au travers des entrées d'annuaires demande cependant un travail considérable de recensement, de transcription et de recoupement manuels. Pour pallier cette difficulté, cet article propose une approche automatique pour construire et visualiser un graphe de connaissances géohistorique des commerces figurant dans des annuaires anciens. L'approche est testée sur des annuaires du commerce parisien du XIX^{e} siècle allant de 1798 à 1914, sur le cas des métiers de la photographie.

Mots-clés. — Graphe de connaissances géohistorique, annuaires anciens, reconnaissance et résolution d'entités nommées, bruit OCR, visualisation spatio-temporelle.

1. Introduction

À partir de la fin du XVIII^e siècle, les annuaires des habitants et des commerces (voir figure 4.1), sortes de « pages blanches » et « pages jaunes » avant l'heure, ont connu un succès croissant et ont été édités pour de nombreuses villes européennes et nord-américaines. Ils recensent les habitants, leurs activités professionnelles et leurs localisations, et constituent des sources historiques extrêmement riches pour suivre les évolutions urbaines, de l'échelle individuelle à celle de la ville. Ainsi, [14] procède à une analyse systématiques de la rubrique *Marchands de tableaux* des annuaires du commerce de Paris allant de 1815 à 1955 pour étudier la dynamique globale de ce commerce et les évolutions de la profession. L'extraction et la structuration des données des annuaires est réalisée manuellement, l'accent étant mis ici sur l'étude de la fiabilité des sources et l'analyse sérielle des données : stocks, flux, et évolution des pratiques de la profession vue au travers des termes choisis pour désigner l'activité des commerces d'art. [4] évalue le potentiel des annuaires des habitants de Berlin de 1880 pour réaliser des études socio-économiques et démographiques en l'absence de données de recensement. Cette étude, réalisée pour une unique date, démontre l'utilisabilité des entrées extraites automatiquement et suggère, en perspectives, de lier les entrées similaires d'une édition à l'autre et d'étendre les analyses aux annuaires du commerce. [6] utilise des annuaires parus entre 1936 et 1990 pour localiser les anciennes stations services de la ville de Providence (Rhode Island) aux États-Unis, afin de détecter des zones potentiellement polluées. L'approche proposée par [6] comprend plusieurs étapes, qu'elle vise à automatiser le plus possible : analyse de la mise en page des annuaires, reconnaissance optique de caractères (OCR), reconnaissance des entités nommées, réalisée ici à l'aide de patrons lexico-syntaxiques, et géocodage à l'aide d'une base d'adresses récente. Le fait de retrouver une même station service dans plusieurs annuaires successifs n'ayant pas d'intérêt pour l'application visée, ce travail n'aborde pas la question du liage des entrées d'annuaires successifs.

Dans cet article, nous proposons d'adapter et d'étendre cette approche pour construire et peupler un graphe de connaissances géohistorique permettant de suivre l'évolution d'un commerce au cours du temps. L'objectif est de se doter de données structurées, permettant le suivi individuel et collectif des commerces d'un type donné au cours du temps et dans l'espace parisien ancien. Nous la mettons en œuvre sur les entrées relatives aux métiers de la photographie, mais elle peut être appliquée à n'importe quelle activité professionnelle représentée dans ces annuaires. Le choix de cette profession particulière a été fait pour deux raisons : d'une part, nous disposions d'un premier recensement effectué manuellement par des historiens de l'art et qui nous a aidé à constituer notre jeu de données initial (cf. section 4.5) ; d'autre part, il nous semblait intéressant d'étudier une profession connaissant de fortes évolutions. La photographie apparaissant et se développant rapidement au cours du XIX^e siècle, le suivi de sa représentation dans les annuaires nous a semblé constituer un cas d'usage pertinent.

Cet article étend notre publication [52] dans les actes de la conférence d'Ingénierie des Connaissances 2023. Par rapport à l'article initial, cette version propose : (1) un état de l'art plus détaillé ; (2) une nouvelle chaîne d'extraction des données dans les

annuaires du commerce parisien du XIXe siècle, couvrant plus d'annuaires, avec une approche plus performante et des données globalement de meilleure qualité (moins d'erreurs d'OCR ou de segmentation des entrées, etc.) ; (3) une présentation plus détaillée du corpus ; (4) une évaluation du géocodage historique ; (5) une approche de liage plus stricte, assurant des liens a priori corrects ; (6) un tutoriel permettant de reproduire l'approche pour d'autres professions. L'article est organisé de la façon suivante : la section 2 présente les travaux antérieurs sur la création de graphes géohistoriques ; la section 3 décrit les questions de compétences associées à notre graphe de connaissances ; la section 4 détaille les étapes de création du graphe ; la section 5 propose une application de visualisation spatio-temporelle du graphe et l'évaluation des questions de compétence ; la section 6 discute des perspectives de ce travail.

2. Travaux antérieurs

La création d'un graphe de connaissances à partir d'annuaires anciens nécessite d'en extraire et structurer les informations textuelles. Cette section passe en revue les travaux relatifs à ces différentes étapes : extraction du texte, reconnaissance et liage des entités nommées.

2.1. Détection de mise en page et transcription

Les approches d'analyse de mise en page visent à identifier et étiqueter les régions homogènes de documents. Dans son état de l'art, [8] distingue trois stratégies. Les stratégies descendantes ou *top-down*, comme XY-cut [38] et ses dérivées [20, 49], appliquent des règles pour diviser progressivement le document en portions de plus en plus petites, jusqu'à atteindre un critère d'arrêt prédéfini ou bien qu'il ne soit plus possible de créer de portion plus petite. Les stratégies ascendantes ou *bottom-up*, comme la méthode docstrum [42] partent des portions élémentaires de documents (des pixels, des mots, etc.) et les regroupent pour créer des régions homogènes, jusqu'à atteindre un critère d'arrêt prédéfini. Certaines approches par apprentissage automatique modernes reprennent ce principe. On peut distinguer celles qui se basent sur des réseaux de segmentation sémantique, comme dhSegment [5] ou Doc-UFCN [9], et celles qui basent sur des réseaux de détection, comme celles disponibles dans la bibliothèque Layout-Parser [45]. Enfin, les approches hybrides combinent des techniques ascendantes et descendantes. Un exemple intéressant de famille d'approches hybrides moderne est celle des approches multimodales comme LayoutLM [55] et ses variantes comme StructuralLM [30]. Grâce à des réseaux de type *transformer*, ces approches tirent profit d'une étape de sur-segmentation préalable (ascendante ou descendante) pour ensuite venir catégoriser et regrouper les éléments constitutifs d'une page. Leur force résidant dans leur capacité à exploiter conjointement les informations textuelles, visuelles et spatiales des documents pour reconnaître des mises en pages très complexes et variées. Cependant, la puissance de ces approches nécessite, au-delà de ressources en calcul plus importantes, d'être spécialisées pour les types de documents concernés, en l'absence de modèles généralistes disponibles à ce jour. Dans le cadre de documents à la mise en page simple, numérisé dans une qualité suffisante, les approches anciennes ne sauraient donc être ignorées en raison simplement de leur âge, si elles permettent

d'atteindre un niveau de qualité acceptable pour une fraction du temps et du coût de traitement nécessaires avec des approches plus récentes.

Concernant la transcription automatique de documents modernes imprimés en français, les outils libres sont relativement abondants et performants. Les systèmes de reconnaissance optique de caractères (OCR) récents, comme Tesseract [48], OCRopus [12], PyLaia [43], Kraken [24], Calamari [54], ou Pero OCR [28] s'appuient sur des architectures à base de réseaux de neurones convolutifs (CNN) et de réseaux *Long short-term memory* (LSTM). Ils obtiennent globalement de bons résultats sur des textes récents, mais sur les textes anciens, pour lesquels moins de données d'entraînement sont disponibles, leurs performances baissent. Pour pallier cette difficulté, Pero OCR intègre une couche pour détecter le style de transcription le plus adapté au texte à traiter [28]. Plus récemment, des percées intéressantes ont été effectuées par des systèmes basés sur des architectures à base de *transformers*, comme ABINet [18], DONUT [25] ou encore TrOCR [32]. Ces approches permettent l'utilisation de modèles de langue sophistiqués, mais doivent encore être adaptées aux types de documents visés, nécessitant un travail supplémentaire qui n'est pas forcément justifié lorsque la performance des approches plus simples (et souvent moins gourmandes en ressources de calcul) suffisent à produire des données de qualité suffisante pour les étapes ultérieures de la chaîne de traitement.

