Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19 ème siècle: application aux métiers de la photographie

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

Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19ème siècle: application aux métiers de la photographie

S. Tual[1], N. Abadie[1], B. Duménieu[2], J. Chazalon[3], E. Carlinet[3]

[1] LASTIG, Univ. Gustave Eiffel, IGN-ENSG
[2] CRH, EHESS
[3] LRDE, EPITA

solenn.tual@ign.fr; nathalie-f.abadie@ign.fr; bertrand.dumenieu@ehess.fr; joseph.chazalon@epita.fr; edwin.carlinet@epita.fr

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 1799 à 1908, sur le cas des métiers de la photographie.

Mots-clés

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

Abstract

Business directories have been published at a high frequency in many European cities throughout the 19^th^ and 20^th^ 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 19^th^ century Parisian trade directories from 1799 to 1908, on the case of photographers.

Keywords

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

1 Introduction

A partir de la fin du XVIII^ème^ siècle, les annuaires des habitants et des commerces (voir figure 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, [3] 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. [4] utilise des annuaires parus entre 1936 et 1990 pour localiser les anciennes stations services de la ville de Providence (Rhode Island) aux Etats-Unis, afin de détecter des zones potentiellement polluées. L'approche proposée par [4] 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'évolu-

tion 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 oeuvre sur les entrées relatives aux métiers de la photographie, mais elle peut être appliquée à n'importe quelle activité professionnelle représentée dans ces annuaires.

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 OCR

Les approches d'analyse de mise en page visent à identifier et étiquetter les régions homogènes de documents. Dans son état de l'art, [7] distingue trois stratégies. Les stratégies descendantes ou *top-down*, comme XY-cut [20] et ses dérivées [11, 26], appliquent des règles pour divisiser 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 [22] ou les approches par apprentissage automatique, 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. Enfin, les approches hybrides combinent des techniques ascendantes et descendantes. Plus récemment, l'essor des réseaux de neurones de type transformer a conduit à la proposition de nombreuses approches multimodales comme LayoutLM [29] et ses variantes comme [16]. Elles tirent parti des informations textuelles, visuelles et spatiales des documents pour reconnaître des mises en pages très complexes et variées, mais nécessitent des ressources importantes pour être mises en oeuvre.

Les systèmes d'OCR récents, comme Tesseract [25], OCRopus [8], Kraken [13], Calamari [28] ou Pero OCR [14] 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 [14].

2.2 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 [19]. Les approches à base de règles utilisent des patrons lexico-syntaxiques combinant catégories grammaticales et entrées de dictionnaires [4, 18]. Sur des corpus spécialisés, lorsque l'on dispose de dictionnaires exhaustifs, elles produisent de bons résultats, mais l'élaboration des patrons constitue un effort important. 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. Les secondes, en revanche, définissent leurs propres descripteurs pour classer les tokens selon leur appartenance à un type d'entités nommées. Les modèles de langue récents peuvent être adaptés à des corpus spécialisés avec relativement peu de données d'entraînement et sont très susceptibles de produire les meilleurs résultats [17].

2.3 Construction de graphes géohistoriques et liage de ressources

De nombreux modèles ont été proposés pour représenter des données spatio-temporelles [24]. Les travaux récents sur la représentation des états passés successifs du territoire, reposent majoritairement sur des modèles de graphes. Ainsi [6, 15, 5] s'inspirent du modèle de graphe spatio-temporel de [10] ; 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 [1], et dans le troisème, 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. Ces trois 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 sucessifs 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. [23] 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 suceptibles de représenter une même entité du monde réel. C'est le type d'approche mis en oeuvre par des outils comme Silk [2] [12] ou LIMES [3] [21]. Les méthodes fondées sur les connaissances exploitent les connaissances fournies par l'ontologie qui décrit les données. Les restrictions désignant des ensembles de propriétés

1. Voir : https ://geoservices.ign.fr/rpg
2. http ://silkframework.org/
3. http ://aksw.org/Projects/LIMES.html

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 [27] ou [2]. Les approches proposées par [6, 15, 5] 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.

