RAPPORT SCIENTIFIQUE DU PROJET POPP

Projet d'océrisation des recensements parisiens

Projet Lauréat CollEx-Persée 2019-2020

Projet porté par Sandra Brée (CNRS, LARHRA) et François Merveille (GED Campus Condorcet), en partenariat avec Thierry Paquet (LITIS).

Financé par le GIS CollEx-Persée, et le co-financement de l'Infrastructure de recherche Progedo et du Grand Équipement Documentaire du Campus Condorcet.

Table des matières

I. Présentation du projet 3
1. Les recensements de population : source pour l'histoire des populations 3
2. Créer des bases de données 5
3. Au-delà de l'intérêt scientifique, un intérêt méthodologique 6
II. Déroulement du projet 6
1. Création de la base de données 6
2. Nettoyage de la base de données 9
3. Créer de nouvelles variables 10
III. Réalisations et perspectives 11
1. Le projet POPP et la science ouverte 11
2. Livrables 12
2.1. Publications et communications 12
2.2. Logiciels libres et annotations ouvertes 14
3. L'analyse historique : perspectives de recherche 14
IV. Bilan 16
1. Réussites et échecs 16
2. Ouverture 17
V. Annexe : l'équipe du projet 18

I. Présentation du projet

Porté par le Grand équipement documentaire du Campus Condorcet et par le Laboratoire de Recherche Historique Rhône-Alpes, (LARHRA – UMR 5190) en association avec le Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes de Rouen (LITIS – EA 4108), le projet POPP a été lauréat de l'appel à projets CollEx-Persée 2019-2020, et a bénéficié de co-financements de l'Infrastructure de recherche Progedo[1] et du Grand Équipement Documentaire du Campus Condorcet.

Ce projet visait à élaborer une vaste base de données (12 millions d'individus) à partir des recensements nominatifs de Paris de 1926, 1931, 1936 et 1946 qui sont les seules listes nominatives de la population parisienne[2]. Ces recensements sont déjà numérisés et en ligne sur le site des archives de Paris[3] mais il s'agissait ici de créer une base de données permettant l'exploitation statistique de ces documents particulièrement riches et qui offrira l'opportunité d'un pas important dans la connaissance de la population urbaine européenne jusqu'alors très peu étudiée. Outre son intérêt pour la connaissance scientifique, ce projet présente également une avancée importance du point de vue méthodologique. En effet, la création de la base de données susmentionnée nécessite l'utilisation de technique de reconnaissance optique de caractère et la collaboration entre chercheurs en sciences sociales et en informatique.

1. Les recensements de population : source pour l'histoire des populations

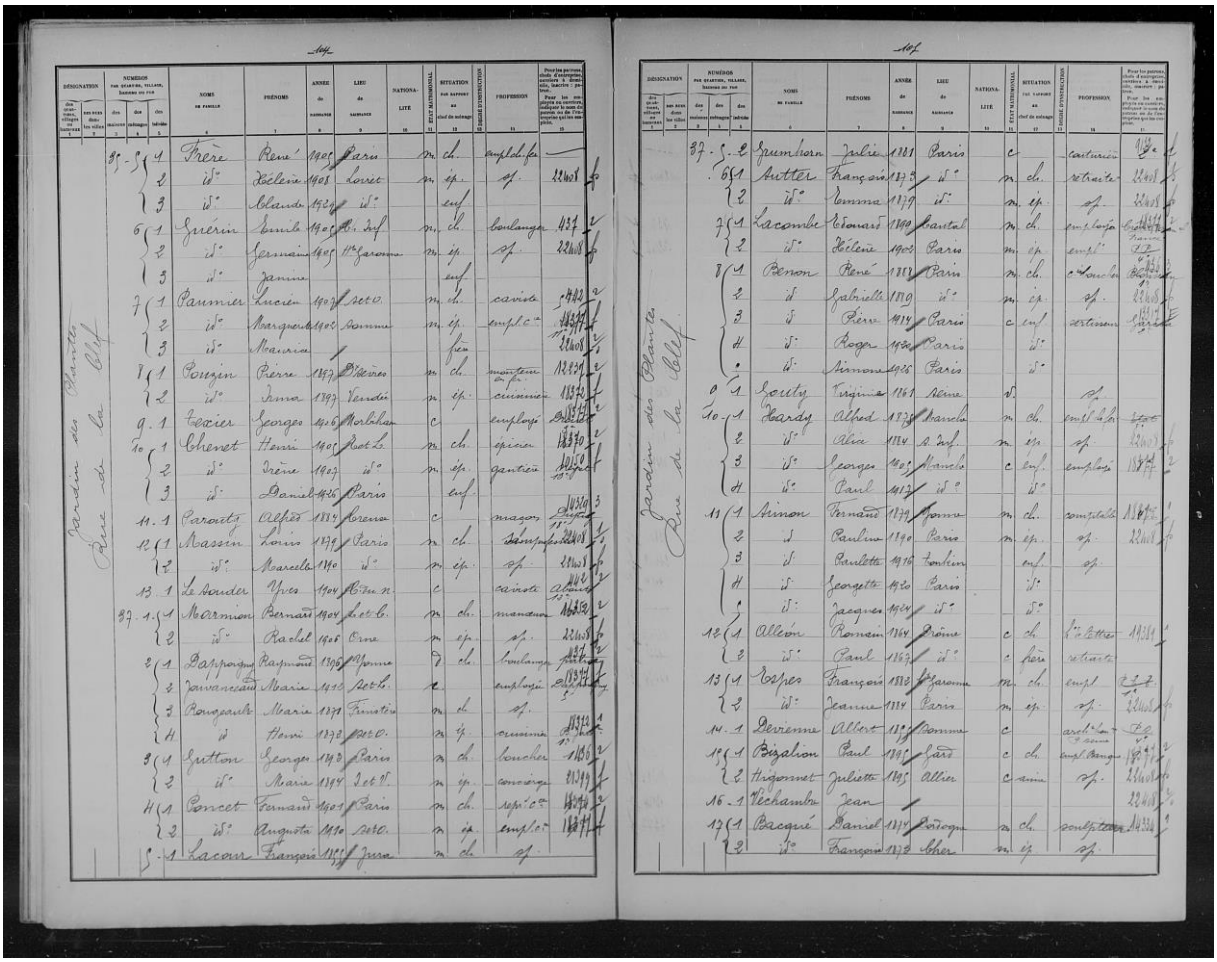
Les recensements de population sont, avec les registres d'état civil, l'une des principales sources de recherches pour la démographie historique et, plus généralement, l'histoire de la population. Les personnes de tous âges et de tous sexes, résidant dans la commune, ou de passage, sont recensées individuellement et nominativement. La liste des habitants est établie par ménage, avec en tête le chef de famille, suivi de sa femme (ou de sa compagne dans le cas de Paris comme on le verra), puis de ses enfants, puis des ascendants ou autres parents appartenant au même ménage, puis éventuellement des domestiques ayant la même résidence, des apprentis vivant dans la même maison, etc. La population est distinguée en trois grands groupes : la population de résidence habituelle, la population comptée à part (population des casernes, internats des lycées, hôpitaux, prisons, congrégations religieuses, etc), les hôtes de passage.

