Reconstituer l'histoire de la France « par le bas » : un siècle de recensements de population décryptés par le projet Socface

Socface est un projet de recherche, soutenu par l'Agence nationale de la recherche (ANR), sur les recensements de la population française de 1836 à 1936. Il mobilise des chercheurs en sciences humaines et sociales, des ingénieurs et des archivistes, et illustre de nombreux aspects de la science ouverte ainsi que les apports et défis de la reconnaissance automatique d'écriture manuscrite[1].

CHRISTOPHER KERMORVANT
Président de la société Teklia

LIONEL KESZTENBAUM
Directeur de recherche à l'Institut national d'études démographiques (INED)

MANONMANI RESTIF
Cheffe de projet du portail FranceArchives, Service interministériel des Archives de France (SIAF)

Le projet Socface[2] a pour ambition de transcrire automatiquement l'ensemble des listes nominatives des recensements de 1836 à 1936 (soit vingt recensements) pour produire, étudier et diffuser une base de données des individus ayant vécu en France durant cette période. Soutenu par l'Agence nationale de la recherche (ANR)[3], ce projet illustre de nombreux aspects de la science ouverte ainsi que les apports et défis de la reconnaissance automatique d'écriture manuscrite.

Il met aussi en évidence l'appétit sans cesse croissant des différents utilisateurs des archives pour les données nominatives : aujourd'hui, l'écrasante majorité des recherches faites dans les services d'archives porte sur ce type de sources. Chaque personne a vocation à être représentée aux archives, dès lors que sa vie a connu quelques événements, heureux ou, le plus souvent, malheureux.

Socface mérite une place à part en raison de son ampleur : il porte en effet sur un corpus très vaste – une même typologie, traitée sur 100 ans, conservée dans près de 100 structures de métropole et d'outre-mer.

Aux origines du projet

L'intérêt croissant pour les données individuelles, en particulier nominatives, est alimenté par les développements techniques (facilité de numérisation, diffusion des images sur le Web, améliorations des techniques de reconnaissance automatique d'écritures, etc.) tout autant qu'il les nourrit : la demande des usagers (chercheurs, généalogistes ou amateurs éclairés) motive les campagnes de numérisation tout comme l'appétence de la recherche quantitative en sciences sociales pour des données « micro » stimule le développement de la reconnaissance automatique d'écriture manuscrite.

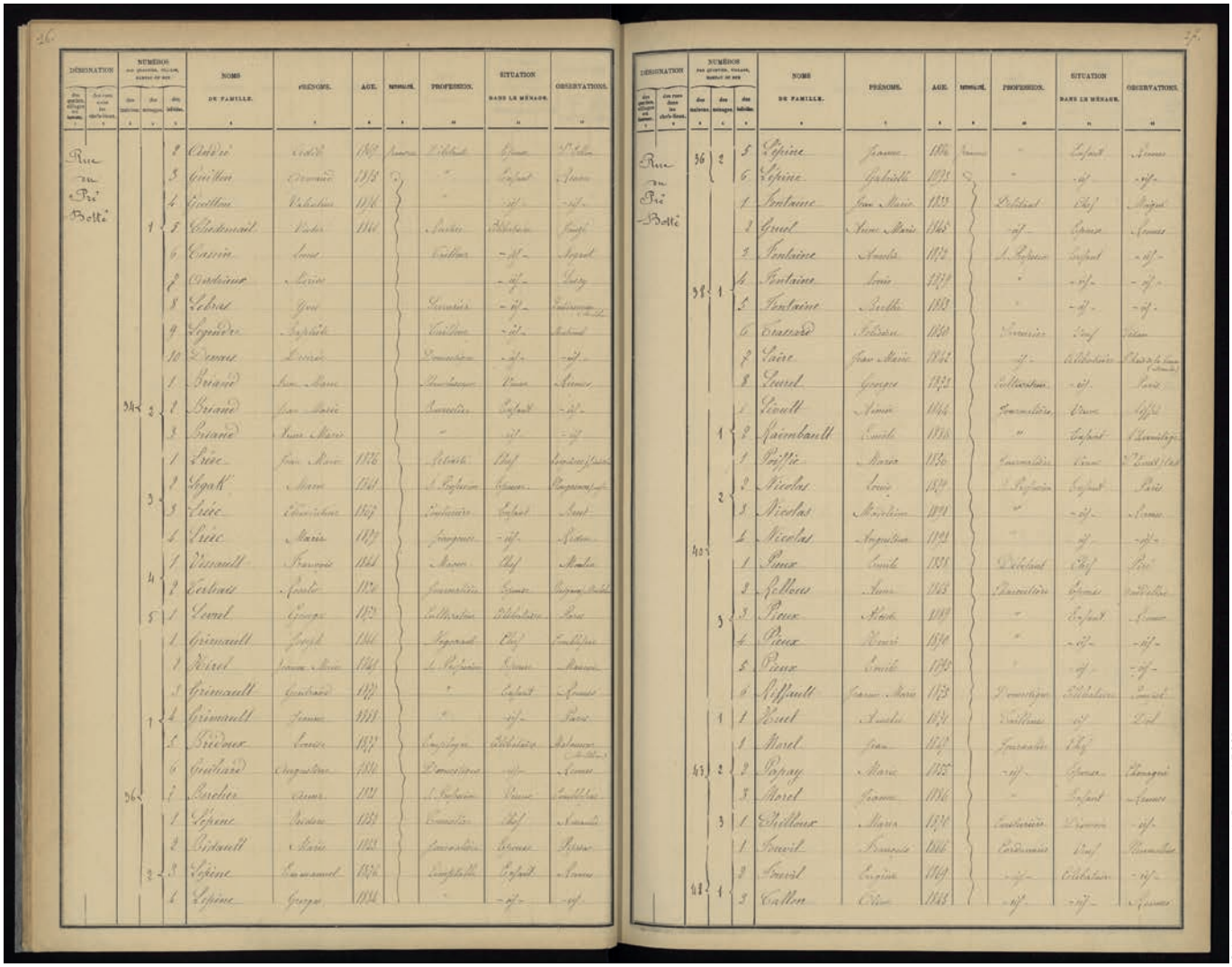
Socface illustre parfaitement ce cercle vertueux autour d'une source unique (les recensements) qui fait partie des rares typologies de documents à avoir été presque intégralement numérisées par les services d'archives, créant un corpus qui devrait dépasser à terme les 10 millions d'images malgré les destructions, volontaires ou accidentelles. Cette numérisation quasi exhaustive était une condition préalable pour qu'un tel projet de recherche puisse être réalisé.