3 Questions de compétence

Ce travail vise à adapter et étendre la chaîne de traitement proposée par [4] 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 pérénité, 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 suffisament générales pour pouvoir s'appliquer à la plupart des types de commerces figurant dans les annuaires.

CQ1. Quelle est l'adresse du commerce X en 1861 ?
CQ2. Combien y a-t-il de commerces de ce type localisés rue de Rivoli en 1856 ?
CQ3. Quels sont les commerces situés dans une zone définie par un polygone ou un rectangle englobant en 1875 ?
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 connaissance géohistoriques 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 connaisances qui reprend et adapte les étapes de la chaîne de traitement de [4] et les approches à base de liage de [6], [15] et [5].

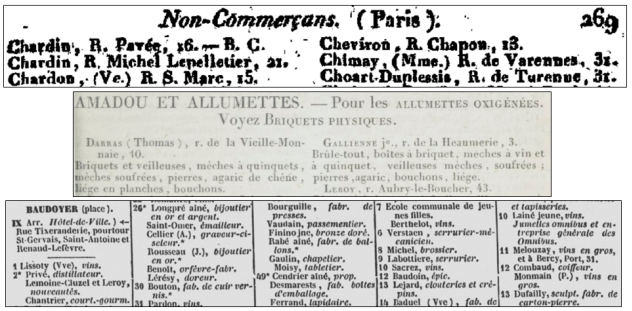

FIGURE 1 – Exemples de mises en pages et d'index différents dans les annuaires *En haut* : Duverneuil et La Tynna 1806 - index par noms ; *Au milieu* : Deflandre 1828 - index par professions ; *En bas* : Bottin 1851 - index par rues.

4.1 Les annuaires du commerce parisien

Le corpus utilisé rassemble des annuaires publiés annuellement entre 1799 et 1908 par différents éditeurs et couvrant 88 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. (voir Figure 1). 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.

4.2 Segmentation de mise en page et OCR

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. Le choix a donc été fait de mettre en oeuvre une approche hybride à base de techniques classiques de nettoyage des scans et d'analyse de leurs mises en page pour détecter les entrées : 1) XY-cuts et classification de régions, 2) Détection des lignes (watershed), 3) Regroupement des lignes en entrées. Sur notre corpus, ces techniques s'avèrent extrêmement performantes et peu coûteuses à mettre en oeuvre.

Enfin, pour extraire le texte de chaque entrée, nous avons utilisé la version "sur étagère" de l'outil Pero OCR.

4.3 Reconnaissance des 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. A celà, s'ajoutent les erreurs de l'OCR. Les approches de reconnaissance d'entités nommées à base de règles semblent donc inappropriées, car elle

nécessiteraient de définir un nombre de règles trop important pour gérer tous les cas.
Nous avons donc adopté une approche supervisée, et adapté un réseau de neurones profond de reconnaissance d'entités nommées utilisant le modèle de langue CamemBERT pour traiter notre corpus. Nous avons procédé à un pré-entrainement non-supervisé sur plusieurs milliers de pages d'annuaires et à un entraînement supervisé sur un corpus annoté 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. Sur notre corpus de test, de 1669 entrées, également bruitées, le modèle obtient ainsi un score de F-mesure globale de 94.1%. Les étapes d'extraction du texte et de reconnaissance des entités nommées et les résultats obtenus sont décrits en détail dans [1].

4.4 Géocodage historique des entrées

La mise en oeuvre des deux étapes précédente a permis de construire une base de données comportant 9 821 898 entrées. Pour doter les entrées d'annuaires de coordonnées, nous avons procédé au géocodage des adresses. Nous avons utilisé le géocodeur historique développé par le groupe de travail Geohistoricaldata [4]. Sa base de données d'adresses a été saisie à partir de différents plans de Paris du XIX^{e} siècle et l'outil favorise les adresses issues de plans dont la date de production est proche de celle des données à géocoder [5].