Les informations contenues dans les listes nominatives communales ont varié dans le temps. Pour les recensements de l'entre-deux-guerres à Paris, on connaît, pour chaque individu (un peu moins de 3 millions chaque année[4]), ses nom et prénom, son année et son lieu de naissance, sa nationalité, son lien avec le chef de famille et sa profession (Figure 1). Ses informations sont riches et permettent d'analyser un nombre considérable de problématiques, de la structure par

[1] www.progedo.fr
[2] En effet, à partir du recensement de 1866, Paris obtient le droit de ne pas établir de liste nominative à partir de ces bulletins individuels en raison de la taille de la ville et du coût de l'opération, et il semble bien qu'il n'y en ait pas eu depuis 1817. En tout état de cause, aucune liste nominative de population ne nous est parvenue pour Paris avant celle de 1926 (Biraben, 1963).
[3] https://archives.paris.fr/s/11/denombrements-de-population/?
[4] 1926 : 2 871 429 habitants, 1931 : 2 891 020 habitants ; 1936 : 2 829 746 habitants.

sexe et âge de la population à la structure professionnelle, de la composition des ménages à l'origine des habitants (voir plus loin). Pourtant, les recensements de population ont fait l'objet de peu de recherches en France.

Figure 1. Liste nominative du recensement de la population de 1926, population de résidence habituelle, quartier de Belleville. Cote D2M8 307 (Archives de Paris).



Les recensements de la population sont des sources administratives qui ont été souvent mal conservées au XIX[e] siècle et de manière assez aléatoire au siècle suivant. Par ailleurs, plusieurs circulaires (notamment celle de 1887) ont permis aux services d'archives de ne pas conserver les listes nominatives dressées à partir des bulletins individuels.

Ceci explique probablement, en partie, la moindre utilisation de ces sources par les historiens démographes français qui se sont d'abord intéressés à l'analyse des registres paroissiaux et d'état civil pour reconstituer la population française de l'Ancien Régime[5]). L'intérêt de ces

[5] Goubert P., 1960, Beauvais et le Beauvaisis de 1600 à 1730.Contribution à l'histoire sociale de la France au XVIIe siècle, Paris, S.E.V.P.E.N. ; Henry L., Fleury M., 1956, Des registres paroissiaux à l'histoire de la population. Manuel de dépouillement et d'exploitation de l'état civil ancien, Paris, Institut national d'études démographiques.

sources est pourtant pointé par l'anglais Peter Laslett dès le début des années 1960[6], notamment pour comprendre la composition des ménages. Ce n'est que depuis la fin des années 2000 qu'un certain regain d'intérêt des démographes historiens français se fait sentir pour les recensements. À commencer par l'enquête sur Charleville[7] pour laquelle Biraben[8] signalait dès le début des années 1960 une série de listes nominatives particulièrement remarquables. L'enquête, toujours en cours, tire, en effet, profit de l'exceptionnelle collection de recensements annuels des habitants de Charleville, conservés pour une période allant de la fin du XVII^{e} au début du XX^{e} siècle et permettant notamment un suivi nominatif des habitants.

Le projet POPP est le premier projet à envisager la création d'une vaste base de données à l'aide des techniques de *Deep Learning* et d'océrisation. À sa suite, le projet ANR Socface[9], portée par Lionel Kesztenbaum (INED) depuis octobre 2021, a pour ambition de transcrire automatiquement l'ensemble des listes nominatives du recensement de 1836 à 1936.

2. Créer des bases de données

L'une des préoccupations majeures de l'histoire quantitative, et notamment en démographie historique, est la création de bases de données de la manière la plus efficace possible. Les progrès de l'intelligence artificielle et du *Deep Learning* ont permis aux démographes historiens d'envisager la collecte d'informations de manière automatique. L'océrisation (de l'anglais OCR, *optical character recognition*), c'est-à-dire la reconnaissance optique des caractères imprimés permet aujourd'hui aux ordinateurs de déchiffrer les formes d'une image numérisée et de les traduire en caractères. Ces techniques sont beaucoup utilisées par les bibliothèques qui ont des plateformes numériques (comme le site Gallica de la BNF par exemple) pour permettre aux usagers de chercher des mots dans les documents ainsi océrisés. La création d'une base de données reposant sur cette technique nécessite cependant de nombreuses autres étapes en plus de l'océrisation comme cela sera développé plus loin. Par ailleurs, les listes nominatives de recensement sont constituées à la main comme visualisé sur la figure 1, ce qui nécessite de faire franchir à la machine une difficulté supplémentaire par rapport aux techniques d'OCR traditionnelles.

L'intérêt majeure de ces techniques pour la création de bases de données en démographie historique, outre le gain de temps évident, est de pouvoir envisager la collecte d'informations pour des populations très conséquentes. Ainsi, ces nouvelles techniques permettent d'envisager de travailler sur des populations vastes, et donc notamment sur les population des grandes villes, longtemps laissées de côté (à quelques exceptions), notamment pour les XIX^{e} et XX^{e} siècles.

[6] Laslett P. Harrison J.,1963, "Clayworth and Cogenhoe," in H.E. Bell, R. L. Ollar, *Historical essays, 1600-1750*.
[7] Boudjaaba F., Gourdon V., Rathier C., 2010, "Charleville's census reports: an exceptional source for the longitudinal study of urban populations in France," Popolazione e Storia, 2, 17–42
[8] Biraben, J-N.,1963, "Inventaire des listes nominatives de recensement en France," *Population*, 18 (2), 305–328.
[9] https://socface.site.ined.fr/

3. Au-delà de l'intérêt scientifique, un intérêt méthodologique

Outre son intérêt pour la connaissance scientifique, ce projet présente également une avancée importance du point de vue méthodologique. En effet, la création de la base de données susmentionnée nécessite l'utilisation de techniques de reconnaissance optique de caractères manuscrits.

Travaillant sur l'Equipex D-FIH, l'équipe du laboratoire LITIS EA 4108[10] (portée par Thierry Paquet) a mis au point des logiciels de lecture optique de plus en plus performants pour extraire des informations boursières (nom d'entreprise, noms de personnes, des montants financiers, des dates…) imprimées dans les annuaires financiers historiques et les cotations boursières afin de constituer des bases de données boursières et financières historiques.

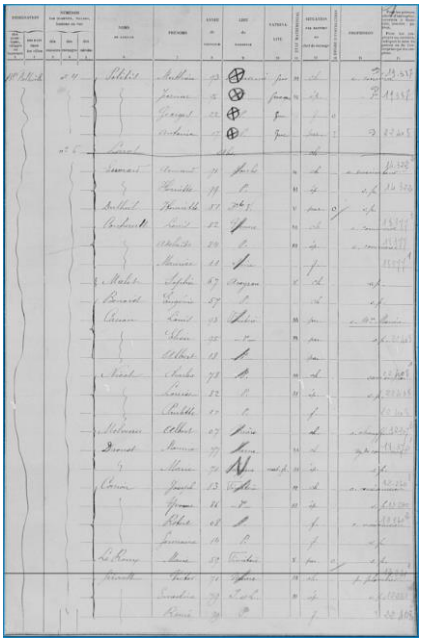
Dans cette continuité, le projet Popp constituait un véritable défi à relever puisque les recensements de population ne sont pas tapuscrits comme les données de la bourse mais remplis à la main.