Cette condition remplie, la gourmandise des historiens pour cette masse de données ne suffisait pas ; encore fallait-il imaginer un système efficace pour extraire le texte contenu dans ces millions d'images. Les progrès considérables de la reconnaissance automatique de l'écriture manuscrite ces dernières années, grâce aux avancées des technologies de l'intelligence artificielle, permettent d'envisager cette extraction. Les documents historiques manuscrits, du Moyen Âge à nos jours, sont désormais à la portée d'une transcription automatique permettant une exploitation directe. Cette reconnaissance automatique prend tout son sens pour des traitements à très large échelle pour lesquels une transcription manuelle, même collaborative, n'est pas envisageable.

Le rôle du collaboratif dans la reconnaissance d'écriture

Pour autant, la reconnaissance d'écriture n'est pas un vase clos, entièrement autonome. En effet, le développement d'un système de reconnaissance d'écriture

1. En anglais : *Handwritten Text Recognition* (HTR).
2. **http://www.socface.org/**
3. **https://anr.fr/Projet-ANR-21-CE38-0013**



Recensement du canton sud-est de Rennes : ville et banlieue, 1896.

performant nécessite une phase d'entraînement des modèles sur des données annotées, par des techniques d'apprentissage automatique supervisé. Les modèles les plus récents, fondés sur des technologies de *Deep Learning*[4], peuvent être entraînés avec un protocole beaucoup plus simple que leurs prédécesseurs. Aujourd'hui, il n'est plus nécessaire de transcrire précisément les documents, en indiquant la position et le contenu des lignes de texte. Il est possible d'entraîner les modèles à partir de données saisies dans un formulaire, comme on le ferait pour un dépouillement d'archives. Ce protocole, beaucoup plus rapide et naturel, permet de faire appel à des volontaires pour réaliser les annotations.

Le projet Socface a ainsi ouvert une dizaine de campagnes d'annotations collaboratives pour créer des données d'entraînement en utilisant la plateforme Callico de Teklia[5].

Par ailleurs, les annotations déjà existantes, réalisées par les cercles généalogiques ou dans les services d'archives départementales, peuvent aussi être utilisées pour entraîner la machine. De fait, la qualité de la reconnaissance est améliorée par tout un ensemble d'informations extérieures : de la liste des noms de famille (et de leur fréquence) jusqu'au nom des lieux-dits de chaque commune, en passant par une estimation grossière des distributions par âge au cours du temps, tout ce qui peut donner à la machine une idée, même vague, de « l'univers des possibles » est précieux.

En ce sens, Socface est très directement un produit de la science ouverte.

Traiter, analyser et diffuser des millions d'images

Le siècle d'histoire française auquel s'intéresse Socface est marqué par des changements spectaculaires souvent résumés par quelques concepts généraux esquissés à grands traits : urbanisation, industrialisation, transition démographique. Pourtant, on connaît encore relativement mal la variation spatiale de ces phénomènes sur le territoire métropolitain, leurs mécanismes et leurs conséquences. L'apport de Socface, en particulier en appariant les individus entre les recensements pour reconstituer leurs trajectoires (migratoires, professionnelles, familiales), est de permettre d'étudier cette hétérogénéité, de saisir comment ces trajectoires rencontrent, ou pas, la « Grande Histoire », comment elles sont influencées par elle et l'influencent en retour.