4.5 Sélection et représentation des entrées en RDF

Pour faciliter la suite du traitement, nous proposons de filtrer les entrées pour ne conserver que celles concernant le type de commerces à étudier. Dans le cas des métiers de la photographie, nous nous sommes appuyés sur une liste de 252 photographes parisiens extraite de l'ouvrage de [9], 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*. Puis nous avons converti en RDF et exporté les 34 062 entrées dont l'attribut "activité" comportait ces mots-clés, à l'aide d'un script R2RML. 26 275 d'entre elles ont pu être géocodées à l'étape précédente, soit environ 70% du jeu de données.

4.6 Liage des entrées

Deux méthodes de liage des entrées ont été implémentées. La première méthode génère les liens par inférences en utilisant les clés déclarées dans l'ontologie qui décrit les données. Les propriétés qui composent les clés sont identifiées à l'aide de Sakey [27] : (1) le numéro de l'entrée, (2) le nom et l'adresse, (3) le nom et l'activité et (4) l'activité et l'adresse. Les clés 2 à 4 sont des 1-quasi-clés identifiées sur les données d'un annuaire. On tolère donc une exception afin de gérer l'existence possible de deux index dans le même annuaire. 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.
La seconde méthode est fondée sur les données. Elle exploite la similarité des valeurs des propriétés des ressources. Elle est mise en oeuvre avec Silk. Après suppression des caractères spéciaux et passage des caractères en minuscule, la distance d'édition Token-Wise est calculée pour les propriétés nom, activité et adresse de chaque paire de ressources. Le score de similarité associé à l'adresse est produit à l'aide d'une moyenne pondérée des résultats obtenus pour les valeurs de LOC et CARDINAL. Enfin, pour les combinaisons de propriétés suivantes - Nom et Activité, Nom et Adresse, Activité et Adresse - le score agrégé retenu correspond à la valeur de la distance Token-Wise la plus faible obtenue par l'une des propriétés de la combinaison. Finalement, seuls les liens dont le score est supérieur à 0.8 sont conservés. 250 622 liens *owl :sameAs* ont été créés avec la méthode fondée sur les connaissances et 357 130 avec la méthode fondée sur les données. Le nombre total des liens calculés et inférés est finalement de 401 852 liens distincts.

5 Visualisation et évaluation

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 Evaluation des questions de compétences

Ainsi, notre première question de compétence peut se vérifier avec la requête suivante, qui renvoie "11 rue suger" :

```
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 { ?e a ont :Entry.
?e rdfs :label ?label.
?e prov :wasDerivedFrom ?directory.
?directory pav :createdOn "1861"@fr.
?e locn :address ?add.
?add locn :fullAddress ?fullAdd.
Filter regex( ?label, "gallino").}
```

Les requêtes SPARQL correspondant aux autres questions de compétences listées sont fournies, avec l'en-

