LOGO
ETABLISSEMENT

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

école doctorale **sciences pour l'ingénieur et microtechniques** SPIM

Titre de la thèse/Thesis title : Approches de Machine Learning appliquées à l'analyse de documents d'état civil/ Machine Learning Approaches applied to civil status document

Laboratoire d'accueil / Host Laboratory : UTBM-CIAD (Connaissances et Intelligence Artificielle Distribuées)

Spécialité du doctorat préparé/Speciality : Informatique/Computer Science

Mots-clefs / Keywords : deep learning, annotation semi-automatique, démographie historique, état-civil

Descriptif détaillé de la thèse / Job description

French version

Depuis une dizaine d'années, le champ des Digital Humanities s'est considérablement développé en France. Pour les historien.nes, cela signifie à la fois l'écriture d'une histoire numérique, à partir d'un corpus d'archives numérisées et la numérisation du métier d'historien.ne, qui doit maîtriser de nouveaux outils, notamment statistiques. Pour les informaticien.ne.s, cela signifie la mise à disposition d'un champ d'application presqu'inépuisable pour le Deep Learning en reconnaissance de l'écriture manuscrite.
Plus spécifiquement, les partenaires du projet font partie du SOSI (Suivi Ouvert des Sciences et de leurs interactions) interdisciplinaire (SHS-SPI) «Observatoire de l'histoire de la population française» labellisé CNRS et coordonné par l'IR* PROGEDO. Cet Observatoire, composé de 6 laboratoires français SHS et SPI se focalise sur les apports du deep learning à la recherche en démographique historique.

Au moyen d'un corpus circonscrit, *i.e.* l'état civil des naissances -désormais ECN- et des décès -ci-après ECD- pour les années 1824 à 1919, ce projet vise, dans un premier temps, à analyser les naissances hors-mariage dans un contexte de développement urbain important et rapide qui est souvent décrit comme favorisant l'émergence ou l'accentuation de l'illégitimité. Les données individuelles fournies par ces actes de naissance permettront de distinguer les naissances des «filles-mères», pour lesquelles aucune présence paternelle ne peut être décelée dans la reconnaissance ou la légitimation de l'enfant, des naissances des couples non mariés (concubins). Ces indications de reconnaissances et de légitimations sont reportées dans les marges des actes de naissance qui constituent une grande richesse : mention également des mariages (depuis 1897), des divorces (depuis 1886) et des décès (depuis 1945), qui permettent d'envisager de multiples directions de recherche en démographie historique. Pour l'heure, l'échelle individuelle des données permettra de croiser les professions ou l'âge des parents des enfants pour affiner la connaissance sur l'illégitimité (Brée 2014 et 2017a). Par ailleurs, l'analyse par méthode de deep learning de ces documents devrait permettra d'autres champs d'investigation tels que : l'analyse de l'évolution des accouchements à et hors domicile, l'analyse des causes de la baisse de la mortalité infantile et l'augmentation de la longévité, …

Ce projet peut être réalisé grâce aux progrès accomplis dans l'apprentissage profond, notamment dans le cadre de la reconnaissance de textes manuscrits. La majorité des articles publiés sur ce sujet commencent par réaliser la segmentation des lignes (Chung 2019 ; Plateau-Holleville, Bonnot et alii 2021), par l'une des trois approches suivantes (Likforman-Sulem 2007) : des méthodes basées sur une projection, qui considèrent les frontières entre les lignes comme des vallées du profil de projection verticale (Papavassiliou 2010). Deuxièmement, les méthodes de regroupement ; elles consistent à regrouper les rangées de composants connectés selon des règles heuristiques (Feldbach 2001). La dernière catégorie est celle des méthodes de brouillage qui utilisent des filtres de flou combinés à la binarisation ou aux contours actifs par exemple (Swaileh 2015). Elle est maintenant généralement traitée par un réseau entièrement convolutif (FCN) comme dans (Grüning 2019).

Ce sujet de thèse s'inscrit dans le contexte de ce projet. Les objectifs sont alors les suivants :

- Adapter les méthodes d'OCR à la spécificité des états civils de naissance

- Développer de nouveaux outils capables de lire les couches successives de texte des annotations marginal.
- Créer de nouvelle base de données d'apprentissage pour l'OCR
- Développer des algorithmes d'extraction d'information à partir de la base de données construite pour des études de démographie historique.

Le doctorant recruté, de par le positionnement interdisciplinaire SPI-SHS qu'il aura acquis, bénéficiera d'un savoir-faire et d'un savoir-être lui permettant de participer aux différentes manifestations nationales et régionales du type "fête de la science" mais aussi "journées du patrimoine", à l'interface entre SPI et SHS, notamment en collaboration avec les services d'archives (archives municipales/archives départementales) et pourra aussi intervenir à destination des publics de collégiens-lycéens (services éducatifs, cordées de la réussite, etc.): démonstration et initiation à l'outil conçu.