2.2. Extraction d'informations et reconnaissance d'entités nommées

De nombreuses approches ont été proposées pour localiser et classer les portions de texte qui désignent des entités de types prédéfinis comme des personnes, des lieux ou des organisations [37].

Les approches à base de règles utilisent des patrons lexico-syntaxiques combinant catégories grammaticales [6, 41] et entrées de dictionnaires [33, 35]. Sur des corpus spécialisés exempts d'erreurs de transcription, lorsque l'on dispose de dictionnaires exhaustifs, elles produisent de bons résultats, mais l'élaboration des patrons exige une expertise et des efforts importants[37].

Les approches supervisées regroupent les techniques d'apprentissage statistique traditionnel et les techniques à base de réseaux de neurones profonds. Comme les approches par patrons, les premières exploitent des descripteurs textuels choisis par un expert [36]. Les secondes, en revanche, définissent leurs propres descripteurs pour classer les tokens selon leur appartenance à un type d'entités nommées. L'état de l'art proposé par [31] montre que les modèles de langue récents, en particulier ceux de type *transformer*, peuvent être adaptés à des corpus spécialisés avec relativement peu de données d'entraînement et qu'ils sont très susceptibles de produire de meilleurs résultats que les approches par règles ou par apprentissage statistique. [11] entraîne un modèle de langue à base de réseaux de neurones profonds dédié au français et issu de la bibliothèque spaCy pour reconnaître des entités nommées dans des entrées d'annuaires de propriétaires parisiens de la fin du XIX^{e} et de début du XX^{e} siècles. Si les résultats obtenus sont satisfaisants, ils peuvent en revanche souffrir des erreurs de transcriptions liées à l'étape d'OCR. Une approche visant à adapter un modèle de reconnaissance d'entités nommées aux erreurs de transcription a donc été proposée

par [2]. L'approche a été testée sur les modèles français de la bibliothèque spaCy et le modèle de reconnaissance d'entités nommées *CamemBERT* publié sur *Hugging Face*[(1)] ; le second s'est révélé meilleur sur les tests effectués. Nous avons donc mis en œuvre cette dernière approche pour la chaîne de traitement présentée dans cet article.

2.3. Construction de graphes géohistoriques et liage de ressources

De nombreux modèles ont été proposés pour représenter des données spatio-temporelles [46]. Les travaux récents sur la représentation des états passés successifs du territoire reposent majoritairement sur des modèles de graphes. Ainsi [7, 15, 29] s'inspirent du modèle de graphe spatio-temporel de [19] ; dans le premier cas, il s'agit de rues de Paris vectorisées à partir de plans à grande échelle levés à différentes périodes du XIX^e siècle, dans le second, des parcelles agricoles issues de plusieurs millésimes du Registre Parcellaire Graphique[(2)], et dans le troisième, d'unités territoriales statistiques produites par Eurostat et d'unités administratives suisses produites par Swisstopo. Ce dernier travail utilise les standards du Web de données pour représenter et publier les graphes créés. C'est également le cas de [10, 23, 53] qui proposent des vocabulaires pour représenter respectivement les évolutions de communes, de bâtiments ou encore de paroisses. La plupart de ces approches de construction de graphes géohistoriques utilisent des séries temporelles de données géographiques dont elles extraient les relations spatio-temporelles à l'aide de méthodes de liage entre états successifs des entités géographiques considérées.

Les approches de liage de données visent à créer des liens de correspondance explicites entre ressources représentant une même entité du monde réel, éventuellement à des temporalités différentes. [44] distingue deux principales catégories de méthodes de liage. Les méthodes fondées sur les données reposent sur l'hypothèse selon laquelle deux ressources présentant des valeurs similaires pour leurs propriétés similaires sont très susceptibles de représenter une même entité du monde réel. C'est le type d'approche mis en œuvre par des outils comme Silk[(3)] [22] ou LIMES[(4)] [39]. Les méthodes fondées sur les connaissances exploitent les connaissances fournies par l'ontologie qui décrit les données. Les restrictions désignant des ensembles de propriétés comme clés d'identification de ressources sont particulièrement utilisées par ces approches. De nombreux travaux sont ainsi dédiés à l'identification des clés pour le liage, comme [50] ou [3].

Les approches de création de graphes spatio-temporels proposées par [7, 15, 29] appartiennent à la première catégorie. Elles reposent essentiellement sur l'évaluation de la similarité de la forme et de la localisation des entités géographiques à lier, représentées dans les jeux de données traités sous la forme de géométries vectorielles. Les données des annuaires du commerce ne comportant pas de géométries, des approches de liage fondées sur les données ou sur les connaissances peuvent être envisagées.

---

(1) `https://huggingface.co/`

(2) Voir : `https://geoservices.ign.fr/rpg`

(3) `http://silkframework.org/`

(4) `http://aksw.org/Projects/LIMES.html`

3. Questions de compétence

Ce travail vise à adapter et étendre la chaîne de traitement proposée par [6] pour construire un graphe de connaissances géohistorique à partir d'annuaires anciens. L'objectif de ce modèle de connaissances est d'aider les historiens à suivre et analyser les évolutions des commerces sur le territoire considéré. Ces évolutions peuvent porter sur la nature même des commerces, sur leurs localisations, sur leur pérennité, sur leurs modes d'organisation, etc. Nous avons donc retenu les questions de compétences suivantes, définies avec les historiens du projet. Il s'agit des questions auxquelles on souhaite a minima pouvoir répondre, et que nous supposons suffisamment générales pour pouvoir s'appliquer à la plupart des types de commerces figurant dans les annuaires.

CQ1. Quelle est l'adresse du commerce X d'après le(s) annuaire(s) publié(s) en YYYY ?
CQ2. Combien y a-t-il de commerces de ce type localisés rue X d'après le(s) annuaire(s) publié(s) en YYYY ?
CQ3. Quels sont les commerces situés dans une zone définie par un polygone ou un rectangle englobant d'après le(s) annuaire(s) publié(s) en YYYY ?
CQ4. Quels commerces ont déménagé au cours de leur existence ?
CQ5. Quels commerces ont été repris par un autre commerçant exerçant la même activité ?

Par ailleurs, les logiques d'organisation spatio-temporelles des commerces peuvent être difficiles à mettre en évidence à l'aide de simples requêtes et nécessitent souvent des analyses spatio-temporelles plus complexes. Par exemple, identifier la multiplication de commerces du même type tenus par les membres d'une même famille dans un même quartier exige d'expliciter à la fois les liens familiaux entre les propriétaires de commerces, la proximité spatiale des commerces sur une période donnée et d'éventuelles logiques de transmissions intra-familiales. Pour faciliter ce type d'analyses complexes, nous proposons donc d'accompagner notre graphe de connaissances géohistorique d'une application de visualisation spatio-temporelle des données.

4. Construction du graphe de connaissances géohistorique

Les informations contenues dans les annuaires peuvent être vues comme des séries temporelles de données semi-structurées sur les commerces qu'elles décrivent. Nous proposons donc une approche d'extraction d'informations et de construction de graphe de connaissances qui reprend et adapte les étapes de la chaîne de traitement de [6] et les approches à base de liage de [15], [29] et [7]. Partant d'une définition d'un corpus de sources primaires, nous procédons tout d'abord à une analyse page par page des parties pertinentes des annuaires pour en extraire le contenu textuel. Grâce à cette matière première augmentée d'informations de mise en page (espaces et sauts), il est ensuite possible de venir simultanément séparer les *entrées* qui constituent ces annuaires, ainsi que leurs attributs : les noms de propriétaires, les activités et les adresses pour la plupart des cas. Ces adresses sont ensuite géocodées à l'aide d'un

outil et d'une base de connaissances géo-historique en tenant compte de la période considérée. Pour des raisons de performance, un filtrage des entrées pertinentes est réalisé, selon la portée de l'étude qu'on souhaite mener (ici le suivi des photographes). Finalement, nous procédons au liage des entrées sélectionnées en comparant les valeurs de leurs attributs ; ceci nous permet de créer les relations au sein de notre graphe de connaissance.