4. `https://api.geohistoricaldata.org/docs/#/Geocoding`

5. Ce jeu de données géocodées est publié sur le dépôt suivant : `https://nakala.fr/10.34847/nkl.98eem49t`

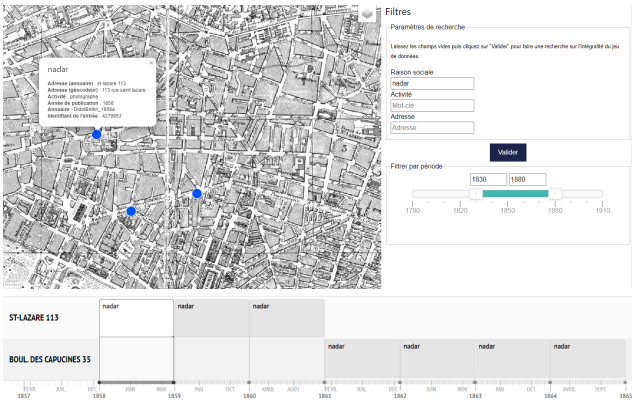

FIGURE 2 – Déménagement du photographe Nadar

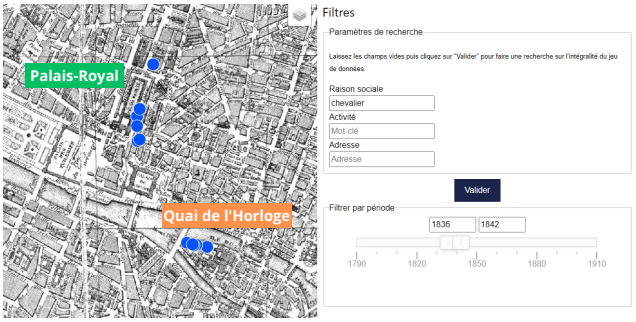

FIGURE 3 – Les ateliers dont les propriétaires se nomment "Chevalier" se concentrent dans deux quartiers.

semble des scripts relatifs à ce travail, l'ontologie et l'application de visualisation :`https://github.com/soduco/ic_2023_photographes_parisiens`

5.2 Visualisation

Pour explorer les données du graphe de façon intuitive, nous avons développé une application de visualisation cartographique et temporelle. 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 2, 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 3 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 4 montre qu'au moins 4 photographes se sont succédés au 59 rue de Rivoli au cours de la seconde moitié du XIX^{e} siècle.

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. Les deux stratégies de liage adoptées permettent de créer suffisament de liens entre entrées représentant un même commerce au cours du temps pour

FIGURE 4 – Phénomène de transmission probable d'un atelier entre photographes.

suivre l'évolution des entrées issues d'éditions successives. Le géocodage des entrées et leur visualisation cartographique permettent en outre d'identifier aisément des phénomènes spatiaux. Trois perspectives à cours terme sont prévues : l'explicitation des relations spatio-temporelles entre entrées, la publication du graphe sur le Web et la mise en oeuvre de l'approche pour d'autres types de commerces.

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.

Références

[1] N. Abadie, E. Carlinet, J. Chazalon, and B. Duménieu. A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories. In S. Uchida, E. Barney, and V. Eglin, editors, *Document Analysis Systems. DAS 2022.*, number 13237 in Document Analysis Systems. DAS 2022., La Rochelle, France, May 2022. Springer, Cham.

[2] Nacira Abbas, Jérôme David, and Amedeo Napoli. Linkex : A Tool for Link Key Discovery Based on Pattern Structures. In *ICFCA 2019 - workshop on Applications and tools of formal concept analysis*, Proc. ICFCA workshop on Applications and tools of formal concept analysis, pages 33–38, Frankfurt, Germany, June 2019. abbas2019a.

[3] Thilo Albers and Kalle Kappner. Perks and Pitfalls of City Directories as a Micro-Geographic Data Source. *Rationality and Competition*, Discussion Paper No. 315, January 2022.

[4] Samuel Bell, Thomas Marlow, Kai Wombacher, Anina Hitt, Neev Parikh, Andras Zsom, and Scott Frickel. Automated data extraction from historical city directories : The rise and fall of mid-century gas stations in Providence, RI. *PLOS ONE*, 15(8) :e0220219, August 2020. Publisher : Public Library of Science.

[5] Camille Bernard, Marlène Villanova-Oliver, and Jérôme Gensel. Theseus : A framework for managing knowledge graphs about geographical divisions and their evolution. *Transactions in GIS*, 2022.

[6] Duménieu Bertrand. *Un système d'information géographique pour le suivi d'objets historiques urbains à travers l'espace et le temps*. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, 2015.

[7] Galal M Binmakhashen and Sabri A Mahmoud. Document layout analysis : a comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6) :1–36, 2019.