English version

Over the past decade, the field of Digital Humanities has significantly developed in France. For historians, this entails both the creation of digital history from a corpus of digitized archives and the digital transformation of the historian's profession, requiring mastery of new tools, particularly in statistics. For computer scientists, it provides an almost inexhaustible field of application for Deep Learning in handwriting recognition.
Specifically, the project partners are part of the interdisciplinary "Open Monitoring of Sciences and its interactions" (in French SOSI) (Social sciences-STEM) "Observatory of French Population History" labeled CNRS, coordinated by the IR* (research infrastructure) PROGEDO. This Observatory, composed of 6 French research centers, focuses on the contributions of deep learning to research in historical demography.
Using a delimited corpus, namely civil registration records for births (referred to as ECN) and deaths (referred to as ECD) from 1824 to 1919, this project aims, initially, to analyze non-marital births in a context of significant and rapid urban development, often described as favoring the emergence or accentuation of illegitimacy. The individual data provided by these birth records will allow for the distinction between births of "unwed mothers," where no paternal presence can be detected in the acknowledgment or legitimation of the child, and births of unmarried couples (cohabiting partners). These indications of acknowledgment and legitimation are recorded in the margins of the birth records, which constitute a great wealth, including mentions of marriages (since 1897), divorces (since 1886), and deaths (since 1945), opening up multiple directions for research in historical demography. Currently, at the individual level, the data will allow for cross-referencing of parents' professions or age to refine knowledge about illegitimacy (Brée 2014 and 2017a). Furthermore, deep learning analysis of these documents should enable other fields of investigation, such as analyzing the evolution of home and non-home births, the causes of the decrease in infant mortality, and the increase in longevity.
This project can be realized thanks to the progress made in deep learning, particularly in the field of handwritten text recognition. The majority of articles published on this topic start by performing line segmentation (Chung 2019; Plateau-Holleville, Bonnot et alii 2021), using one of the following three approaches (Likforman-Sulem 2007): projection-based methods that consider the boundaries between lines as valleys in the vertical projection profile (Papavassiliou 2010), clustering methods that group rows of connected components according to heuristic rules (Feldbach 2001), and blurring methods that use blur filters combined with binarization or active contours, for example (Swaileh 2015). This is now commonly addressed by a fully convolutional network (FCN), as in (Grüning 2019).
This thesis topic is situated within the context of this project. The objectives are as follows:
• Adapt OCR methods to the specific characteristics of birth records.
• Develop new tools capable of reading the successive layers of text in marginal annotations.
• Create new training databases for OCR.
• Develop information extraction algorithms from the constructed database for historical demography studies.
The recruited Ph.D. student, through their interdisciplinary position in SPI-SHS, will acquire expertise and skills that will allow them to participate in various national and regional events such as "Fête de la Science" and "Journées du Patrimoine" acting as an interface between SPI and SHS, particularly in collaboration with archival services (municipal archives/departmental archives). They will also be able to engage with secondary school students (educational services, success initiatives, etc.) by providing demonstrations and introductions to the developed tool.

Références bibliographiques / Bibliography

Bluche T., Ney H. and Kermorvant C., "Tandem HMM with convolutional neural network for handwritten word recognition," in International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2013, pp. 2390–2394.

Bonneuil Noël, *Transformation of the French Demographic Landscape 1806-1906*, Oxford, Clarendon Press, 1997.

Brée Sandra, « Incidence de la fécondité illégitime sur la fécondité générale à Paris au XIX^{e} siècle », dans *Espace, Population, Société*, 2014/1, 2014.

Brée Sandra, *Paris l'infécondé, La limitation des naissances en région parisienne au XIX^{e} siècle*, Paris, INED, coll. « Études et enquêtes historiques », 2017a.

Brée Sandra, « Les accouchements hors domicile à Paris au 19^{e} siècle », XVI^{e} colloque de la CUDEP, Aix-en-Provence (mai 2013), 2017b, pp. 265-279.

Charrier Philippe, Clavandier Gaëlle, Gourdon Vincent *et alii* (dir.), *Morts avant de naître. La mort périnatale*, Tours, Presses universitaires Francois Rabelais, coll. « Perspectives historiques », 2018.

Chung J., Delteil T., "A computationally efficient pipeline approach to full page offline handwritten text recognition," in Workshop on Machine Learning, WML@ICDAR, 2019, pp. 35–40.

Coquenet Denis, Chatelain Clément, Paquet Thierry, *End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network*, à paraître dans IEEE-PAMI, 2022. doi.org/10.1109/TPAMI.2022.3144899, pre-print https://arxiv.org/abs/2012.03868.

Constum Thomas, Kempf Nicolas, Paquet Thierry *et alii*, « Recognition and information extraction in historical handwritten tables: toward understanding early 20th century Paris census », à paraître dans proceeding of IAPR - Document Analysis Systems (DAS), mai 2022 (accepté).

Feldbach M., Tönnies K. D. , "Line detection and segmentation in historical church registers," in 6th International Conference on Document Analysis and Recognition, ICDAR, 2001, pp. 743–747.