II. Déroulement du projet

1. Création de la base de données

L'objectif du projet est de passer d'images numérisées fournies par les archives de Paris à une base de données exploitables statistiquement (figure 2).

Figure 2. De l'image numérisée à la base de données



name	Noms	Prenoms	Annee de naissance	Lieu de naissance	Nationalite	Etat matrimonial	Sitation par rapport au chef de menage	Degre d'instruction	Profession	ID entreprise
07_0014	POLITIS	MATHIEU	93	ROUMANIE	GREC	M	CH	¤	$E CINEMNA$	19387
07_0014	¤	JEANNE	96	P	GRECQUE	M	EP	¤	D'	19387
07_0014	¤	GEORGES	22	P	$GAUE$	¤	F	O	¤	¤
07_0014	¤	ANTOINE	7	P	$GRCE$	¤	PAR	¤	D	22408
07_0014	EUVRARD	ARMAND	91	DOUBS	¤	M	CH	¤	E DESSINATEUR	¤
07_0014	¤	HENRIETTE	99	P	¤	M	EP	¤	$S PR$	14322
07_0014	DUTHEIL	HENRIETTE	51	$HTE V$	¤	V	PAR	O	$S P$	¤
07_0014	COUHAULT	LOUIS	82	YONNE	¤	M	CH	¤	E COMMERCE	$=37$
07_0014	¤	ADELAIDE	84	P	¤	M	EP	¤	E COMMERCE	1837
07_0014	¤	MAURICE	11	SEINE	¤	¤	F	¤	¤	18377
07_0014	MALET	SOPHIE	67	AVEYRON	¤	V	CH	¤	$S P$	¤
07_0014	CARIOU	LOUIS	93	FINISTERE	¤	M	PAR	¤	$E MRE MARINE$	¤
07_0014	¤	ELISA	95	D°	¤	M	PAR	¤	$S P$	22408
07_0014	¤	ALBERT	18	D°	¤	¤	PAR	¤	¤	¤
07_0014	¤	LOUISE	82	P	¤	M	EP	¤	$S P$	22408
07_0014	¤	PAULETTE	10	P	¤	¤	F	¤	¤	22408
07_0014	MOLVEAU	ALBERT	7	NIEVRE	¤	¤	CH	¤	O CHAUFFEUR	12307
07_0014	DAOUST	MAURICE	77	MARNE	¤	M	CH	¤	$VOY DE COMMERCE$	1
07_0014	CARIOU	JOSEPH	83	FINISTERE	¤	M	CH	¤	O MECANICIEN	¤
07_0014	¤	YVONNE	86	D°	¤	M	EP	¤	$S P$	12260
07_0014	¤	ROBERT	8	P	¤	¤	F	¤	O MECANICIEN	¤
07_0014	GIRAULT	VICTOR	71	NIEVRE	¤	M	CH	¤	P PLOMBIER	¤
07_0014	¤	ERNESTINE	79	$I ET L$	¤	M	EP	¤	$S P$	122401
07_0014	¤	RENEE	99	P	¤	¤	F	¤	¤	22408
07_0015	GIRAULT (CUIS	GEORGES	9	P	¤	¤	F	¤	$O PLONBIER$	21
07_0015	MAES	ALPHONSE	81	NORD	¤	C	CH	¤	$E ASSURANCE$	¤
07_0015	¤	SIMEON	1	D°	¤	C	F	¤	$G DE CAFE$	18373
07_0015	LASSELIN	LUCIE	82	D°	¤	C	A	¤	$S P$	22408
07_0015	HIRAT	MARCEL	76	P	¤	¤	CH	¤	$E COMPTABL$	18375
07_0015	CARPENTIER	HECTOR	71	¤	FR	M	CH	¤	O MENUISIER	13
07_0015	¤	MARIE	79	¤	FR	M	EP	¤	$S P$	11159
07_0015	¤	PAULE	5	¤	FR	¤	F	¤	$O ESSAYEUSE$	9133
07_0015	GUIRANDE	MARIE	77	CORREZE	¤	C	CH	¤	¤	22408
07_0015	¤	FRANCOIS	90	D°	¤	C	PAR	¤	O COIFFEUR	10390
07_0015	GRILLET	LEON	56	SEINE	¤	M	CH	¤	O CORDONNIER	$=3$
07_0015	$VILMIER$	MARIA	71	$BELGOQUE$	$NAT FR$	M	A	¤	$O BLANCHISSEURE$	¤
07_0015	DODARD	PAUL	81	MANCHE	¤	M	CH	¤	¤	224083
07_0015	¤	MARIA	83	¤	¤	M	EP	¤	$S P$	22408
07_0015	¤	ROGER	13	¤	¤	¤	F	¤	¤	¤
07_0015	LEONARDON	FRANCOIS	87	P	¤	M	CH	¤	$E T CRP$	¤
07_0015	¤	AUGUSTINE	92	P	¤	M	EP	¤	$S P$	22408
07_0015	MAUNOIR	AUGUSTIN	8	P	¤	¤	$B-F$	¤	OUPEUR ENCHAISSUF	¤
07_0015	PARPAIS	$FELICITE$	65	P	¤	V	CH	¤	$F DE MENAGE$	¤
07_0015	CLAQUIN	$MOEL$	90	FINISTERE	¤	M	CH	¤	$O CORDONNIE$	1
07_0015	¤	$MARGUERITE$	95	D°	¤	M	EP	¤	$S P$	10147

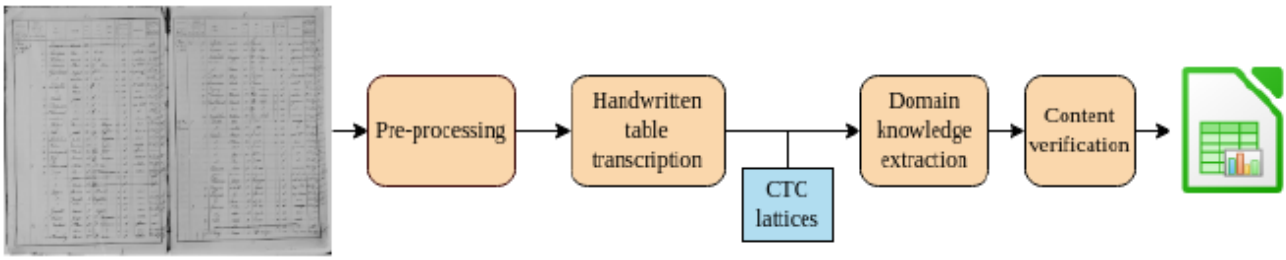
[10] https://www.litislab.fr/

Dans le cadre du projet POPP, nous avons choisi de travailler dans un premier temps sur les trois premiers recensements (1926, 1931 et 1936) car ils ont la même structure et qu'ils permettaient de travailler sur une période cohérente : l'entre-deux-guerres. Cette structure fixe est constituée de tableaux de 30 lignes (voir figures 1 et 2). Elle facilite le processus de reconnaissance de l'information manuscrite puisque chaque cellule contient un type d'information spécifique attendu (un nom, une date, une adresse, etc.) qui peut être modélisé grâce à un dictionnaire ou une expression régulière qui est utilisée pour piloter le processus de reconnaissance.

