Intelligence artificielle et la reconnaissance de formes et d'écritures manuscrites
Journées annuelles du cluster 3 de l'EquipEx Biblissima+
7-8 mars 2024, Campus Condorcet (Aubervilliers)
Proposition de communication

Titre

Retour(s) d'expérience(s) d'indexation de registres d'emprunteurs de bibliothèques parisiennes : chaîne de production et de diffusion des archives de PRET19

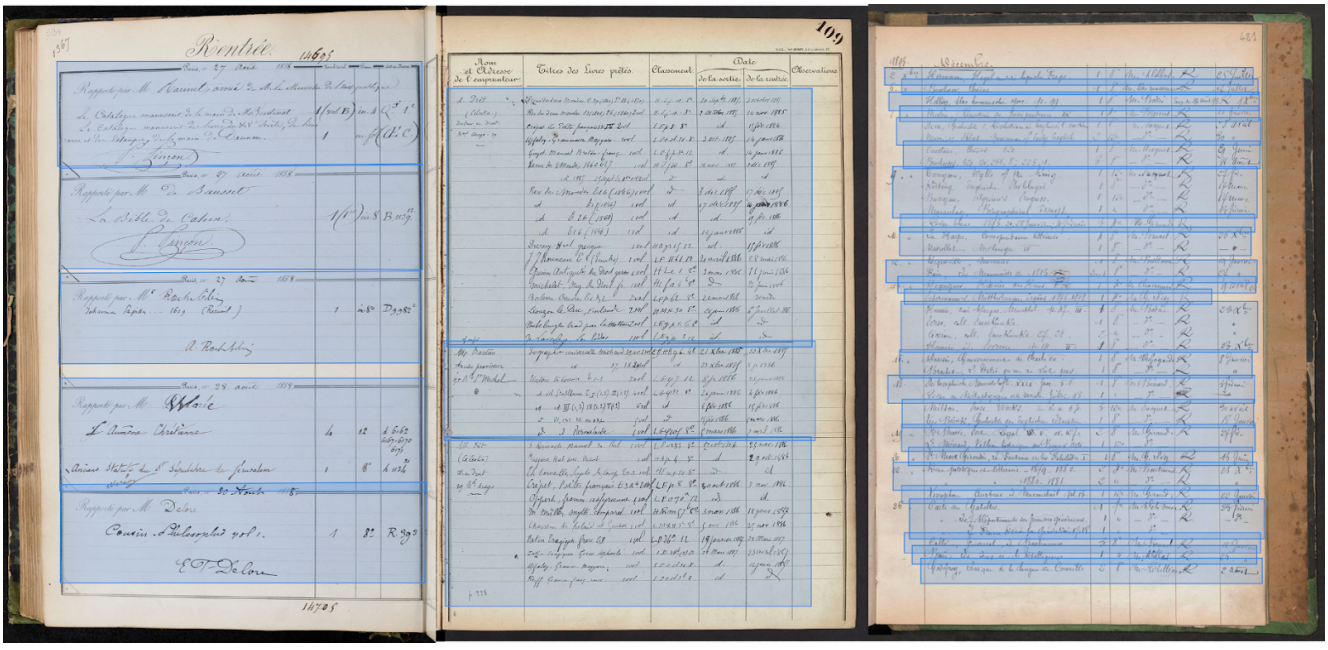
Auteurs

— TEKLIA : Eva Bardou, Christopher Kermorvant, Yoann Schneider
— PRET19 : Léa Périssier, Marie-Thérèse Petiot, Viera Rebolledo-Dhuin

Speakers

— Christopher Kermorvant
— Viera Rebolledo-Dhuin

Résumé



Le projet CollEx-Persée PRET19 (Projet de Répertoire des Emprunteurs et Titres empruntés au XIXe siècle à l'université [1]) vise à numériser et indexer les registres de prêt de plusieurs grandes bibliothèques parisiennes du XIXe siècle en vue de constituer une base de données sous Heurist, dont la modélisation pourra éventuellement servir à d'autres corpus similaires. La base PRET19 doit permettre d'interroger le contenu des registres par nom et qualité d'emprunteur, par bibliothèque, date de prêt et si possible par auteur/titre des documents empruntés. L'objectif est d'alimenter la prosopographie du monde savant et académique, d'étudier à travers le prêt en bibliothèque un mode de circulation des livres et des hommes, de repérer ainsi d'éventuels réseaux ou effets de communauté autour d'une collection, d'analyser les pratiques de tel intellectuel ou de tel autre, qu'il ait une position dominante dans son