[8] Thomas M Breuel. The OCRopus open source OCR system. In *Document Recognition and Retrieval XV*, volume 6815, page 68150F. Int. Soc. for Optics and Photonics, 2008.

[9] Marc 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*. Number 2. Archives nationales, Pierrefitte-sur-Seine, 2015.

[10] Del Mondo Géraldine. *Un modèle de graphe spatio-temporel pour représenter l'évolution d'entités géographiques*. PhD thesis, Université de Brest, 2011.

[11] Jaekyu Ha, Robert M Haralick, and Ihsin T Phillips. Recursive xy cut using bounding boxes of connected components. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 952–955. IEEE, 1995.

[12] Robert Isele, Anja Jentzsch, and Christian Bizer. Efficient multidimensional blocking for link discovery without losing recall. In *WebDB*, 2011.

[13] Benjamin Kiessling. Kraken-an universal text recognizer for the humanities. In *Alliance of Digital Humanities Organizations (ADHO), Éd., Actes de la conférence Digital Humanities*, Utrecht, The Netherlands, 2019.

[14] Jan Kohút and Michal Hradiš. TS-Net : OCR trained to switch between text transcription styles. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, *Document Analysis and Recognition – ICDAR 2021*, pages 478–493. Springer Int. Publishing, 2021.

[15] Aurélie Leborgne, Adrien Meyer, Henri Giraud, Florence Le Ber, and Stella Marc-Zwecker. Un graphe spatio-temporel pour modéliser l'évolution de parcelles agricoles. In *Conférence internationale francphone en analyse spatiale et géomatique SAGEO*, 2019.

[16] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. StructuralLM : Structural pre-training for form understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 6309–6318, Online, August 2021. Association for Computational Linguistics.

[17] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1) :50–70, 2020.

[18] Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iriss Eshkol-Taravella, and Damien Nouvel. Casen : a transducer cascade to recognize french named entities. *TAL*, 52(1) :69–96, 2011.

[19] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3—26, 2007.

[20] George Nagy and Sharad C Seth. Hierarchical representation of optically scanned documents. In *" International conference on Pattern Recognition*, 1984.

[21] Axel-Cyrille Ngonga Ngomo. Orchid–reduction-ratio-optimal computation of geo-spatial distances for link discovery. In *International Semantic Web Conference*, pages 395–410. Springer, 2013.

[22] Lawrence O'Gorman. The document spectrum for page layout analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 15(11) :1162–1173, 1993.

[23] François Scharffe, Alfio Ferrara, and Andriy Nikolov. Data linking for the semantic web. *International Journal on Semantic Web and Information Systems*, 7(3) :46–76, 2011.

[24] Willington Siabato, Christophe Claramunt, Sergio Ilarri, and Miguel Ángel Manso-Callejo. A survey of modelling trends in temporal gis. *ACM Computing Surveys (CSUR)*, 51(2) :1–41, 2018.

[25] Ray Smith. An overview of the tesseract OCR engine. In *Int. Conf. on Doc. Analysis and Recognition*, volume 2, pages 629–633. IEEE, 2007.

[26] Phaisarn Sutheebanjard and Wichian Premchaiswadi. A modified recursive xy cut algorithm for solving block ordering problems. In *2010 2nd International Conference on Computer Engineering and Technology*, volume 3, pages V3–307. IEEE, 2010.

[27] Danai Symeonidou, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. SAKey : Scalable Almost Key Discovery in RDF Data. In Springer Verlag, editor, *In proceedings of the 13th International Semantic Web Conference, ISWC 2014*, volume Lecture Notes in Computer Science of *The Semantic Web – ISWC 2014*, pages 33–49, Riva del Garda, Italy, October 2014. Editions Springer.

[28] Christoph Wick, Christian Reul, and Frank Puppe. Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Digital Humanities Quarterly*, 14(1), 2020.

[29] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming 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*, pages 1192–1200, 2020.