Les tableaux ayant été remplis à la main, la disposition dans les colonnes est peu respectée. Des mots peuvent être écrits sur deux colonnes et d'autres entre les lignes. De plus, comme les tableaux ont été remplis par plusieurs auteurs, certaines colonnes n'ont pas été remplies de la même manière. Par exemple, la colonne lieu de naissance est parfois remplie avec la ville et le département de naissance, avec le département seulement, ou avec le pays de naissance seulement.

Quatre étapes principales sont nécessaires pour passer de l'image numérisée à la base de données en format CSV (figure 3) : le prétraitement, la reconnaissance de l'écriture manuscrite, l'intégration des connaissances du domaine et la vérification du contenu. L'ensemble du processus décrit ci-dessous est développé dans le *data paper* du projet (Constum et al., 2022)[11].

Figure 3. Étapes de la création de la base de données



Source : (Constum *et al.*, 2022)

- **Prétraitement** : l'étape de prétraitement est consacrée à la détection des tableaux, au détourage des deux tableaux dans l'image de la double page et à la détection des lignes dans ces tableaux.

- **Reconnaissance par lignes** : comme cela a déjà été mentionné, la structure en lignes du tableau a été respectée la plupart du temps. En revanche, la structure en colonnes n'a pas toujours été respectée, faute d'espace pour écrire. Il est très fréquent qu'une information relative à une colonne croise la colonne suivante. Cela signifie qu'il n'est pas possible de détecter les cellules du tableau avant la reconnaissance de leur contenu textuel. C'est donc l'étape de reconnaissance qui permet *a posteriori* la segmentation en colonnes, en respectant une convention d'annotation qui introduit un symbole de saut de colonne, ainsi qu'un symbole qui

[11] https://doi.org/10.1007/978-3-031-06555-2_10

encode une cellule vide. Deux symboles que la machine reconnaît également, pourvu que les données annotées comportent ces informations afin que la machine puisse les apprendre également.

- **Annotation de jeux de données** : pour ajuster le modèle de reconnaissance de texte, un premier ensemble de données d'entraînement était nécessaire. Bien que la structure des tableaux soit extrêmement stable pour chaque page des recensements, il existe une grande variabilité dans les styles d'écriture, la couleur du fond, le type d'encre. Notons que les images ont été fournies par le service des archives de la ville de Paris sous forme de fichiers JPEG-2000 dont la qualité de compression est excellente. Nous avons donc annoté une double page pour chacun des 80 quartiers du recensement de 1926, afin de créer un jeu de données générique aussi représentatif que possible de l'ensemble du corpus. Ce premier jeu de données contient donc 160 pages composées de 4800 lignes de test manuscrites. De plus, afin de mener des expériences avec un seul type d'écriture[12], nous avons annoté un autre jeu de données composé de 49 pages (1470 lignes) provenant du seul quartier de Belleville et écrit par une seule personne[13].

- **Reconnaissance de l'écriture manuscrite : le modèle optique** : Nous avons d'abord entraîné notre modèle sur le jeu de données générique et obtenu un CER (taux d'erreur sur les caractères) de 7,08 % et un WER (taux d'erreur sur les mots) de 19,05 % sur l'ensemble de test. Ce premier résultat est correct, bien que le taux d'erreur soit supérieur aux performances de l'état de l'art sur d'autres jeux de données (2% de CER sur RIMES, et 4% sur la base IAM), certes de qualité plus contrôlée. En ce qui concerne la reconnaissance des mono-écrivains, nous avons d'abord évalué le modèle entraîné avec le jeu de données générique sur le jeu de données Belleville et obtenu 6,42% de CER ; puis, en spécialisant le modèle sur l'ensemble du jeu de données Belleville, nous avons réduit le CER à 3,65%.

- **Auto-entrainement** : Nous avons également utilisé l'auto-entrainement, une technique qui consiste à utiliser un modèle déjà entrainé sur des données étiquetées, appelé enseignant, pour générer des pseudo-annotations qui sont ensuite utilisées pour entrainer un second modèle, appelé étudiant. Ce type de technique est particulièrement utile lorsqu'une très grande quantité de données non étiquetées est disponible, ce qui est le cas dans notre projet. L'ensemble de données non étiquetées comprend dans ce cas les 2,4 millions d'images de lignes sélectionnées aléatoirement à partir du recensement de 1926. En utilisant l'auto-entrainement du système de reconnaissance, nous avons pu améliorer sensiblement le CER pour le faire passer de 7,08% à 4,52%.

- **Modèle linguistique** : Afin d'améliorer et valider dans certains cas la sortie du modèle optique, nous utilisons une combinaison de règles de grammaire et de dictionnaires, en fonction de la colonne à considérer. Un dictionnaire décrit l'ensemble des mots possibles pour une cellule du tableau spécifique (telle que la colonne prénom, ou nom, par exemple), une

[12] L'intérêt est de spécialiser la machine sur un type d'écriture donné et donc de connaître les niveaux de reconnaissance maximum. L'objectif à long terme serait de pouvoir spécialiser un modèle sur une écriture donnée en lui montrant très peu d'images transcrites à la main, voire uniquement des images non transcrites.

[13] Le recensement complet de Belleville a été écrit par trois auteurs, mais leur style d'écriture est très similaire et peut donc être considéré comme un style d'écriture unique.

grammaire est "un formalisme permettant de définir une syntaxe et donc un langage formel, c'est-à-dire un ensemble de mots admissibles sur un alphabet donné". Nous avons ainsi pu créer des dictionnaires en utilisant plusieurs sources[14] d'information et les premières pages annotées, pour certaines colonnes, et des expressions régulières décrivant la syntaxe des contenus attendus d'autres colonnes telles que le lieu de naissance, ou le métier. L'une des plus grandes difficultés dans l'élaboration de ces grammaires a été de tenir compte des nombreuses abréviations utilisées par les différents auteurs pour désigner un même département ou une même profession. Par exemple, les noms des départements tels que "Hautes-Pyrénées" ou "Hautes-Charentes" peuvent être abrégés en "Hte Pyrénées" ou "H. Charentes" ; "femme de chambre" peut être abrégé en "F de C".

Enfin, grâce aux dictionnaires et grammaires, nous avons pu mettre en place une règle de vérification/rejet qui compare la sortie optique et la sortie de la grammaire. Les séquences présentant une trop grande divergence entre le meilleur chemin optique et le meilleur chemin de la grammaire sont mis en évidence par le système au moyen d'un symbole spécifique ('$'). Ainsi, la base de données produite à ce stade comporte des informations reconnues avec une forte confiance, tandis que les éléments reconnus entre les symboles '$' sont plus incertains (ils présentent une certaine incertitude). Une phase de « nettoyage » permet aux historiens de corriger / valider ces informations incertaines issues de la reconnaissance, comme nous allons le voir ci-après.