1. https://www.collexpersee.eu/projet/pret19/

champ d'action ou non, sinon de se pencher sur la génétique des oeuvres, etc. dans l'optique d'une histoire matérielle et sociale des savoirs, attentive aux « minorités ».
Le corpus compte environ 23 500 vues aux contenus et formats divers : plusieurs types de mises en page, grande variété des écritures manuscrites, et très nombreuses ratures, sans compter la diversité des langues. Nous avons priorisé le travail sur les emprunteurs ; le travail sur les données relatives aux documents empruntés se fera dans un second temps.

Pour extraire les informations sur les emprunteurs de ces différents registres, nous avons développé une chaîne de traitement en trois étapes, exécutée sur la plateforme Arkindex : localisation, reconnaissance et indexation.
Pour les étapes de localisation et de reconnaissance, un modèle de *deep learning* a été entraîné à partir d'annotations créées grâce à la plateforme open-source Callico [2]. L'étape de localisation consiste à détecter, dans les images des pages de registres, les zones correspondant à chaque emprunteur. Cette zone contient les informations sur l'emprunteur et une liste d'ouvrages prêtés. Le modèle de détection a été entraîné en utilisant la librairie YoloV8 [3]. Une fois la zone détectée, les informations sur l'emprunteur sont extraites par un modèle de reconnaissance d'écriture manuscrite basé sur la librairie DAN [4]. Les informations extraites sont en priorité les nom, prénom et civilité. L'étape d'indexation consiste à rechercher dans des référentiels les noms des emprunteurs reconnus sur les pages. La recherche est réalisée dans une base de données ElasticSearch [5] par une requête combinant plusieurs critères : un appariement approché sur le nom de famille avec en option un appariement sur la première lettre du prénom s'il est présent. La recherche est réalisée en priorité dans un référentiel propre à chaque bibliothèque puis, en cas d'échec, dans les autres référentiels. À l'issue du processus de détection, reconnaissance et indexation, entre 80 % et 85 % des emprunteurs sont identifiés automatiquement.

La numérisation des registres et la construction de la base de données afférente s'accompagnent de la mise en place d'un site développé sous Heurist. La présentation de ce projet en cours de développement permettra de partager un retour d'expérience et d'échanger avec d'autres porteurs de projets.

CV

Christopher Kermorvant est ingénieur et docteur en Informatique, spécialisé en intelligence artificielle appliquée à la compréhension automatique des documents. Après une formation doctorale à l'EPFL (Suisse) et à l'Université de Lyon/Saint-Etienne, il a été chercheur post-doctoral à l'Université de Montréal dans le laboratoire de Yoshua Bengio. De 2005 à 2015, il dirige l'équipe de recherche de l'éditeur A2iA spécialisé en reconnaissance d'écriture manuscrite. En 2018, il fonde la société TEKLIA afin d'offrir des services de reconnaissance automatique de documents basés sur l'IA pour les institutions culturelles et patrimoniales. TEKLIA consacre 30% de son activité à des projets de recherche partenariale et diffuse ses logiciels, données et modèles en open-source.

Viera Rebolledo-Dhuin est maître de conférences en histoire moderne et contemporaine à l'UPEC. Après avoir soutenu une thèse sur les faillites des libraires-éditeurs parisiens du XIX^e^ siècle et les réseaux de crédit afférents (2011), elle a participé à deux projets ANR : DEF19 et FIDUCIAE (2014-2019). Le premier relève d'un dictionnaire prosopographique des éditeurs français du XIX^e^ siècle, l'autre vise à étudier les modalités des échanges marchands et à sonder l'évolution de l'inter-personnalité de ces échanges de part et d'autre de la *Grande transformation* (Karl Polanyi). Ses travaux se situent au croisement de l'histoire du livre et de l'histoire du crédit. Elle co-pilote le projet CollEx-Persée PRET19, qui lui permet de faire le lien entre ces différents champs de recherche, considérant le prêt en bibliothèque comme une forme de crédit du livre.

2. https://gitlab.teklia.com/callico/callico
3. https://github.com/ultralytics/ultralytics
4. https://github.com/FactoDeepLearning/DAN
5. https://github.com/elastic/elasticsearch