4.1. Les annuaires du commerce parisien

Le corpus utilisé rassemble 144 annuaires publiés annuellement entre 1798 et 1914 par différents éditeurs et couvrant 85 années. Leurs contenus varient donc d'une édition à l'autre, en termes d'informations disponibles, d'organisation (index par noms, rues ou professions), de mise en page, de police d'écriture, etc. Ainsi, la Figure 4.1 présente quelques exemples d'annuaires, représentatifs de cette diversité d'organisation et de mise en page selon les éditions. Ils sont conservés dans différentes bibliothèques parisiennes et ont été scannés indépendamment les uns des autres, avec des niveaux de qualité variables. Les entrées des index par noms comportent généralement le nom du commerce ou de son propriétaire, le type d'activité exercée, d'éventuels titres honorifiques ou médailles professionnelles, le nom de la rue et le numéro et éventuellement une précision sur le type du local, comme « atelier », « entrepôt » ou « boutique », lorsque plusieurs adresses sont fournies. L'entrée de l'annuaire Didot-Bottin de 1860 « *Aubert (Mme), couturière, Guénégaud, 10* » est un exemple typique d'entrée des index par noms.

FIGURE 4.1. Exemples de mises en pages et d'index différents dans les annuaires. *En haut* : Duverneuil et La Tynna 1806 – index par noms ; *Au milieu* : Deflandre 1828 – index par professions ; *En bas* : Bottin 1851 – index par rues.

4.2. Détection de mise en page et transcription

Les annuaires à traiter présentent différentes mises en pages selon les éditions et selon les index. Cependant, celles-ci restent relativement homogènes : les entrées sont toujours organisées en colonnes (de 1 à 5 selon les éditions) et éventuellement séparées par des titres. Cette connaissance a priori forte sur les données à traiter a permis la mise en œuvre d'un étage d'extraction du contenu textuel et de sa mise en page selon les 4 étapes suivantes.

(1) Détecter la mise en page macroscopique jusqu'à repérer les différentes colonnes de texte, à l'aide d'une méthode basée sur des découpages horizontaux et verticaux successifs (XY-cuts [38]) et de classification de régions (titres vs texte). Sur notre corpus composé de documents exclusivement textuels, ces techniques s'avèrent extrêmement performantes et peu coûteuses à mettre en œuvre.
(2) Détecter et reconnaître les lignes de texte au sein de chaque colonne de texte, à l'aide du système PERO OCR « sur étagère », c'est-à-dire sans adaptation à nos données. Ce système exploite pour la détection de lignes et de blocs de texte une architecture purement convolutive, *ParseNet* [27], qui prédit pour chaque pixel sa probabilité d'appartenir à la ligne de base ou aux extrémités de la ligne de texte, ainsi que les valeurs probables de hauteur de ligne, de hauteur des caractères descendants et des caractères ascendants.
(3) Le module de reconnaissance du système est quant à lui composé d'un étage convolutif (architecture VGG [47]) et d'un étage de décodage (couches LSTM [21]). Les principales innovations de ce système résident dans son protocole d'entraînement qui, d'une part, s'appuie sur une technique de *pseudo-labeling* et d'augmentation de données pour faire face à une faible quantité de données annotées [26], et d'autre part, sur un module d'adaptation dynamique au style de transcription afin de faire face aux éventuelles incohérences entre les conventions d'annotation [28].
(4) Produire une sortie structurée et ordonnée en pages, puis colonnes, et enfin lignes. Grâce aux dimensions de colonnes et des lignes qui sont préservées, les marges gauches et droites de chaque ligne sont connues. Nous produisons donc, à l'issue de cette étape de détection de la structure et de transcription, une liste de lignes de texte, dans leur ordre probable de lecture, auxquelles sont associées les informations suivantes : leur transcription, leurs coordonnées, leurs marges gauches et droites à l'intérieur de leur colonne parent et relativement à la taille de cette colonne, l'identifiant de leur colonne parent et l'identifiant de leur page parent.

Cette approche de détection de la structure de la page a succédé à une approche plus rudimentaire qui reposait sur un découpage de la page en entrées selon 3 passes : tout d'abord la page était progressivement découpée avec une variante *XY-cuts* jusqu'à obtenir des blocs de texte homogènes, puis les lignes de texte étaient séparées avec une méthode à base de *watershed*, et finalement les lignes étaient regroupées en *entrées* en fonction de leurs espaces en début et fin de ligne. Malgré sa vitesse élevée, cette

première approche pénalisait le système OCR entraîné pour un détecteur de ligne spécifique. De plus, même cette méthode disposait d'une performance raisonnable grâce à la possibilité de calibrer ses paramètres, elle ne tirait par profit de l'information textuelle riche des entrées (qui disposent d'une syntaxe très régulière), et elle ne permettait pas de gérer simplement les entrées à cheval sur plusieurs colonnes ou pages. Nous avons donc abandonné cette première approche purement visuelle de séparation des entrées, pour déployer une approche intégrant les informations textuelles et visuelles (sauts de ligne, de colonne et de page, ainsi que les espaces de début et fin de ligne) dans le processus d'extraction des entités nommées, afin de procéder simultanément à la séparation des entrées et à l'extraction de leurs attributs à l'aide d'un réseau unique. Cette approche sera décrite dans la section 4.3.

La détection des colonnes de texte mentionnée en étape 1 ne nécessite finalement aucun apprentissage, mais elle possède un certain nombre de paramètres calculés à partir de la taille des caractères. Bien qu'elle n'ait pas fait l'objet d'une comparaison rigoureuse avec d'autres approches, principalement à cause du faible nombre d'approches « traditionnelles » *open-source*, nous avons pu l'utiliser avec succès pour traiter les dizaines de milliers de pages du corpus avec un taux d'échec assez faible. De plus, grâce à une analyse de la régularité de la mise en page détectée en post traitement, nous avons pu automatiquement identifier certaines pages problématiques (détection de 2 colonnes dans une grande plage de pages à 3 colonnes par exemple), et forcer manuellement la valeur de la taille des caractères grâce au contexte de collection lorsque c'était nécessaire. Cette pratique a permis de réduire le nombre de cas d'échec significativement. Ce premier module est intégré au code source original du module de traitement de données[(5)].

Le système PERO OCR mentionné aux étapes 2 et 3 a fait l'objet d'une réutilisation sans adaptation (*finetuning*) à nos données. Nous avons utilisé les sources de la version `0.6.1` disponible à l'été 2023, ainsi que les modèles entraînés par les créateurs de PERO OCR en novembre 2022 pour la reconnaissance de textes imprimés modernes. Ce système a été choisi pour sa bonne performance lors d'expériences préliminaires [2] : sur un jeu de données de 8 765 entrées manuellement corrigées [1], PERO OCR obtenait les meilleurs résultats avec un taux d'erreur de transcription niveau caractère de 3,78 % sans aucune normalisation. Finalement, l'enrichissement mentionné à l'étape 4 a pour charge de modifier les données textuelles produites pour y intégrer les informations de sauts (de ligne, colonne et page) ainsi que les dimensions des espaces en début et fin de ligne. Nous stockons les valeurs des distances entre le bord gauche (resp. droit) de la colonne à laquelle appartient une ligne et le début (resp. la fin) de cette ligne. Ces valeurs sont normalisées par la largeur de la colonne. Cet outil très simple a fait l'objet d'une distribution séparée pour faciliter le travail par lot[(6)].