2. Nettoyage de la base de données

Comme nous venons de le voir, les taux d'erreur de l'OCR sont très bas pour la base de données POPP. Cependant, il reste des erreurs qui peuvent être de différents types. L'ensemble des corrections apportées par l'équipe SHS (composée de Marion Leturcq (démographe et économiste, INED), Yoann Doignon (géographe et démographe, MCFC, Université de Strasbourg), Baptiste Coulmont (Sociologue, ENS Saclay) et Victor Gay (économiste, MCFC TSE, Toulouse) est décrite dans un *data paper* spécifique (en cours[15]) qui a été présenté à la conférence de l'*European Society of Historical Demography* à Madrid au printemps 2022.

À l'issue de la reconnaissance automatique, il subsiste des décalages de colonnes. Dans certains cas, la machine n'a pas détecté les ruptures de colonnes et l'information d'une colonne apparaît fusionnée avec celle d'une une autre. D'autres erreurs peuvent être spécifiques à chaque colonne et chaque variable a été corrigée indépendamment[16].

Plus globalement, il nous a ensuite fallu adapter la base à l'analyse statistique. En effet, malgré les consignes pour remplir les listes nominatives de recensement, les agents qui ont rempli ces

[14] La base de données des personnes décédées depuis 1970 (INSEE) pour les prénoms, deux bases de Victor Gay pour les départements et les naissances (Gay, 2021a, 2021b). Pour chaque colonne, nous avons ajouté tous les mots ou expression relevés dans les premiers jeux de données ajoutés (notamment les nombreuses abréviations). Enfin, j'ai fourni des jeux de données collectés à partir d'autres recherches sur Paris et sa région.

[15] Brée S. et al., *POPP. Project for the Oceration of the Paris Population censuses: elaboration of a database for historical demography.*

[16] Brée S. et al., *POPP. Project for the Oceration of the Paris Population censuses: elaboration of a database for historical demography.*

listes ne l'ont pas fait de la même manière : les départements, liens au chef de ménage et professions sont abrégés différemment, et il a donc fallu uniformiser les données.

3. Créer de nouvelles variables

La base de données comprend toutes les informations contenues dans le recensement. Plusieurs variables ont cependant été ajoutées pour faciliter l'analyse statistique.

- **Variable "absent".** Cette variable indique si l'individu est absent au moment du recensement. On sait que les personnes sont "absents" parce que leur ligne est barrée et que la mention "absent" est indiquée dans une ou plusieurs cases. Lorsque les personnes sont "absentes", les informations fournies sont en général très peu nombreuses. Habituellement, lors du dénombrement de la population, les personnes présentes au moment du recensement et vivant dans le lieu sont différenciées des autres et la population officielle distingue la population présente de l'ensemble de la population.

- **La variable sexe.** La variable sexe n'existe pas dans le recensement de la population : alors que cette information est demandée dans les bulletins individuels[17], elle n'est pas reportée dans les tableaux des listes nominatives (mais bien dans les tableaux statistiques). Nous l'avons recréé sur la base des prénoms, des situations de ménage et des professions.

- **Lieu de naissance** : Comme déjà mentionné, la colonne "lieu de naissance" a été divisée en quatre variables afin de pouvoir distinguer les villes, les départements et les pays. Une dernière variable est "autres lieux" pour mettre tous les mots qui n'ont pas été reconnus automatiquement.

- **Professions et statut professionnel**. Dans les statistiques officielles, les personnes actives sont classées en fonction de leur profession mais aussi de leur statut dans cette profession : patron, employé, ouvrier (et famille et domestiques). Lorsque cela était mentionné dans la colonne profession, nous avons extrait cette information et l'avons placée dans une variable de statut professionnel.

- **Nom de l'employeur et code professionnel**. Comme nous l'avons déjà mentionné, la dernière colonne où les rédacteurs étaient censés mettre le nom de l'employeur, mettait les codes professionnels. Cette colonne a donc été divisée en deux variables : le nom de l'employeur et le code de la profession.

- **Typologie des ménages.** Nous avons créé une typologie des ménages inspirée de adaptée à nos questions de recherche, notamment pour distinguer les couples mariés et les couples cohabitants.

[17] Les listes de recensements sont dressées par les maires des communes à partir des bulletins individuels et de ménages renseignés par les individus et qui contiennent plus d'informations que les seules reportées dans la liste nominative. Une partie de ces informations est traitées de manière agrégées pour les publications statistiques des recensements (la variable « sexe » est ainsi utilisée dans ces publications).

III. Réalisations et perspectives

1. Le projet POPP et la science ouverte

Au moment de la rédaction de ce bilan, nous avons pu travailler sur trois recensements. Nous avons choisi de nous concentrer sur les trois recensements de l'entre-deux-guerres (1926, 1931 et 1936) car leur structure était la même et parce que la période était cohérente. Nous espérons pouvoir travailler sur le recensement de 1946 par la suite.

La base de données sera, dans un premier temps, exploitée par les chercheurs et chercheuses associées au projet (voir conférence et ouvrages) puis déposée auprès de l'IR* Progedo[18] afin d'être archivées et diffusée par l'Adisp.

Ce projet vise plus globalement trois publics.

En premier lieu, la communauté de chercheurs en sciences sociales qui utilise des méthodes quantitatives pourra travailler sur une base de données déjà constituée sans le coût de sa construction pour répondre à un très grand nombre de problématiques différentes en démographie historique, histoire de la famille, histoire sociale, histoire économique, histoire des professions, histoire du genre et bien d'autres.

En deuxième lieu, les bibliothèques – et en premier lieu le Grand Équipement Documentaire du Campus Condorcet (notamment à travers François Merveille, co-porteur du projet) qui accueille les collections documentaires de l'INED – et l'ensemble de la communauté scientifique qui souhaitent aujourd'hui indexer des données tabulaires issues de publications papier pourront profiter de ces avancées techniques. Les techniques utilisées pourront donc être réutilisées dans le cadre d'autres projets d'exploitation et mises en valeur de numérisations, notamment grâce aux *data paper* publiés, ainsi que les bases annotées disponibles en ligne (voir livrables). Plusieurs groupes de recherche sont également en cours de création (voir plus loin) pour valoriser les connaissances acquises dans le cadre du projet POPP.

Enfin, le projet vise un public bien plus large puisque la base de données sera versée aux Archives de Paris afin qu'elles créent une recherche nominative pour les recensements de population (comme c'est déjà le cas pour les recrutements militaires[19]) permettant de retrouver un seul individus (sans connaître son adresse) parmi les près de 3 millions de parisiens de l'Entre-deux-guerres.

[18] L'infrastructure de recherche PROGEDO a pour but de développer la culture des données, d'impulser et structurer une politique des données d'enquêtes pour la recherche en sciences sociales.
[19] https://archives.paris.fr/s/17/etats-signaletiques-et-des-services-militaires/?

