ObHisPOP - Observatoire de l'histoire de la population française : grandes bases de données et intelligence artificielle

Historienne démographe, chargée de recherche CNRS au Laboratoire de recherche historique Rhône-Alpes (LARHRA, UMR 5190, CNRS / ENS de Lyon / Université Grenoble Alpes / Université Lumière Lyon 2 / Université Jean Moulin Lyon 3), Sandra Brée s'intéresse à l'histoire démographique des populations, principalement urbaines et banlieusardes, pendant la période 1880-1940. Elle porte le SOSI Observatoire de l'histoire de la population française : grandes bases de données et intelligence artificielle.

L'un des outils essentiels de l'histoire quantitative, et notamment de la démographie historique, est la constitution de grandes bases de données. Celle-ci implique la gestion d'une masse de données souvent difficile à maîtriser, à l'échelle des équipes de recherche, par des méthodes classiques de collecte. La démographie historique française a toujours été ouverte à l'innovation et a longtemps été pionnière dans le monde. Or, les progrès très récents de l'intelligence artificielle et du *Deep Learning* (apprentissage profond) permettent à présent d'envisager la collecte d'informations de manière automatique. L'intérêt majeur de ces techniques pour la création de bases de données en démographie historique, outre le gain de temps évident, est d'envisager de travailler sur des populations plus vastes, et notamment sur les populations des grandes villes, longtemps laissées de côté, notamment pour les XIXe et XXe siècles.

Le premier projet déposé en France pour créer une base de données de démographie historique grâce aux nouveaux progrès du *Deep Learning* et de l'océrisation[1] est le projet POPP (Projet d'océrisation des recensements de la population parisienne)[2]. L'ambition du projet POPP, qui a débuté en septembre 2020 grâce à un co-financement du CollEx-Persée, de l'IR* Progedo et de l'Humathèque Condorcet, était de créer une base de données permettant l'exploitation statistique des listes nominatives de la population de Paris de 1926, 1931 et 1936[3] (Figure 1). La population de Paris comptait alors près de trois millions d'habitants et chaque recensement comprend environ 50 000 images contenant généralement deux doubles pages, soit 60 individus au maximum. Sur la base d'un rythme — assez élevé – de quarante-cinq secondes à une minute pour relever les informations concernant chaque individu, créer la base de données à la main prendrait entre 652 500 et 870 000 minutes, soit de 18 600 à 24 900 semaines de 35 heures, c'est-à-dire environ 500 ans pour une personne seule ou un siècle pour une équipe de cinq personnes. Ces données suffisent à montrer le caractère irréaliste d'une telle démarche dans le cadre d'un projet scientifique historique.

Plusieurs possibilités s'offrent aux chercheurs et chercheuses lorsqu'ils et elles veulent travailler sur des populations nombreuses. La première est de ne relever qu'une partie des individus, en effectuant des sondages. Cette technique est intéressante et a engendré des recherches particulièrement fructueuses, mais elle ne permet pas, par exemple, de travailler sur des sous-ensembles de populations, contrairement aux relevés exhaustifs.

Plus récemment, les progrès de l'intelligence artificielle ont été mobilisés pour la recherche en sciences sociales, et notamment en histoire. L'océrisation des documents, c'est-à-dire la reconnaissance optique des caractères, est largement utilisée pour reconnaître et rechercher des mots dans des ensembles imprimés ou, plus récemment, manuscrits. Les bases de données quantitatives nécessitent, en plus de la reconnaissance optique des caractères, la reconnaissance des systèmes de tableurs des sources présentées ainsi (comme les listes nominatives des recensements de la population par exemple, ou les données statistiques publiées) ou encore l'utilisation de *Named Entity Recognition*, qui est un système permettant d'extraire de l'information dans des textes en reconnaissant des entités nommées et de leur attribuer des étiquettes telles que « nom », « lieu », « profession ». La