[(5)] `https://github.com/soduco/directory-annotator-back/blob/main/doc/cli.md`
[(6)] `https://github.com/soduco/processor-ocr-pero`

4.3. Extraction d'informations et reconnaissance d'entités nommées

Si les éléments constitutifs des entrées d'annuaires restent globalement les mêmes, leur présentation, en revanche, varie d'une édition à l'autre. En effet, la structure des entrées peut présenter des variations dans l'ordre des éléments à extraire (« raison sociale », adresse et activité, essentiellement) et dans le niveau de détail qu'elles contiennent (principalement en ce qui concerne la description d'activité). À cette difficulté s'ajoute le bruit des inévitables erreurs de transcriptions, invalidant de fait toute approche d'extraction d'information à base de règles qui nécessiterait la spécification manuelle d'un très grand nombre de cas imprévisibles.

Nous avons donc adopté une technique moderne d'extraction d'informations basée sur le principe de la reconnaissance d'entités nommées, qui consiste concrètement à catégoriser les fragments de texte fournis à la méthode. Il s'agit donc d'une tâche de classification que nous avons implémentée à l'aide du modèle de langue CamemBERT [34] basé sur une architecture de type *transformer encoder*. Ce type d'architecture présente l'avantage d'être très rapide à spécialiser avec peu de données, tout en présentant une excellente robustesse au bruit OCR. En effet, nous avons montré [2] qu'il était possible d'atteindre un F-score en détection stricte des entités attendues (c'est-à-dire en détectant exactement leur position) de plus de 92 % avec seulement 50 exemples d'entrées annotées, dès lors qu'un pré-entraînement sur des données bruités réalistes était réalisé au préalable.

La variante de base de cette approche s'entraîne de la façon suivante. Le modèle de base CamemBERT [34] que nous utilisons a déjà fait l'objet d'un double entraînement : il a été pré-entraîné de façon faiblement supervisée avec une tâche de reconstruction (*masked language modeling*) sur le jeu de données massif OSCAR[(7)], puis entraîné de façon plus spécifique pour la reconnaissance d'entités nommées sur le jeu de données WikiNER-FR [40]. Cependant, le modèle de langue capturé ainsi que les entités disponibles ne sont pas exactement alignés sur les propriétés du corpus des annuaires que nous avons traité. Ensuite, nous avons à notre tour procédé à nouveau à 2 passes d'entraînement : tout d'abord avec la même tâche de reconstruction faiblement supervisée sur plusieurs milliers de pages d'annuaires (plusieurs milions de lignes de texte automatiquement reconnues par le système présenté en section 4.2), puis de façon fortement supervisée sur un corpus annoté (de 6 373 entrées) avec les types d'entités que l'on cherche à reconnaître : PER pour les noms de commerces ou de personnes, ACT pour le type d'activité, LOC pour les noms de rues, CARDINAL pour les numéros, TITRE pour les distinctions et FT pour les précisions sur les adresses. Pour limiter les effets négatifs des erreurs d'OCR sur les résultats de reconnaissance des entités nommées, nous avons entraîné le modèle sur du texte bruité lors des 2 passes. Sur notre corpus de test, de 1 669 entrées également bruitées, le modèle obtient ainsi un F-score global de 94,1 %. Les étapes d'extraction du texte et de reconnaissance des entités nommées et les résultats obtenus sont décrits en détail dans [2].

[(7)] `https://oscar-project.org/`

Cette variante a été ensuite étendue pour permettre de procéder simultanément à la reconnaissance des entités nommées (tâche d'extraction d'informations) et à la séparation des entrées (tâche de segmentation). L'avantage majeur de cette approche est de permettre de réaliser une opération supplémentaire de la chaîne de traitement, la séparation des entrées, sans surcoût de calcul, car le même réseau est utilisé. Nous avons comparé plusieurs stratégies de labellisation des tokens, activant sélectivement la possibilité de détecter les entités nommées et séparer les entrées [16]. La meilleure configuration repose sur l'injection de deux types de tokens supplémentaires : un token de rupture `<break>` utilisé pour les sauts de ligne, de colonne et de page ; et un groupe de tokens d'espaces, permettant de quantifier la taille des zones libres à gauche et à droite de chaque ligne. Pour chaque côté, 3 tokens ont été utilisés ; correspondant à des valeurs des plages de valeurs différentes. Par exemple, un des tokens possibles sert à encoder les espaces droits plus petits que 1 % de la largeur de la colonne, tandis qu'un autre peut encoder les espaces gauches qui sont plus que 2,5 % plus grands que ceux de la ligne précédente. Ces nouveaux tokens permettent de capturer des éléments essentiels de la mise en page des annuaires, tout en servant de support potentiel à des marqueurs de segmentation ; le modèle encodeur CamemBERT prédisant en effet, pour chaque token de la séquence, un label. Grâce à cette approche, nous avons pu mesurer en conditions expérimentales un F-score de 99,2 % pour la prédiction des séparateurs d'entrées, tout en maintenant un F-score moyen de 94,0 % pour la détection des entités nommées. Cette variante a été utilisée pour calculer les résultats présentés dans cet article, et son code est disponible librement[8]. Toutefois, il faut noter que la qualité des lignes extraites par l'OCR est légèrement inférieure à celle des lignes validées à la main pour les expériences, et qu'il conviendrait de ré-entraîner le système sur des données réelles pour améliorer sa performance.

Pour terminer, mentionnons une dernière variante prometteuse qui n'a pas été intégrée à la chaîne de production finale. Elle permet de prédire la structure hiérarchique des entrées directement, principalement pour faire face aux cas à multiples adresses et pour faciliter la mise en correspondance des différentes composantes de chaque adresse. Nous nous sommes intéressés à la possibilité de mettre en place une reconnaissance d'entités nommées imbriquées [51], et avons montré qu'il était possible de prédire des données structurées sans perte de performance dans le cas des documents étudiés.

4.4. Géocodage historique des entrées

La mise en œuvre des deux étapes précédente a permis de construire une base de données de 22 743 928 entrées, chacune composée d'au moins un nom de commerce ou de personne, un type d'activité et une adresse. Les adresses sont des énoncés courts contenant un nom de voie et, le plus souvent, un numéro d'immeuble. Placer ces entrées sur une carte est un préalable à l'analyse spatiale de données dont la répartition dans la ville et les trajectoires individuelles qu'elles dessinent témoignent de l'activité urbaine de Paris sur plus de 100 ans. La cartographie de entrées d'annuaires passe par leur géocodage de leurs adresses, nécessairement automatisé étant donné la masse de

[8] `https://github.com/soduco/processor-ner`

données considédérée. Géocoder, c'est à dire assigner des coordonnées géographique à chaque adresse, est une opération triviale lorsqu'il s'agit d'adresses contemporaines, automatisable facilement en s'appuyant sur des bases de données géographiques à jour comme OpenStreetMap, la Base Adresse Nationale ou encore le Géoportail de l'IGN. Ces bases d'adresses sont cependant inadaptées au géocodage des entrées d'annuaire car les adresses qu'elles contiennent ont souvent disparu ou ne désignent plus le même immeuble.

Pour résoudre ce problème, nous avons mobilisé un outil de géocodage historique développé par le groupe de travail GeoHistoricaldaData [13]. Cet outil spécialise le géocodeur open-source Pelias[(9)] pour le cas des adresses anciennes. Il est appuyé sur une base de données d'adresses anciennes contituée en agrégeant des jeux de points d'adresses issus de quatres plans à grande échelle de Paris au XIX^e siècle, préalablement géoréférencés et traités à l'occasion de travaux pré-existants en humanités-numériques [(10)]. Chaque point d'adresse est ensuite associé à un intervalle temporel correspondant à la période de production du plan dont il a été extrait. Le géocodeur s'appuie sur le moteur de recherche ElasticSearch et expose une API REST[(11)] pouvant être interrogée, sous la forme d'une requête associant un énoncé d'adresse ainsi qu'une date cible. Lors de la recherche, l'outil favorise alors les adresses issues de plans dont la date de production est proche de celle des données à géocoder [(12)].