2. Livrables

2.1. Publications et communications

Dès le début du projet, nous avons ouvert un carnet Hypothèse[20] dans lequel était présenté le projet, l'équipe et les communications et publications. Sandra Brée a également communiqué régulièrement sur les avancées du projet à travers son compte Twitter[21]. Une interview « paroles de chercheurs » a également été réalisée pour le site du GIS CollEx-Persée[22].

Le projet a donnée lieu a plus d'une dizaine de communications dans des conférences diverses : tant pour présenter le projet à un public d'informaticiens spécialistes de l'océrisation qu'à des historiens ou chercheurs en sciences sociales intéressés par la question, ainsi que les premiers résultats issus de l'analyse des bases de données spécifiques aux quartiers Belleville et Chaussée d'Antin qui avaient fait l'objet des premiers traitements.

Un premier *data paper* est paru (côté informatique) et un second (côté SHS) a été présenté lors de la Conférence de l'*European Society of Historical Demography* à Madrid et sera prochainement soumis à une revue. Enfin, deux articles issues des communications sur les quartiers Belleville et Chaussée d'Antin sont en cours de soumission (voir détail ci-dessous). Les résultats de l'analyse de la base complète seront présentés lors de la conférence.

Publications

Constum, T. et al. (2022). "Recognition and Information Extraction in Historical Handwritten Tables: Toward Understanding Early 20th Century Paris Census", in Uchida, S., Barney, E., Eglin, V. (eds) *Document Analysis Systems. DAS 2022. Lecture Notes in Computer Science*, vol 13237. Springer, Cham. https://doi.org/10.1007/978-3-031-06555-2_10

À paraître/en cours

Brée S. et al., *POPP. Project for the Oceration of the Paris Population censuses: elaboration of a database for historical demography.*

Brée S. "Les célibataires dans les recensements de la population de Paris de l'Entre-deux-guerres", pour un numéro spécial sur les célibataires dans la *Revue d'Histoire Moderne et Contemporaine*

Publications en ligne

Sandra Brée et François Merville, « #FocusProjet : Popp, projet d'océrisation des recensements de la population » parisienne, site Collex-Persee, https://www.collexpersee.eu/focusprojet-popp-projet-docerisation-des-recensements-de-la-population-parisienne/

[20] https://popp.hypotheses.org/
[21] https://twitter.com/SandrBree
[22] https://www.collexpersee.eu/parole-de-chercheurs-sandra-bree/

Sandra Brée et François Merville , « Popp, projet d'océrisation des recensements de la population parisienne », dite du GED Campus Condorcet, https://gedcondorcet.hypotheses.org/2082

Sandra Brée, « Parole de chercheurs », site Collex-Persee, https://www.collexpersee.eu/parole-de-chercheurs-sandra-bree/

Sandra Brée, « Focus. POPP. Projet d'Océrisation des recensements de la Population Parisienne », La lettre du LARHRA, n°17, année 2021.

Communications

(seuls les noms des communicants sont précisés)

Thomas Constum "Reconnaissance et extraction d'informations dans des tableaux manuscrits historiques : vers une compréhension des recensements de Paris de l'Entre-deux-guerres", *Conférence "Documents anciens et reconnaissance automatique des écritures manuscrites"*, Paris, École des Chartes, juin 2022.

Sandra Brée et Thierry Paquet, « Le projet POPP : Projet d'Océrisation des recensements de la Population Parisienne », *séminaire DHAI*, 7 juin 2022.

Constum Thomas, "Recognition and information extraction in historical handwritten tables: toward understanding early 20th century Paris census", *15th IAPR International Workshop on Document Analysis System*, La Rochelle, France, 22-25 mai 2022

Sandra Brée, « The POPP Project. Project for the Oceration of the Paris Population Census (1926-1946)", *Conférence de l'European Society of Historical Demography*, Madrid, Espagne, 2-5 mars 2022.

Sandra Brée, '*Mariages à la parisienne*', Cohabiting couples in Interwar Paris", *Conférence de l'European Society of Historical Demography*, Madrid, Espagne, 2-5 mars 2022.

Sandra Brée, "Concubinage et célibat dans les recensements de la population de Paris de l'Entre-deux-guerres", Séminaire Familles : alliances, transmission, reproduction sociale, territoires. XVIIIe-XXe siècle, EHESS, Campus Condorcet, Aubervilliers.

Sandra Brée, « The POPP Project. Project for the Oceration of the Paris Population Census (1926-1946)", *IUSSP International Population Conference* (5/10 décembre 2021, Hyderabad Inde/en ligne)

Sandra Brée, "Les 'jeunes' dans les ménages parisiens de l'Entre-deux-guerres" à partir des premiers résultats issus du projet POPP », *colloque de la Société de la Démographie Historique "Entre l'enfance et l'entrée dans la vie adulte. Normes et pratiques dans les sociétés médiévales, modernes et contemporaines"*, Aubervilliers, Campus Condorcet, 4 et 5 novembre 2021.

Sandra Brée, "Les célibataires dans les recensements de la population de Paris de l'Entre-deux-guerres" à partir des premiers résultats issus du projet POPP, colloque "Le Genre des célibats", Paris, 1er octobre 2021.

Sandra Brée, « le projet POPP », *festival éPOPées du Campus Condorcet*, Aubervilliers, 30 septembre 2021.

Sandra Brée, « Le projet POPP : Projet d'Océrisation des recensements de la Population Parisienne », *colloque Humanistica*, en ligne, 10-12 mai 2021.

Sandra Brée, « The POPP Project. Project for the Oceration of the Paris Population Census (1926-1946)", *Conférence de la Population Association of America*, en ligne, 5-8 mai 2021.

2.2. Logiciels libres et annotations ouvertes

Le projet POPP a donné lieu à la réalisation de deux composants logiciels qui ont été mis en libre accès :

- Modèle optique de réseau de neurones profond reposant sur la plateforme open source pytorch: hrefhttps://github.com/FactoDeepLearning/VerticalAttentionOCR
- Module de définition de dictionnaires et d'expression régulières compilable en automates d'états finis et intégrable dans le décodage du modèle optique reposant sur les composant open-source Kaldi et Thrax : https://gitlab.com/projet-popp/sigra/

Les données annotées pour l'apprentissage du modèle optique sont également mises en libre accès en suivant ce lien :https://gitlab.com/projet-popp/sigra/