Un second produit direct du projet sera de diffuser librement ces données pour permettre à tout un chacun d'y accéder. Pour les archives, cette mise à disposition d'un grand volume de données, tant dans

4. Connu en français sous le nom « apprentissage profond », le *Deep Learning* est une branche de l'intelligence artificielle. L'apprentissage profond est un procédé d'apprentissage automatique utilisant de très grands réseaux de neurones, possédant des centaines de couches et des milliards de paramètres (neurones). Cette technique a permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale, de la vision par ordinateur et du traitement automatisé du langage.

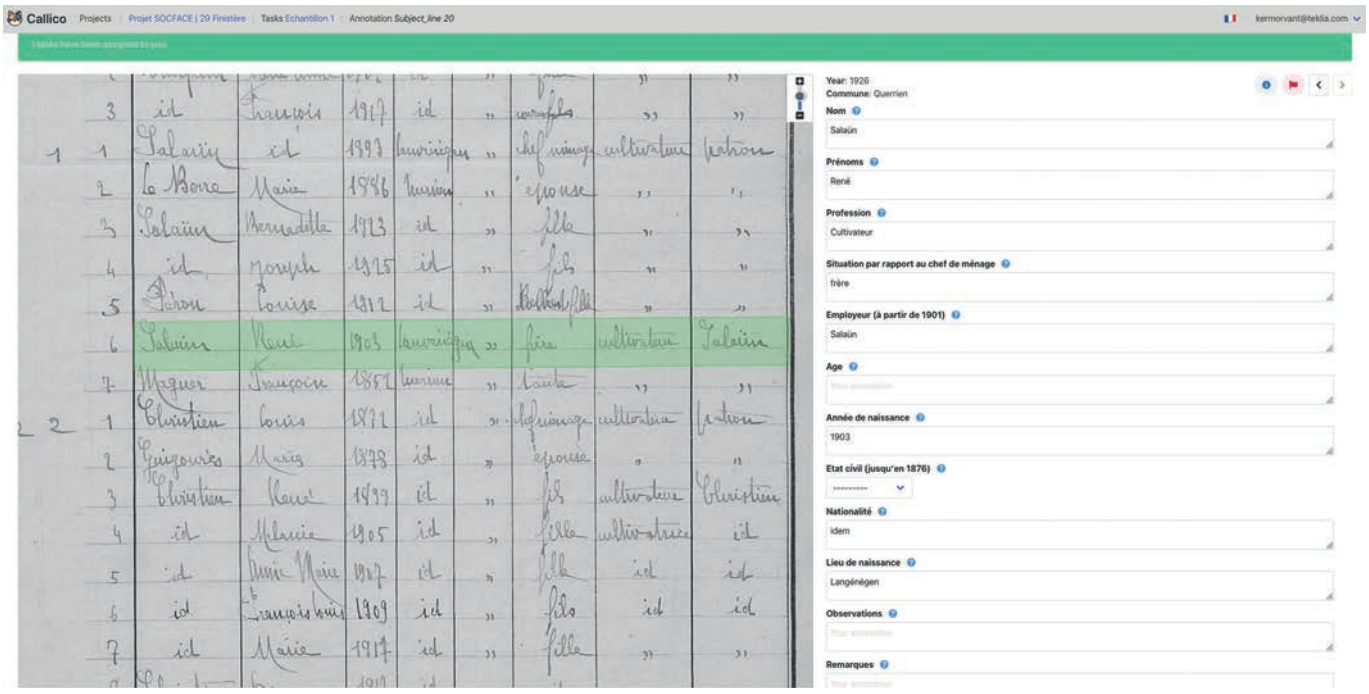
5. **https://callico.teklia.com**

Transcription automatique d'une page de recensement.

la base de noms de FranceArchives[6] que sur les sites Web des services d'archives, représente une formidable opportunité de développer de nouveaux services pour leurs publics attachés à la micro-histoire individuelle. Elle ouvre aussi des perspectives de mutualisation du réseau des archives pour augmenter le stock des métadonnées archivistiques interopérables.

À terme, Socface représentera un prodigieux effet levier. D'un côté, il poussera, inévitablement, à la numérisation des recensements manquants, voire à leur identification. De l'autre, il pourra constituer un socle sur lequel mettre en œuvre d'autres dépouillements de sources à grande échelle. Plus largement, il devrait favoriser la concertation entre les archivistes et le monde de la recherche, les premiers pouvant réinterroger leurs politiques de numérisation, par exemple en développant une dimension nationale autour de typologies d'envergure, tandis que le second devra être plus attentif à reverser aux services d'archives les données qu'il produit. ■

6. **https://francearchives.gouv.fr/fr/basedenoms**



Annotation d'une page de recensement à l'aide de Callico.