L'ensemble des entrées ont été ainsi géocodées, la date cible était celle de publication de l'annuaire du commerce dont est issue l'entrée. Afin de réduire le volume d'adresses à géocoder, les adresses identiques issues d'annuaires différents ont été réduites à une unique requête de géocodage si l'écart temporel entre les annuaires était de 10 ans ou moins.

Sur quelques 6 900 000 adresses distinctes, environ 96 % ont été effectivement géocodées. Toutefois, une évaluation qualitative menée sur l'annuaire Didot de 1845 et en cours de publication tends à montrer que la qualité du géocodage varie fortement selon si l'adresse se trouve à l'intérieur ou à l'extérieur des frontières administratives de Paris. Si quelques adresses sont placées hors Paris (282, soit 0,0005 %), du fait de la faible couverture de ces espaces par la base de données du géocodeur, seules 40 % d'entre elles sont correctement géocodées. Dans Paris, 67 % des adresses sont bien localisées, mais cela correspond à 88 % de toutes les entrées de l'annuaire [(13)].

[(9)] `https://github.com/pelias`

[(10)] Notamment par le projet ANR ALPAGE et l'initiative GeoHistoricalData

[(11)] Accessible à cet URI `https://api.geohistoricaldata.org/docs/#/Geocoding`

[(12)] Ce jeu de données géocodées est publié sur le dépôt suivant : `https://nakala.fr/10.34847/nkl.98eem49t`

[(13)] Cela tends à montrer que les adresses les mieux localisées sont également les plus fréquentes au sein des annuaires. De plus, la présence de bruit OCR et les erreurs de reconnaissance des adresses dans les entrées aboutit à des adresses erronées qui sont pourtant comptabilisées dans ces mesures.

4.5. Sélection et représentation des entrées en RDF

Pour simplifier la suite du traitement, nous proposons de filtrer les entrées pour ne conserver que celles concernant le type de commerces à étudier. Ceci facilite à la fois le passage à l'échelle de l'étape de liage en réduisant fortement le nombre d'entrées à traiter et limite les erreurs de liage pouvant intervenir en cas d'entrées homonymes dans les annuaires. De plus, nous nous restreignons aux entrées des seuls index par noms, afin de nous abstraire des biais de classification des activités introduits par les éditeurs d'annuaires.

Pour sélectionner les entrées concernant les métiers de la photographie, nous nous sommes appuyés sur une liste de 252 photographes parisiens extraite de l'ouvrage de M. Durand [17], qui couvre la période 1820-1910. Nous avons recherché les entrées associées à ces photographes et recencé les activités mentionnées dans ces entrées pour en retenir trois mots-clés que l'on suppose représentatifs des entrées décrivant des photographes : *photo*, *daguer* et *opti*.

Le premier permet de récupérer les photographes, le second les daguerréotypistes et le troisième les opticiens, producteurs de lentilles dont beaucoup se sont spécialisés dans la fabrication voire l'exploitation directe de matériel de daguerréotypie et de photographie. Cette dernière profession est antérieure aux deux autres et figure dans les annuaires dès le début de la période étudiée, comme en témoigne la figure 4.2. Le daguerréotype est l'un des premiers procédés photographiques, créé par Nicéphore Niépce et Louis Daguerre, à qui il doit son nom. Mis au point en 1835, ce procédé est publié ouvertement en détails en 1839, ce qui permet l'essor de son exploitation commerciale. La toute première entrée d'annuaire en faisant mention, dans notre sélection, apparaît dans l'annuaire Didot de 1841 : *« Gérard et Cie, articles pour daguerréotypes, 46 q. des Orfèvres »*. Cette technique connaît un succès rapide jusqu'au milieu des années 1850, comme le montre l'augmentation du nombre d'entrées d'annuaires sur la figure 4.3. Concurrencée par d'autres procédés photographiques, commercialement plus intéressants – plus rapides, moins chers, permettant des reproductions, etc. – comme l'ambrotype, le ferrotype, le panotype ou la photographie à l'albumine, la daguerréotypie disparaît progressivement des annuaires dans les années qui suivent. Les dernières entrées recensées vers la fin du siècle concernent des fabricants de passepartout pour daguerréotype. Le terme *photographe* supplante ainsi rapidement celui de *daguerréotype* dans les annuaires. Eugène Delannoy[(14)] et François Louis Victor Trottier sont ainsi les premiers de notre sélection à apparaître sous cette apellation, dans l'annuaire Didot de 1848 [(15)]. Les échelles verticales des histogrammes présentés en figures 4.2, 4.3 et 4.4 sont différentes pour des raisons de lisibilité, mais les effectifs des photographes sont équivalents à ceux des daguerréotypistes dès 1855 et vont croître considérablement à la fin du XIXe siècle.

[(14)] NB : Eugène Delannoy apparaît dans le Didot de 1847 avec la mention « daguerréotypes ».

[(15)] Les occurences antérieures du mot-clé « photo » rencontrées dans notre sélection ne correspondent pas au terme « photographe ».

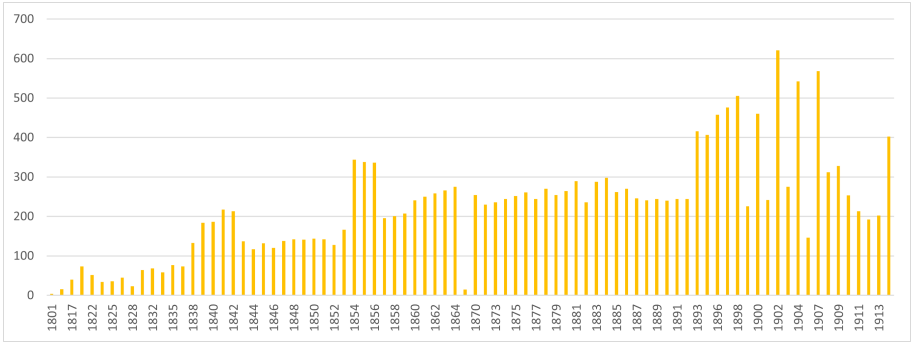

FIGURE 4.2. Répartition temporelle des entrées d'annuaires comportant le mot-clé *opti*.

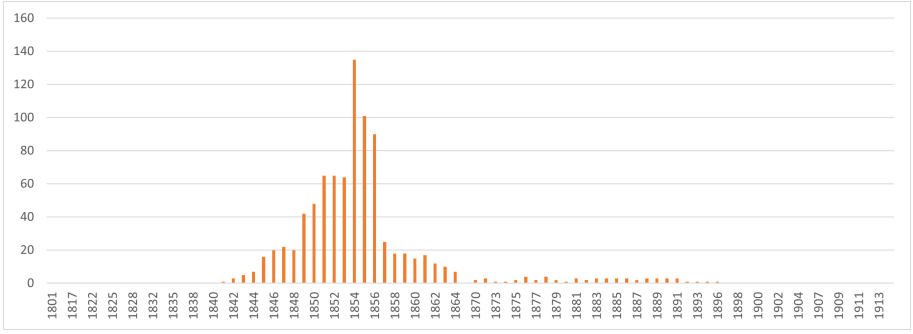

FIGURE 4.3. Répartition temporelle des entrées d'annuaires comportant le mot-clé *daguer*.

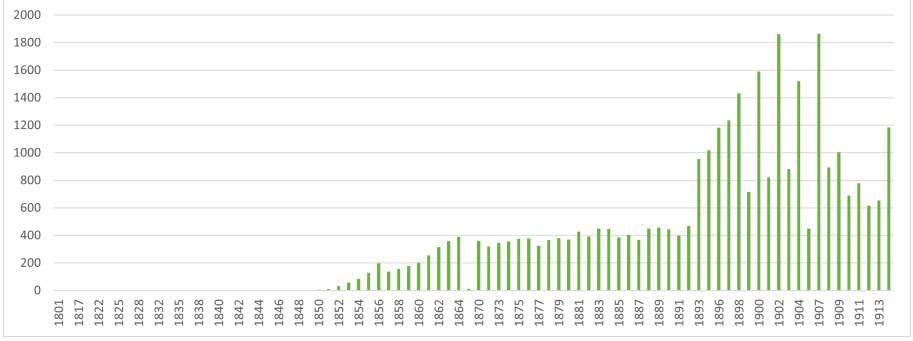

FIGURE 4.4. Répartition temporelle des entrées d'annuaires comportant le mot-clé *photo*.