3. L'analyse historique : perspectives de recherche

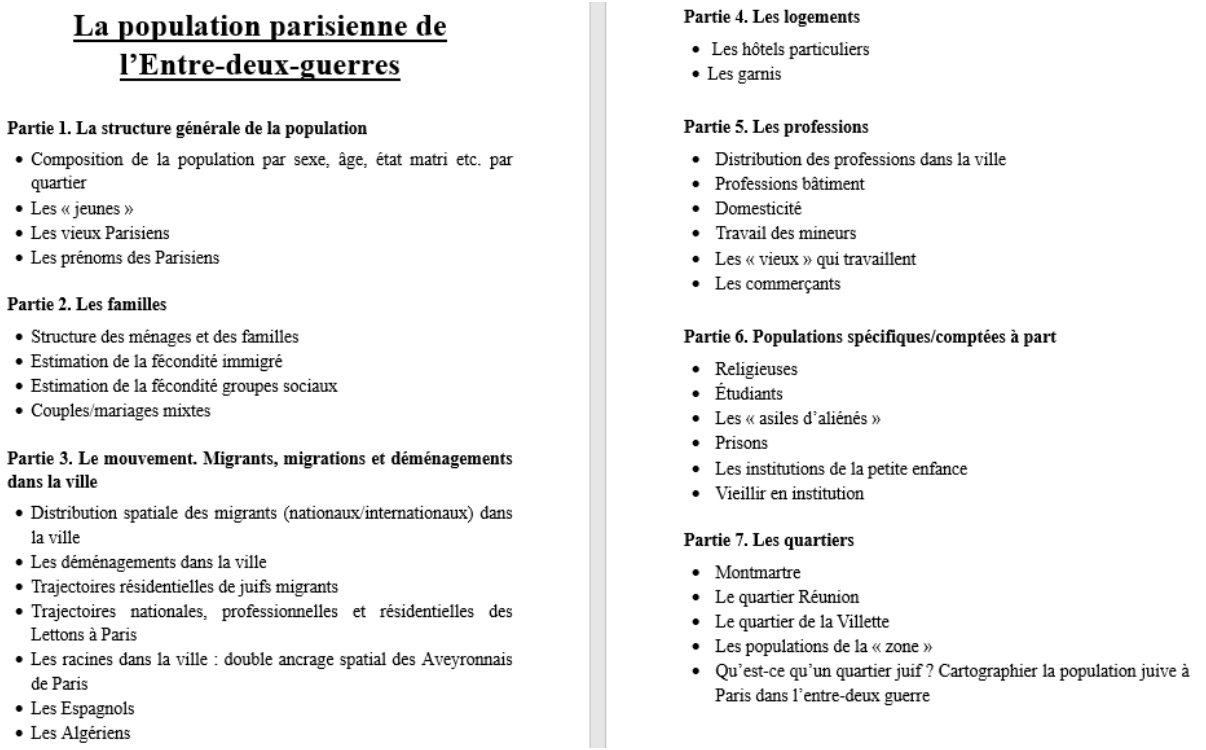
La base de données permettra une grande quantité d'analyses sur :

- La structure de la population de Paris pendant l'entre-deux-guerres (sexe, âge, statut matrimonial).
- La structure des ménages et des familles (taille des familles et des ménages, structure des ménages, nombre d'enfants, familles monoparentales, familles recomposées...)
- Mariages et partenariats (couples mariés et couples cohabitants : âge, écart d'âge, professions... ; distribution spatiale de la cohabitation, etc.)
- Natifs/migrants : part et distribution spatiale des migrants provinciaux et étrangers dans la population. Quelles sont leurs occupations professionnelles ? Avec qui vivent-ils ? Mariages mixtes de migrants de différentes origines ou avec des Parisiens, etc.
- Les professions : répartition spatiale dans la ville ; analyses spécifiques d'une profession : avec qui vivent-ils ? combien d'enfants ? ; distance entre domicile et travail
- Les populations comptées à part : analyse des population des prisons, des hospices, des asiles, des établissements de la petite enfance, des congrégations religieuses etc.

La conférence et l'ouvrage tirés de ces travaux seront particulièrement cohérents grâce à l'analyse de cette base commune, ce qui est rare en histoire.

Une conférence est prévue au cours de l'année 2024. Elle devrait réunir une quarantaine de communications sur trois jours présentées par des chercheurs et chercheuses en histoire, démographie, sociologie, économie notamment. La figure 4 présente un programme prévisionnel de la conférence.

Figure 4. Programme prévisionnel des communications de la conférence « La population parisienne de l'entre-deux-guerres ».



La population parisienne de l'Entre-deux-guerres

Partie 1. La structure générale de la population

- Composition de la population par sexe, âge, état matri etc. par quartier
- Les « jeunes »
- Les vieux Parisiens
- Les prénoms des Parisiens

Partie 2. Les familles

- Structure des ménages et des familles
- Estimation de la fécondité immigré
- Estimation de la fécondité groupes sociaux
- Couples/mariages mixtes

Partie 3. Le mouvement. Migrants, migrations et déménagements dans la ville

- Distribution spatiale des migrants (nationaux/internationaux) dans la ville
- Les déménagements dans la ville
- Trajectoires résidentielles de juifs migrants
- Trajectoires nationales, professionnelles et résidentielles des Lettons à Paris
- Les racines dans la ville : double ancrage spatial des Aveyronnais de Paris
- Les Espagnols
- Les Algériens

Partie 4. Les logements

- Les hôtels particuliers
- Les garnis

Partie 5. Les professions

- Distribution des professions dans la ville
- Professions bâtiment
- Domesticité
- Travail des mineurs
- Les « vieux » qui travaillent
- Les commerçants

Partie 6. Populations spécifiques/comptées à part

- Religieuses
- Étudiants
- Les « asiles d'aliénés »
- Prisons
- Les institutions de la petite enfance
- Vieillir en institution

Partie 7. Les quartiers

- Montmartre
- Le quartier Réunion
- Le quartier de la Villette
- Les populations de la « zone »
- Qu'est-ce qu'un quartier juif ? Cartographier la population juive à Paris dans l'entre-deux guerre

Les communications présentées lors de la conférence seront regroupées, sous la forme de chapitres, dans un ouvrage sur la population parisienne de l'Entre-deux-guerres dirigé par Sandra Brée.

Sandra Brée est, par ailleurs, en train de rédiger le mémoire inédit de son Habilitation à Diriger des Recherches qui porte sur les ménages parisiens pendant l'entre-deux-guerres et qui repose sur l'analyse de la base POPP.

IV. Bilan

1. Réussites et échecs

Tout d'abord, l'effort humain a été presque correctement estimé (2 hommes/an) :

- un ingénieur consacré à la reconnaissance d'images et à l'apprentissage automatique
- un ingénieur consacré au post-traitement et à la modélisation des connaissances

Nous avons consacré moins de temps que prévu à la reconnaissance de l'écriture manuscrite sans doute du fait que nous maitrisions déjà une solution performante de reconnaissance, mais plus de temps que prévu a été consacré à la modélisation des connaissances et du langage, afin de modéliser les abréviations et les expressions les plus utilisées sur les champs adresses et métiers (sans prétendre à l'exhaustivité).

Nous n'avons pu traiter que trois recensements mais le taux d'erreurs est l'un des plus bas observés dans la littérature. En effet, le recensement de 1946 n'a pu être traité dans le temps imparti au projet car la structure des tableaux est différente des trois autres et aurait nécessité des adaptations que nous n'avons pas réalisé faute de temps.