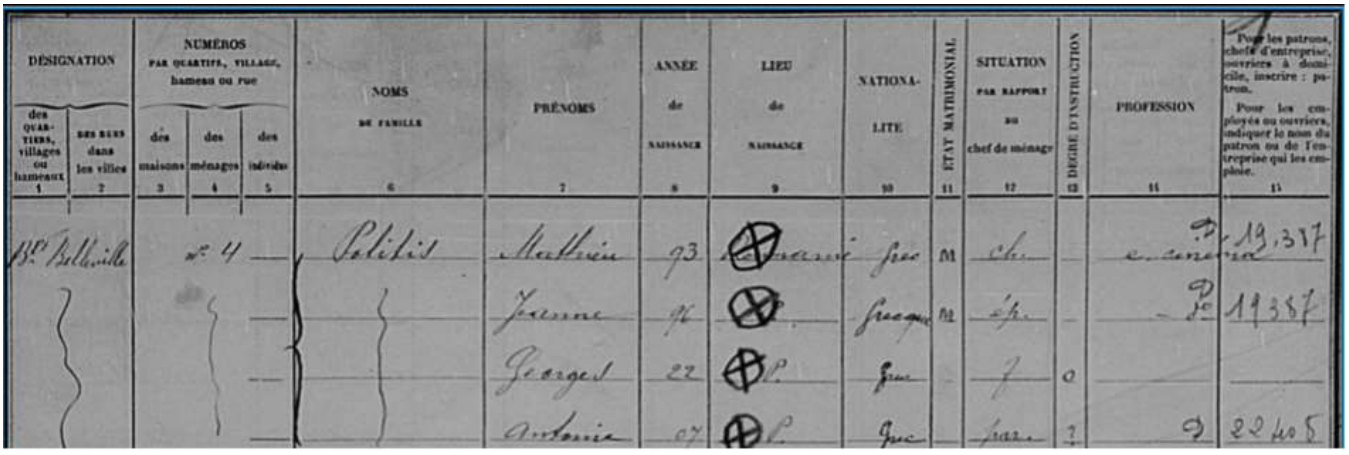

Figure 1. Extrait de la liste nominale du recensement de la population de 1926, population de résidence habituelle, arrondissement de Belleville. Cote D2M8 307 (Archives de Paris)

1. L'océrisation désigne le recours à l'Optical character recognition (OCR), connu en français sous le nom de Reconnaissance optique des caractères (ROC).
2. Responsable : Sandra Brée (CNRS, Laboratoire de recherche historique Rhône-Alpes - LARHRA), avec la collaboration de François Merveille (Humathèque Condorcet) et Thierry Paquet (Université de Rouen, Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes - LITIS).
3. Seules quatre listes nominatives de recensements ont été dressées pour Paris : en 1926, 1931, 1936 et 1946. Pour des raisons de coûts, mais aussi parce que les tableaux des listes de recensement de 1946 ne sont pas structurés de la même manière que les précédents, le recensement de 1946 n'a pas été traité dans le cadre du projet POPP.

Figure 2. Acte de mariage de Paris de 1910 (14e arrondissement) et les annotations correspondantes des entités nommés dans PIVAN (logiciel créé par le LITIS). Source : Constum T., Preel L., Larcher T., Tranouez P., Paquet T., Brée S. 2024, « End-to-end information extraction in handwritten documents: understanding Paris marriage records from 1880 to 1940 », *IDCAR*.

reconnaissance des entités nommées est très intéressante pour la constitution de bases de données en démographie historique car elle permet de lire et de reconnaître des mots et de les attribuer « directement » à des colonnes (variables). Cette technique est ainsi particulièrement appropriée pour travailler sur des actes d'état civil, qui sont l'une des sources principales de données utilisées en démographie historique (avec les recensements de population). Elle est ainsi à la base du projet Exo-POPP dont le but est de constituer une très large base de données à partir de l'ensemble des actes de mariage de Paris et de sa banlieue entre 1880 et 1940 (Figure 2).

C'est pour partager les connaissances acquises pendant les projets POPP et Exo-POPP que l'idée d'un consortium réunissant des équipes souhaitant élaborer des bases de démographie historique grâce à l'Intelligence Artificielle est née. L'ambition de ce consortium était donc de partager les techniques informatiques et de discuter des difficultés techniques liées à la création de très grandes bases de données avec des équipes composées, à la fois, de chercheurs et chercheuses en sciences humaines et sociales et en informatique. Ce consortium a été baptisé SoDHIA (Sources de Démographie Historique et Intelligence Artificielle) et a bénéficié d'un premier financement de CNRS Sciences humaines & sociales (appel à projets Sepia). Le consortium a été inauguré lors des journées d'étude de lancement du projet en janvier 2023. À la suite de ces premières journées, les six équipes impliquées[4] ont souhaité constituer un réseau autour de la création de bases de démographie historique grâce à l'Intelligence Artificielle. Le SOSI ObHisPop a ainsi été fondé en juin 2023 sur les bases de ce premier consortium. La première journée annuelle a eu lieu en novembre 2023 au Campus Condorcet (Figure 3).