Nous avons finalement extrait 52 231 entrées dont l'attribut « activité » comporte ces mots-clés. Elles ont été converties à l'aide d'un script R2RML pour former un premier graphe RDF.

4.6. Liage des entrées

Pour lier les entrées représentant un même commerce dans différents annuaires, nous avons procédé à une recherche des propriétés des entrées pouvant faire office de clés au sein de chaque annuaire. Les propriétés qui composent les clés sont identifiées à l'aide de Sakey [50]. La propriété adresse est créée par concaténation des valeurs des entités de type LOC et CARDINAL, préalablement à l'exécution de Sakey. Tous les caractères ont été passés en minuscules et les éléments de ponctuation situés en début et fin de chaînes ont été supprimés.

Ainsi, les clés identifiées pour le liage sont : (1) l'identifiant de l'entrée au sein de l'annuaire, (2) le nom et l'adresse, (3) l'activité et l'adresse, et (4) le nom, l'activité et l'année de parution de l'annuaire. Les clés (2) et (3) sont des n-quasi-clés identifiées sur les données d'un annuaire. La valeur de n varie de 1 à 3 selon les annuaires : on tolère donc quelques exceptions afin de gérer l'existence possible de doublons liés à des erreurs d'OCR lors des étapes précédentes. La clé (2) va ainsi nous permettre de lier des entrées représentant des commerces dont l'intitulé d'activité peut avoir varié au cours du temps. La clé (3) va permettre de mettre en évidence des situations de transmission de commerces d'un propriétaire à un autre. La clé (4) est une n-quasi-clé dont la valeur de n dépend du nombre d'entrées, potentiellement élevé, possédant plusieurs adresses. Là encore, un certain nombre d'exceptions, plus élevé, est toléré afin d'autoriser cette situation qui ne pose pas de problèmes pour le liage. Cette clé reste en effet très utile car elle permet d'identifier des commerces ayant potentiellement déménagé. En revanche, dans la mesure où les valeurs d'activité sont assez peu variées, il semble préférable de s'assurer que les intervalles temporels entre entrées liées à l'aide de cette clé restent faibles. Dans le cas contraire, des liens entre homonymes exerçant la même activité à 100 ans d'intervalle pourraient être créés[(16)]. Le corpus d'annuaires utilisé offre une couverture temporelle de la période relativement homogène. En revanche, on constate une période peu ou pas couverte de 1864 à 1870. Nous avons donc choisi d'autoriser un écart temporel maximal de 7 ans entre deux entrées liées via cette clé.

Pour permettre l'application de cette dernière condition, nous avons mis en œuvre cette étape de liage avec Silk. Après suppression une étape de transformation des caractères en minuscule, les propriétés nom, activité et adresse de chaque paire de ressources sont comparées à l'aide de mesures de similarité de chaînes de caractères dont le seuil est fixé à zéro, afin de s'assurer de ne produire que des liens corrects. Les identifiants des entrées d'annuaires sont comparés à l'aide d'un test d'égalité des valeurs. Enfin, les valeurs d'année de publication des annuaires sont comparées à l'aide d'une mesure qui calcule l'écart entre deux valeurs numériques, dont le seuil est

[(16)] Ce cas d'homonymie pourrait également arriver pour des photographes exerçant à la même période, mais dans ce cas, les éditeurs d'annuaires prennent généralement garde à les distinguer en précisant leurs prénoms par exemple.

fixé à 7. Pour les combinaisons de propriétés suivantes – Nom, Activité et Année de publication, Nom et Adresse, Activité et Adresse – le score agrégé retenu correspond à la valeur de mesure de similarité la plus faible obtenue par l'une des propriétés de la combinaison. Finalement, seuls les liens dont le score est supérieur à zéro sont conservés.

232 749 liens *owl :sameAs* ont ainsi été calculés. Le nombre total des liens calculés et inférés est finalement de 395 239 liens distincts.

5. Visualisation et évaluation

Le graphe final est généré à l'aide d'un second script R2RML, qui permet d'obtenir des données plus structurées. Les liens créés à l'étape précédente y sont ajoutés.

Nous avons évalué le graphe de deux façons. D'une part, nous avons traduit les questions de compétences en requêtes SPARQL, afin de nous assurer que nous obtenions les réponses attendues. Le graphe est accessible ici : `https://dir.geohistoricaldata.org/`. D'autre part, nous avons développé une application de visualisation spatio-temporelle, qui permet d'analyser visuellement les données, sans avoir à écrire de requêtes.

5.1. Évaluation des questions de compétences

Ainsi, notre première question de compétence, qui recherche l'adresse d'un commerce précis à une année donnée, peut se vérifier avec la requête suivante :

```
PREFIX locn:<http://www.w3.org/ns/locn#>
PREFIX ont:<http://rdf.geohistoricaldata.org/def/directory#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX prov:<http://www.w3.org/ns/prov#>
PREFIX pav:<http://purl.org/pav/>
SELECT ?fullAdd
WHERE {GRAPH <http://rdf.geohistoricaldata.org/id/directories/photographes>
   {?e a ont:Entry.
    ?e rdfs:label ?label.
    ?e prov:wasDerivedFrom ?directory.
    ?directory pav:createdOn 1861.
    ?e locn:address ?add.
    ?add locn:fullAddress ?fullAdd.
    Filter regex(?label, "Dubois").}}
```

Listing 1. Requête pour retrouver l'adresse du commerce « Dubois » en 1861.

Cette requête renvoie l'adresse du photographe Dubois, situé au 25 boulevard de Sébastopol en 1861.

Les requêtes SPARQL correspondant aux autres questions de compétences listées sont disponibles, avec l'ensemble des scripts relatifs à ce travail, l'ontologie et l'application de visualisation sur le dépôt suivant : `https://github.com/soduco/atelier_graphes_geohistoriques_annuaires/`. Il fournit, en outre, un tutorial

qui permet de reproduire l'ensemble de la chaîne de traitement pour d'autres professions. Des tests préliminaires ont été réalisés dans ce sens, et des graphes sur les notaires, les magasins de nouveautés ou encore les graveurs de cartes sont accessibles sur le même point d'accès que le graphe des photographes, ainsi que sur l'application de visualisation.

5.2. Visualisation

Pour explorer les données du graphe de façon intuitive, nous avons développé une application de visualisation cartographique et temporelle[17]. Elle permet de filtrer les données par nom de commerce, par adresse, par activité et par intervales temporels et facilite l'identification d'éventuelles corrélations spatiales et temporelles et la réponse à certaines questions de compétence. Ainsi, en figure 5.1, la frise temporelle associée au photographe Nadar permet de constater que son atelier a déménagé, en 1860, de la rue Saint-Lazare au 113 boulevard des Capucines. La figure 5.2 montre que les ateliers des frères de la famille Chevalier sont concentrés quai de l'Horloge ; seul l'un des neveux, Charles, déménage en 1831 au Palais Royal. Enfin, la figure 5.3 montre qu'au moins 4 photographes se sont succédés au 59 rue de Rivoli au cours de la seconde moitié du XIXe siècle.