Les adresses n'ont pas été traitées car elles sont renseignées de manières très inégale selon les agents recenseurs, et sans suivre la structure tabulaire préétablie. Faute de temps ces informations n'ont pas été soumises à la reconnaissance automatique. Nous avons fait appel à une société spécialisée (NUMEN) pour recueillir les informations concernant les adresses (numéros et noms des rues).

De premiers tests de désambiguïsation des personnes entre les différents recensements montrent qu'une recherche multi-critères approximative apparait robuste aux erreurs de reconnaissance et aux critères non renseignés. Il sera cependant délicat de quantifier plus précisément les performances en rappel et précision en l'absence de vérité terrain qui semble impossible à établir à l'échelle des trois recensements. Il sera cependant possible de retrouver une partie des individus présents dans deux ou trois des recensements et donc de les suivre. Ce couplage des recensements, et celui prévu avec la base EXO-POPP (actes de mariage pour la période 1880-1940) permettra de travailler sur l'histoire de vie des individus et de rendre plus dynamique tant les recensements, critiqués car ne donnant des informations qu'à un instant t ; que les actes de mariage portant également sur un seul évènement. Ce couplage permettra d'observer ce qu'on ne peut voir dans les recensement et l'état civil, à savoir la cohabitation avant le (re)mariage ou après un divorce ou encore l'âge de départ des enfants du logement familial parmi bien d'autres choses.

La mise à disposition en sources ouvertes des logiciels développés pour l'occasion est propice à la réutilisation par d'autres projets similaires, notamment le traitement des recensements en France sur cette même période par le projet ANR SocFace[23].

[23] https://socface.site.ined.fr/

2. Ouverture

L'interaction entre IT et SHS a été étroite et fructueuse. L'entente a été si bonne et les résultats si encourageants que Thierry Paquet et Sandra Brée ont déposé une ANR qui a été sélectionnée en 2021 pour continuer leur travail d'océrisation de données historiques de population. Le projet **EXO-POPP**[24] porte sur la création d'une base de données à partir d'actes de mariage de Paris et de sa banlieue entre 1880 et 1940.

Au-delà, le projet POPP a beaucoup intéressé les historiens et notamment les historiens démographes. Sandra Brée est donc actuellement en train de créer un consortium francophone autour de la création de base de données historiques en collaboration avec des équipes SHS et d'informaticiens. Ce projet, SoDHIA (Sources de la démographie historique et intelligence artificielle) a obtenu un financement SEPIA à hauteur de 3000 euros. Les équipes partenaires sont : LARHRA (UMR 5190), Laboratoire LITIS (EA 4108), Centre Roland Mousnier (UMR 8596), Femto-St (UMR 6174), SAGE (UMR 7363), Universités de Louvain-la-Neuve et de Gent. Sandra Brée crée également un groupe de recherche international sur le sujet, avec des collègues belges, espagnols, hollandais, anglais et canadiens.

[24] https://exopopp.hypotheses.org/

V. Annexe : l'équipe du projet

POPP est un projet collaboratif imaginé par Sandra Brée (LARHRA, CNRS), porté avec François Merveille (GED-Campus Condorcet), grâce à la collaboration indispensable de Thierry Paquet (Université de Rouen, Litis), Thomas Constum (CNRS, LITIS) et Nicolas Kempf (CNRS, LITIS), et le financement du Collex-Persee, du Campus Condorcet et de Progedo.

Sandra Brée est historienne et démographe, chargée de recherche au CNRS (section 33) et affiliée au LARHRA. Ses recherches en démographie historique portent sur le déclin de la fécondité à Paris au XIXe siècle, la transition démographique urbaine ou encore l'évolution de la nuptialité et de la divortialité en France depuis la Révolution. Elle est également chargée des médias sociaux de la Société de Démographie Historique et des Annales de Démographie Historique et chargée de l'animation du département « données historiques, économiques et financières » de Progedo.

François Merveille est bibliothécaire au grand équipement documentaire du campus Condorcet. Depuis septembre 2019, il est responsable de la coopération pour le grand équipement. Avant de rejoindre le Campus Condorcet, il a été responsable de la bibliothèque de l'Institut des Hautes Études de l'Amérique Latine, bibliothèque de référence Européenne dans le domaine des sciences humaines sur l'aire géographique Latino-Américaine. En 2017, il a participé à une mission de coopération portant sur une opération de numérisation de l'université de Santiago de Cuba. Il est également membre du comité scientifique du projet de tourisme bleu "Odyssea Caraïbes blue grow multi-destination" cofinancé par le programme Interreg Caraïbes.

Thierry Paquet est Professeur à l'université de Rouen et Directeur du laboratoire LITIS (EA 4108). Il est spécialiste en reconnaissance de l'écriture manuscrite, en reconnaissance de formes pour données séquentielles, en analyse et reconnaissance d'images de documents et des modèles probabilistes et réseaux de neurones. Il participe actuellement à de nombreux projets dans le domaine de la lecture automatique de documents.

Thomas Constum est un ingénieur récemment diplômé de l'INSA de Rouen en spécialité informatique. Au sein du projet POPP, Thomas Constum est chargé de l'apprentissage d'un modèle de reconnaissance d'écriture ainsi que de la préparation des données pour l'apprentissage. La préparation des données consiste à localiser dans chaque image les tableaux présents et d'extraire le contenu textuel présents dans ceux-ci. Afin de faciliter la reconnaissance d'écriture, il est nécessaire de détecter pour chaque élément textuel son emplacement dans la structure logique du tableau. Thomas Constum utilise une intelligence artificielle entraînée à la détection de lignes de texte ainsi que des méthodes de traitement d'image classique. Une fois les données préparées, il procèdera à l'entraînement du modèle de reconnaissance d'écriture sur ces données.

Nicolas Kempf est un ingénieur en informatique récemment diplômé de l'INSA de Rouen Normandie. Au sein du projet POPP, il travaille en coordination avec Thomas Constum, pour réaliser l'ensemble de la chaîne de traitement nécessaire au projet. Il s'occupe principalement de la mise en place du module de reconnaissance de texte et de ce qui l'entoure. En sortie de la segmentation des images effectuées par Thomas Constum, il développe un module de formatage des images afin que ces dernières puissent être annotées. Une fois les images annotées et apprises par le modèle, il procédera à la correction de la sortie du module de reconnaissance, à l'aide d'un ensemble de dictionnaires et de règles.

Pierrick Tranouez est Ingénieur de Recherche à l'université de Rouen Normandie, dans le laboratoire LITIS. Un de ses domaines de recherche est la valorisation numérique du patrimoine documentaire, qu'il a enseigné dans le Master Histoire Civilisation Patrimoine, Valorisation du Patrimoine. Il s'intéresse plus particulièrement à l'application de méthodes d'apprentissage artificiel appliquées à l'analyse d'images de documents (layout analysis, structural analysis, HTR). Il anime notamment les projets DocExplore et PIVAJ. Pour en savoir plus.

Clément Chatelain est maître de conférences à l'INSA de Rouen. Il effectue ses recherches au LITIS (EA 4108) dans le domaine du machine learning et des réseaux de neurones, avec des applications aux images de documents et à l'imagerie médicale. Depuis 2021 il est directeur du labcom ANR LLisa.