Frinken V., Peter T., Fischer A. *et alii*, "Improved handwriting recognition by combining two forms of hidden markov models and a recurrent neural network," in Computer Analysis of Images and Patterns, CAIP, ser. Lecture Notes in Computer Science, vol. 5702, 2009, pp. 189–196.

Grüning T., Leifert G., Strauß T. *et alii* "A two-stage method for text line detection in historical documents," International Journal on Document Analysis and Recognition, *IJDAR*, vol. 22, no. 3, pp. 285–302, 2019.

Heyberger L., « L'industrialisation de Belfort : une conséquence positive du siège de 1870-1871 ? Approche par l'histoire anthropométrique », dans Robert Belot (dir.) *1870 de la guerre à la paix, Strasbourg-Belfort*, Paris, Hermann, 2013, p. 207-217.

Heyberger L., “Received wisdom versus reality: height, nutrition, and urbanization in mid-nineteenth-century France”, *Cliometrica*, 8, 2014, 1, p. 115-140 (DOI 10.1007/S11698-013-0095-1).

Knerr S., Augustin E., “A neural network-hidden markov model hybrid for cursive word recognition,” in International Conference on Pattern Recognition, ICPR, 1998, pp. 1518–1520.

Lemercier Claire, Zalc Claire, *Quantitative Methods in the Humanities*, Charlottesville, University of Virginia Press, 2019.

Likforman-Sulem L., Zahour A., Taconet B., “Text line segmentation of historical documents: a survey,” International Journal on Document Analysis and Recognition, IJDAR, vol. 9, no. 2-4, 2007, pp. 123–138.

Manning Patrick, *Big Data in History*, Basingstoke, Palgrave MacMillan, 2013.

Michael J., Labahn R., Grüning T. *et alii* “Evaluating sequence-to-sequence models for handwritten text recognition,” in International Conference on Document Analysis and Recognition, ICDAR, 2019, pp. 1286–1293.

Papavassiliou V., Stafylakis T., Katsouros V. *et alii*, “Handwritten document image segmentation into text lines and words,” Pattern Recognition, vol. 43, no. 1, 2010, pp. 369–377.

Péroz Francis, *La Santé dans le Territoire de Belfort au XIX^e siècle*, thèse d’histoire, Université de Tours, 1997.

Plateau-Holleville Cyprien, Bonnot Enzo, Gechter Franck, Heyberger Laurent, « French Vital Records data gathering and analysis through machine learning algorithm and image processing features recognition », dans *Journal of Data Mining and Digital Humanities*, 2021 : https://doi.org/10.46298/jdmdh.7327 et https://hal.archives-ouvertes.fr/hal-03189188v3 (mis en ligne le 15 juillet 2021).

Plötz T., Fink G. A., “Markov models for offline handwriting recognition: a survey,” International Journal on Document Analysis and Recognition, IJDAR, vol. 12, no. 4, 2009, pp. 269–298.

Rygiel Philippe, « introduction générale », dans *Historien à l’âge du numérique*, Villeurbanne, ENSSSIB, 2018, pp. 7-35.

Stuner B., Chatelain C., Paquet T., “Handwriting recognition using Cohort of LSTM and lexicon verification with extremely large lexicon,” Multimedia Tools and Applications, 2020.

Swaileh W., Ait-Mohand K., Paquet T., “Multi-script iterative steerable directional filtering for handwritten text line extraction,” in 5th International Workshop on Multilingual OCR, ICDAR, 2015, pp. 1241–1245.

Yousef M., Hussain K.F., Mohammed U. S., “Accurate, data-efficient, unconstrained text recognition with convolutional neural networks,” *Pattern Recognition*, vol. 108,2020, p. 107482.

Profil demandé / Applicant profile

Ayant une formation niveau master en informatique et ayant suivi des modules en lien avec le machine learning, le.a doctorant.e devra avoir, en outre, de solides compétences en conception/programmation orientée objet, en base de données, en traitement d'image. De plus, une ouverture d'esprit et une appétence pour les travaux multidisciplinaires sera grandement apprécié. Enfin niveau d'anglais B2 minimum est exigé.

Having a Master's degree in computer science and having completed modules related to machine learning, the Ph.D. candidate should also possess solid skills in object-oriented design/programming, database management, and image processing. Moreover, an open-mindedness and a keen interest in multidisciplinary work will be greatly appreciated. Lastly, a minimum B2 level of English proficiency is required.

Preferred selection criteria:
- Computer Science background (OOP, DB, Machine Learning
- English Level

Personal characteristics:
- open minded (multidisciplinary context)
- pro-active

Financement : UBFC contrat doctoral CDD de 3 ans

Dossier à envoyer pour le **1er septembre 2023**
Début du contrat : 1er Octobre 2023

Direction de la thèse:/ **Thesis Supervisor**
GECHTER Franck, franck.gechter@utbm.fr

Encadrement de la thèse : **co-directeur(s) et co-encadrant(s)**
HEYBERGER Laurent, laurent.heyberger@utbm.fr

Applicants are invited to submit their application to the PhD supervisors.
Application must contain the following documents:
- CV
- Cover letter
- At least 1 reference letter