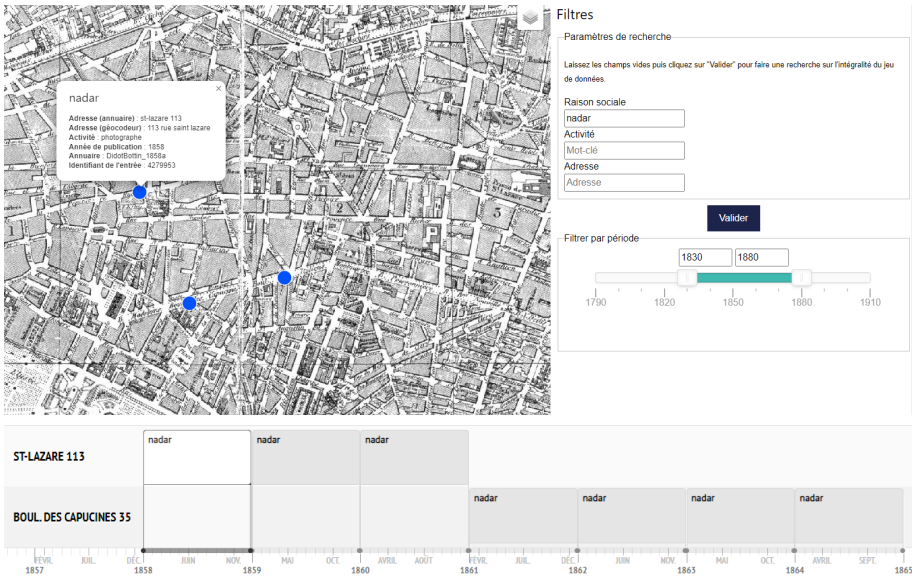

FIGURE 5.1. Déménagement du photographe Nadar

[17] `https://soduco.geohistoricaldata.org/atelier_graphes_geohistoriques_annuaires/`

FIGURE 5.2. Localisation des ateliers dont les propriétaires se nomment « Chevalier ».

FIGURE 5.3. Phénomène de transmission probable d'un atelier entre photographes.

6. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons proposé une approche pour créer et analyser un graphe de connaissances géohistoriques sur des commerces d'un type donné, à partir d'annuaires du commerce anciens. Nous l'avons appliquée ici au cas des métiers de la photographie qui émergent à Paris au cours du XIX^{e} siècle. La stratégie de liage adoptée permet de créer des liens entre entrées représentant un même commerce au cours du temps et de suivre l'évolution des entrées issues d'éditions successives. Cependant,

le fait de se cantonner à des valeurs identiques de propriétés lors du liage restreint fortement le nombre de liens créés. Une perspective à court terme sera donc de procéder à un liage moins strict. Mais cela nécessite de mettre en place une approche de vérification des liens créés efficace, car l'introduction de liens erronés dans le graphe pourrait engendrer des inférences indésirées. Le géocodage des entrées et leur visualisation cartographique permettent en outre d'identifier aisément des phénomènes spatiaux. Enfin, cette approche a été mise en oeuvre pour d'autres types de commerces, lors d'un atelier regroupant plusieurs historiens et historiens de l'art. Les graphes produits à cette occasion sont disponibles sur le point d'accès SPARQL et via l'application de visualisation. L'analyse critique de l'approche de création des graphes de connaissances géohistoriques professionnels et des analyses qu'ils permettent, du point de vue des historiens, feront l'objet d'une publication future.

7. Remerciements

Ce travail a été soutenu financièrement par l'Agence Nationale de la Recherche dans le cadre du projet SODUCO (ANR-18-CE38-0013) et par le Ministère des Armées – Agence de l'innovation de défense.

Bibliographie

[1] N. Abadie, S. Bacciochi, E. Carlinet, J. Chazalon, P. Cristofoli, B. Duménieu & J. Perret, « A Dataset of French Trade Directories from the 19th Century », in *A Dataset of French Trade Directories from the 19th Century (FTD) (1.0.0) [Data set]. Document Analysis Systems (15th IAPR International Workshop on) (DAS), La Rochelle, France*, Zenodo, 2022.

[2] N. Abadie, E. Carlinet, J. Chazalon & B. Duménieu, « A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories », in *Document Analysis Systems (DAS)* (Cham) (S. Uchida, E. Barney & V. Eglin, éds.), Document Analysis Systems. DAS 2022., vol. 13237, Springer, 2022.

[3] N. Abbas, J. David & A. Napoli, « Linkex: A Tool for Link Key Discovery Based on Pattern Structures », in *ICFCA 2019 - Workshop on Applications and tools of formal concept analysis* (Frankfurt, Germany), 2019, p. 33-38.

[4] T. N. Albers & K. Kappner, « Perks and pitfalls of city directories as a micro-geographic data source », *Explorations in Economic History* **87** (2023), article no. 101476.

[5] S. Ares Oliveira, B. Seguin & F. Kaplan, « dhSegment : A Generic Deep-Learning Approach for Document Segmentation », in *16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, p. 7-12.

[6] S. Bell, T. Marlow, K. Wombacher, A. Hitt, N. Parikh, A. Zsom & S. Frickel, « Automated data extraction from historical city directories: The rise and fall of mid-century gas stations in Providence, RI », *PLOS ONE* **15** (2020), n° 8, article no. e0220219.

[7] C. Bernard, M. Villanova-Oliver & J. Gensel, « Theseus: A framework for managing knowledge graphs about geographical divisions and their evolution », *Transactions in GIS* **26** (2022), n° 8, p. 3202-3224.

[8] G. M. Binmakhashen & S. A. Mahmoud, « Document layout analysis : a comprehensive survey », *ACM Computing Surveys (CSUR)* **52** (2019), n° 6, article no. 109 (36 pages).

[9] M. Boillet, C. Kermorvant & T. Paquet, « Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks », in *25th International Conference on Pattern Recognition (ICPR)*, 2021, p. 2134-2141.

[10] L. Bourel, N. J. Hernandez, N. Aussenac-Gilles & W. Charles, « HHT: une ontologie modulaire pour représenter l'évolution des territoires en Histoire », in *33e Journées Francophones d'Ingénierie des Connaissances (IC)*, AFIA, 2022, p. 131-136.

[11] C. Brando & F. Mélanie-Becquet, « Annuaires de propriétaires et des propriétés de Paris (1898, 1903, 1913, 1923) : du papier à la carte », in *2e Journée SoDUCo-BNF* (Paris, France), 2022.

[12] T. M. Breuel, « The OCRopus open source OCR system », in *Document Recognition and Retrieval XV*, vol. 6815, Int. Soc. for Optics and Photonics, 2008.

[13] R. Cura, B. Duménieu, N. Abadie, B. Costes, J. Perret & M. Gribaudi, « Historical collaborative geocoding », *ISPRS International Journal of Geo-Information* **7** (2018), no 7, p. 262.

[14] F. De Maupeou & L. Saint-Raymond, « Les "marchands de tableaux" dans le Bottin du commerce : une approche globale du marché de l'art à Paris entre 1815 et 1955 », *Artl@ s Bulletin* **2** (2013), no 2, article no. 7.

[15] B. Duménieu, « Un système d'information géographique pour le suivi d'objets historiques urbains à travers l'espace et le temps », Thèse, École des Hautes Etudes en Sciences Sociales, 2015.

[16] B. Duménieu, E. Carlinet, N. Abadie & J. Chazalon, « Entry Separation using a Mixed Visual and Textual Language Model: Application to 19th century French Trade Directories », `https://arxiv.org/abs/2302.08948`, 2023.

[17] M. Durand, *De l'image fixe à l'image animée : 1820-1910. Tome 2 : actes des notaires de Paris pour servir à l 'histoire des photographes et de la photographie*, vol. 2, Archives nationales, Pierrefitte-sur-Seine, 2015.

[18] S. Fang, H. Xie, Y. Wang, Z. Mao & Y. Zhang, « Read Like Humans : Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition », in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, p. 7098-7107.

[19] D. M. Géraldine, « Un modèle de graphe spatio-temporel pour représenter l'évolution d'entités géographiques », Thèse, Université de Brest, 2011.

[20] J. Ha, R. M. Haralick & I. T. Phillips, « Recursive XY-cut using bounding boxes of connected components », in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 2, IEEE, 1995, p. 952-955.

[21] S. Hochreiter & J. Schmidhuber, « Long short-term memory », *Neural computation* **9** (1997), no 8, p. 1735-1780.

[22] R. Isele, A. Jentzsch & C. Bizer, « Efficient Multidimensional Blocking for Link Discovery without losing Recall », in *Proceedings of the 14th International Workshop on the Web and Databases 2011, WebDB 2011, Athens, Greece, June 12, 2011*, 2011.