Affiche de la Journée annuelle du SOSI ObHisPop (Campus Condorcet, novembre 2023)

L'objet du SOSI ObHisPop - Observatoire de l'histoire de la population française : Grandes bases de données et intelligence artificielle est d'aider à la construction, la finalisation, la conservation, la diffusion et l'exploitation d'enquêtes historiques sur la population française de manière pérenne. Il a quatre objectifs :

4. Équipe des projets POPP - Projet d'océrisation des recensements de la population parisienne (1926-1946) et EXO-POPP - Extraction optique des entités nommées manuscrites pour les actes de mariage de la population de Paris (1880-1940) : Laboratoire de Recherche Historique Rhône Alpes (LARHRA, UMR 5190, CNRS / ENS de Lyon / Université Grenoble Alpes / Université Lumière Lyon 2 / Université Jean Moulin Lyon 3) ; Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes - LITIS (EA4108) / Normastic
Équipe de l'enquête sur la population de Charleville du XVIIe siècle à la fin du XIXe siècle : Centre Roland Mousnier (CRM, UMR8596, CNRS / Sorbonne Université)
Équipe POPSTRAS - Océrisation du fichier domiciliaire pour une histoire démographique de Strasbourg (1871-1939) : laboratoire Sociétés, acteurs, gouvernement en Europe (SAGE, UMR7363, CNRS / Université de Strasbourg) ; laboratoire Identité et Différenciation de l'Espace, de l'Environnement et des Sociétés (IDEES, UMR6266, CNRS / Université de Caen Normandie / Université Le Havre Normandie / Université de Rouen Normandie)
Équipe de l'enquête sur la population de Belfort : Institut Franche-Comté Électronique Mécanique Thermique et Optique – Sciences et Technologies (FEMTO-ST, UMR6174, CNRS / SUPMICROTECH-ENSMM / Université de Franche-Comté / Université de technologie de Belfort-Montbéliard) ; laboratoire Connaissances et Intelligence Artificielle Distribuées (CIAD)
Équipe de l'enquête GENPAR : Centre Roland Mousnier
Équipe des projets « Ineqkil » et « EPIBEL » : Université catholique de Louvain et Université de Gand

1. Le premier est d'aider des projets ayant déjà obtenu des financements pour construire, grâce à l'intelligence artificielle, des bases de données de démographie historique à perdurer dans le temps. Souvent, les projets financés permettent de créer des bases de données, mais les équipes ont rarement le temps de les exploiter pleinement ou de les entretenir au-delà du temps du projet financé.
2. Le deuxième objectif est de permettre le transfert de technologies vers deux types d'enquêtes : d'une part, des enquêtes qui visent à créer une grande base de données de démographie historique grâce à l'intelligence artificielle ; d'autre part, des enquêtes lancées dans les décennies 2000 et 2010, voire auparavant, selon la méthode classique de la saisie manuelle des données, de manière à ne pas perdre le travail considérable déjà accompli.
3. Le troisième objectif, en collaboration avec l'IR* Progedo, est de documenter, conserver et diffuser les enquêtes de l'Observatoire.
4. Enfin, le dernier objectif est de créer une dynamique scientifique autour des bases de données en démographie historique en mettant à disposition des équipes du SOSI des outils et des connaissances pour en créer de nouvelles. À terme, il s'agira en outre, grâce à cet élan, d'envisager la mise en relation de certaines de ces bases, selon une perspective d'interopérabilité.

Les bases créées dans le cadre de l'ObHisPop seront donc déposées et diffusées *via* l'ADISP - Archives de données issues de la statistique publique (Quetelet-Progedo-Diffusion). Si elles sont indispensables à l'analyse en démographie historique, elles pourront également être des ressources particulièrement riches pour l'ensemble des chercheurs et chercheuses en sciences sociales. Ces bases seront également diffusées au-delà du monde académique. Ainsi, la base POPP sera versée aux Archives de Paris afin de permettre une interrogation nominative dans les 150 000 images constituant les recensements de 1926 à 1936, offrant la possibilité de retrouver des individus, sans connaître leur adresse, dans l'immense botte de foin que forme la population parisienne pendant l'entre-deux-guerres.

Le soutien de CNRS Sciences humaines & sociales a permis aux équipes du SOSI de se structurer. En plus des réunions en visioconférences pluriannuelles, les équipes de l'ObHisPop se réunissent au moins une fois dans l'année. Cette journée d'étude permet de faire le point sur les différents projets du SOSI, et notamment sur les progrès des techniques informatiques pour créer les bases de données. Les équipes informatiques sont en constante communication et l'équipe du LITIS, précurseur en la matière, partage les outils mis en place pour les premiers projets POPP et Exo-POPP. Les équipes en sciences humaines et sociales échangent, quant à elles, sur les exploitations ultérieures à la création des bases par les informaticiens, ainsi que sur leurs difficultés, et l'ingénieure de recherche du SOSI, accueillie au sein de l'IR* Progedo, partenaire du SOSI, aide les équipes des différents projets à corriger et adapter les bases de données à l'analyse statistique.

contact&info
▶ Sandra Brée,
LARHRA
sandra.bree@cnrs.fr