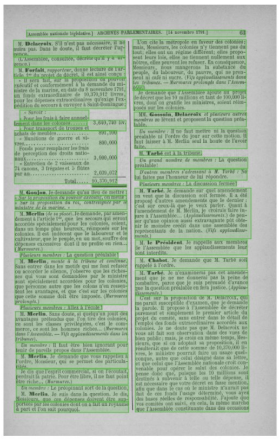
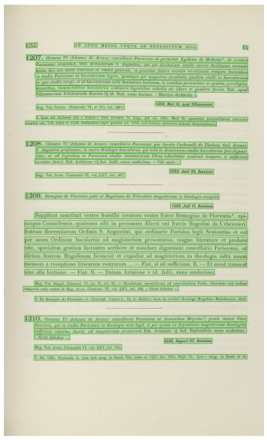
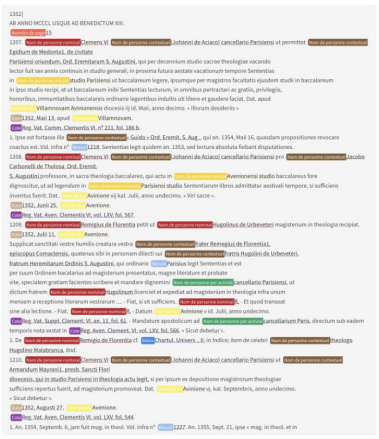
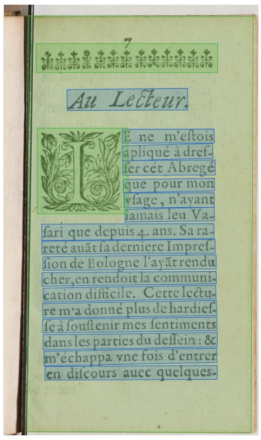
Intelligence artificielle et la reconnaissance de formes et d'écritures manuscrites Journées annuelles du cluster 3 de l'EquipEx Biblissima+
7-8 mars 2024, Campus Condorcet (Aubervilliers)
Proposition de communication

« Un simple OCR ? » : outils, algorithmes et méthodes de reconnaissance de document en humanité numérique.

Christopher Kermorvant
TEKLIA – Paris, France.

La copie, la transcription ou l'édition de textes sont au cœur des activités de recherche en humanité et en sciences humaines et sociales. Depuis l'introduction de l'informatique comme un nouvel outil au service des chercheurs, l'établissement d'une édition électronique ou l'utilisation de corpus déjà collectés et édités sont désormais des étapes incontournables des projets en humanités numériques. Si l'édition manuelle est possible à l'échelle de quelques centaines de pages, l'utilisation de systèmes de reconnaissance automatique de textes, imprimés ou manuscrits, devient rapidement indispensable pour le passage à l'échelle permettant une analyse statistique ou une lecture « distante ».

Bien que la technologie OCR soit utilisée de manière industrielle pour des documents contemporains depuis plusieurs décennies, les faibles performances de ces mêmes systèmes sur les documents historiques ont été largement identifiées et analysées. Cependant, l'étape de transcription automatique est souvent négligée lors de la planification d'un projet de recherche : considérée comme une formalité, cette étape est soit externalisée à un prestataire soit réalisée en interne en prévoyant l'usage d'outils

open-source (Tesseract, Kraken), de plateformes open-source (eScriptorium) ou commerciales (Transkribus, Arkindex). La transcription automatique est souvent pensée comme une étape préalable et indépendante des recherches qui seront ensuite menées sur son résultat. Cependant, dans la majorité des projets, ce qui semble n'être qu'un simple traitement OCR se révèle en pratique plus complexe qu'initialement prévu si l'on cherche à générer une version électronique des textes permettant les traitements et analyses subséquents.

Nous illustrerons la nécessité de définir les traitements de reconnaissance automatique en adéquation avec les objectifs d'exploitation à travers trois projets. Le premier projet, menée en collaboration avec la bibliothèque interuniversitaire de la Sorbonne et l'UAR Persée[1], a consisté en l'évaluation de trois OCR différents sur un échantillon des archives parlementaires de la Révolution française. Le second projet mené en collaboration avec le laboratoire ITEM de l'université Sorbonne Nouvelle[2] visait à réaliser une transcription automatique de treize ouvrages du XVIIème siècle en français et en italien traitant de la théorie de l'art pour une exploitation en TEI. Enfin, le dernier projet, réalisé en collaboration avec la bibliothèque interuniversitaire de la Sorbonne[3], visait à transcrire et structurer une édition critique d'actes et de lettres relatifs à l'ancienne université de Paris, publiée entre le XVII^{e} et le XIX^{e} siècle.

A travers ces trois projets, nous mettrons en évidence les outils (plateforme de gestion, de traitement et d'annotation de documents), les algorithmes de traitement (OCR, analyse de la mise en page, transformation du texte, extraction d'entités nommées, indexation) et les méthodes nécessaires afin de réaliser une transcription automatique de document guidée par les objectifs d'analyse.

CV

Christopher Kermorvant est ingénieur et docteur en Informatique, spécialisé en intelligence artificielle ap- pliquée 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 acti- vité à des projets de recherche partenariale et diffuse ses logiciels, données et modèles en open-source.

[1] https://archives-parlementaires.persee.fr/

[2] https://arterm.hypotheses.org/

[3] https://oresm.hypotheses.org/1634