[23] T. Kauppinen, J. Väätäinen & E. Hyvönen, « Creating and using geospatial ontology time series in a semantic cultural heritage portal », in *5th European Semantic Web Conference (ESWC)* (Tenerife, Canary Islands, Spain), Springer, 2008, p. 110-123.

[24] B. Kiessling, « Kraken-an universal text recognizer for the humanities », in *Éd., Actes de la conférence Digital Humanities* (Utrecht, The Netherlands), 2019.

[25] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han & S. Park, « OCR-Free Document Understanding Transformer », in *Computer Vision – ECCV 2022* (Cham) (S. Avidan, G. Brostow, M. Cissé, G. M. Farinella & T. Hassner, éds.), Springer Nature, 2022, p. 498-517.

[26] M. Kišš, K. Beneš & M. Hradiš, « AT-ST: Self-training Adaptation Strategy for OCR in Domains with Limited Transcriptions », in *Document Analysis and Recognition – ICDAR 2021*, vol. 12824, Springer International Publishing, 2021, p. 463-477.

[27] O. Kodym & M. Hradiš, « Page Layout Analysis System for Unconstrained Historic Documents », in *Document Analysis and Recognition – ICDAR 2021* (J. Lladós, D. Lopresti & S. Uchida, éds.), vol. 12822, Springer International Publishing, Cham, 2021, p. 492-506.

[28] J. Kohút & M. Hradiš, « TS-Net : OCR Trained to Switch Between Text Transcription Styles », in *Document Analysis and Recognition – ICDAR 2021* (J. Lladós, D. Lopresti & S. Uchida, éds.), Springer Int. Publishing, 2021, p. 478-493.

[29] A. Leborgne, A. Meyer, H. Giraud, F. Le Ber & S. Marc-Zwecker, « Un graphe spatio-temporel pour modéliser l'évolution de parcelles agricoles », in *Conférence internationale francophone en analyse spatiale et géomatique SAGEO*, 2019.

[30] C. Li, B. Bi, M. Yan, W. Wang, S. Huang, F. Huang & L. Si, « StructuralLM: Structural Pre-training for Form Understanding », in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (Online), Association for Computational Linguistics, 2021, p. 6309-6318.

[31] J. Li, A. Sun, J. Han & C. Li, « A Survey on Deep Learning for Named Entity Recognition », *IEEE Transactions on Knowledge and Data Engineering* **34** (2020), nº 1, p. 50-70.

[32] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li & F. Wei, « TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models », *Proceedings of the AAAI Conference on Artificial Intelligence* **37** (2023), nº 11, p. 13094-13102.

[33] A. Mansouri, L. S. Affendey & A. Mamat, « Named entity recognition approaches », *TAL* **52** (2008), nº 1, p. 339–344.

[34] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah & B. Sagot, « CamemBERT: a Tasty French Language Model », in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, p. 7203-7219, `https://arxiv.org/abs/1911.03894`.

[35] D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol-Taravella & D. Nouvel, « Casen: a transducer cascade to recognize french named entities », *TAL* **52** (2011), nº 1, p. 69—96.

[36] A. McCallum & W. Li, « Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons », in *Proceedings of the seventh conference on Computational Natural language learning at HLT-NAACL 2003*, 2003, p. 188-191.

[37] D. Nadeau & S. Sekine, « A survey of named entity recognition and classification », *Lingvisticae Investigationes* **30** (2007), nº 1, p. 3—26.

[38] G. Nagy & S. C. Seth, « Hierarchical representation of optically scanned documents », in *Seventh International Conference on Pattern Recognition, Proceedings, Volume 1* (Montreal, Canada), 1984, p. 347-349.

[39] A.-C. Ngonga Ngomo, « Orchid–reduction-ratio-optimal computation of geo-spatial distances for link discovery », in *The Semantic Web -- ISWC 2013* (Sydney, Australia), Springer, 2013, p. 395-410.

[40] J. Nothman, N. Ringland, W. Radford, T. Murphy & J. R. Curran, « Learning multilingual named entity recognition from Wikipedia », *Artificial Intelligence* **194** (2013), p. 151-175.

[41] D. Nouvel, J.-Y. Antoine, N. Friburger & A. Soulet, « Recognizing Named Entities using Automatically Extracted Transduction Rules », in *5th Language and Technology Conference* (Poznan, Poland), 2011, p. 136-140.

[42] L. O'Gorman, « The document spectrum for page layout analysis », *IEEE Transactions on pattern analysis and machine intelligence* **15** (1993), nº 11, p. 1162-1173.

[43] J. Puigcerver, « Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? », in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Kyoto,Japan), vol. 01, 2017, p. 67-72.

[44] F. Scharffe, A. Ferrara & A. Nikolov, « Data linking for the semantic web », *International Journal on Semantic Web and Information Systems* **7** (2011), nº 3, p. 46-76.

[45] Z. Shen, R. Zhang, M. Dell, B. C. G. Lee, J. Carlson & W. Li, « Layoutparser: A unified toolkit for deep learning based document image analysis », in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Proceedings, Part I 16* (Lausanne, Switzerland), Springer, 2021, p. 131-146.

[46] W. Siabato, C. Claramunt, S. Ilarri & M. Á. Manso-Callejo, « A survey of modelling trends in temporal GIS », *ACM Computing Surveys (CSUR)* **51** (2018), nº 2, p. 1-41.

[47] K. Simonyan & A. Zisserman, « Very Deep Convolutional Networks for Large-Scale Image Recognition », in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio & Y. LeCun, éds.), 2015.

[48] R. Smith, « An overview of the Tesseract OCR engine », in *Int. Conf. on Doc. Analysis and Recognition*, vol. 2, IEEE, 2007, p. 629-633.

[49] P. Sutheebanjard & W. Premchaiswadi, « A modified recursive XY-cut algorithm for solving block ordering problems », in *2nd International Conference on Computer Engineering and Technology*, vol. 3, IEEE, 2010.

[50] D. Symeonidou, V. Armant, N. Pernelle & F. Saïs, « SAKey: Scalable Almost Key Discovery in RDF Data », in *Proceedings of the 13th International Semantic Web Conference, ISWC 2014* (Riva del

Garda, Italy), vol. Lecture Notes in Computer Science, The Semantic Web – ISWC 2014, vol. 8796, Springer Verlag, 2014, p. 33-49.

[51] S. Tual, N. Abadie, J. Chazalon, B. Duménieu & E. Carlinet, « A Benchmark of Nested Named Entity Recognition Approaches in Historical Structured Documents », in *Document Analysis and Recognition - ICDAR 2023* (Cham), Springer Nature Switzerland, 2023, p. 115-131.

[52] S. Tual, N. Abadie, B. Duménieu, J. Chazalon & E. Carlinet, « Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du xixe siècle : application aux métiers de la photographie », in *34e Journées francophones d'Ingénierie des Connaissances (IC 2023)@ Plate-Forme Intelligence Artificielle (PFIA 2023)* (Strasbourg), 2023.

[53] D. Vinasco-Alvarez, J. Samuel, S. Servigne & G. Gesquière, « Towards limiting semantic data loss in 4D urban data semantic graph generation », in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 8, 2021, p. 37-44.

[54] C. Wick, C. Reul & F. Puppe, « Calamari – A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition », *Digital Humanities Quarterly* **14** (2020), no 2.

[55] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei & M. Zhou, « 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*, 2020, p. 1192-1200.

Abstract. — Business directories have been published at a high frequency in many European cities throughout the xixth and xxth centuries. This corpus of historical sources is unique because of its volume and the opportunity it gives to follow urban transformations through the professional activities of the inhabitants, from the individual scale to that of the entire city. However, the spatio-temporal analysis of businesses of a given type through directory entries requires a considerable amount of manual work. To overcome this difficulty, this article proposes an automatic approach to construct and visualise a geohistorical knowledge graph of businesses listed in old directories. The approach is tested on xixth century Parisian trade directories from 1798 to 1914, on the case of photographers.

Keywords. — Geohistorical knowledge graph, old directories, named entity recognition and linking, OCR noise, spatio-temporal visualization.
