eScriptorium : une application libre pour la transcription automatique des manuscrits

Développée en 2019, l'application eScriptorium dote le logiciel Kraken d'une interface graphique et facilite la conduite de campagnes de transcription automatique.

Cela fait longtemps que la transcription automatique des documents imprimés (OCR) et manuscrits (HTR) intéresse le monde de la recherche et celui des institutions patrimoniales. Le développement de processus s'appuyant sur l'intelligence artificielle et l'augmentation des capacités de calcul ont récemment ouvert de nouvelles perspectives. Dès le début des années 2000, des campagnes d'OCR ont été mises en place pour traiter les imprimés. Pour les manuscrits en revanche, ce n'est qu'à partir du milieu des années 2010 que les choses ont commencé à changer avec l'apparition de logiciels disponibles en ligne sur abonnement comme Transkribus ou en *open source* comme eScriptorium. C'est le groupe de recherche SCRIPTA PSL[1] qui développe, depuis 2019, l'application eScriptorium dont la vocation principale était de doter le logiciel Kraken[2] d'une interface graphique facilitant son utilisation. Kraken est un logiciel de transcription automatique développé en *open source* en 2015 par Benjamin Kiessling et conçu initialement pour proposer une meilleure prise en charge des textes non latins, en particulier arabes. Aujourd'hui, le groupe bénéficie des contributions d'autres infrastructures ou projets de recherche qui ont adopté l'application. Ce fut le cas du projet LectAuRep (Inria/Archives nationales) jusqu'en 2022 ou encore du groupe OpenITI (université du Maryland).

UN ESPACE DE TRAVAIL POUR GÉRER LES ÉTAPES ESSENTIELLES D'UNE CAMPAGNE DE TRANSCRIPTION

L'application eScriptorium sert d'espace de travail pour gérer les étapes essentielles d'une campagne de transcription. Celles-ci sont relativement simples : charger des images (y compris en les extrayant d'un fichier PDF ou d'un serveur IIIF), analyser la mise en page en localisant des ensembles de lignes de texte auxquelles on peut assigner des types, et enfin transcrire. Ces deux dernières étapes peuvent être réalisées à la main ou bien à l'aide de Kraken. À l'issue du processus, des triades composées d'une image, des coordonnées des lignes ou des ensembles, et de la transcription peuvent être exportées dans des formats standards (XML ALTO et PAGE) et servir à générer par exemple des éditions numériques. Ce sont aussi ces triades qui permettent de créer des modèles à l'aide de Kraken, avec ou sans l'intermédiaire d'eScriptorium. Les modèles sont des fichiers qui enregistrent une représentation abstraite des informations telles qu'elles ont été apprises par le logiciel au contact d'exemples de transcription. Cette abstraction permet à un logiciel comme Kraken de générer un texte à partir de l'analyse d'une image. En plus de ces actions essentielles, eScriptorium propose d'autres fonctionnalités pour la gestion de projet : création d'équipe, partage des transcriptions, images et modèles, regroupement des images en « documents », eux-mêmes rangés dans des « projets », étiquetage des documents, suivi de la progression, etc.

LA PRODUCTION DE MODÈLES, PRINCIPAL DÉFI À RELEVER

Pour utiliser eScriptorium, l'application doit être déployée sur un serveur Web installé sur un ordinateur personnel ou sur une machine dédiée. Les capacités de calcul du matériel employé font ensuite la différence au moment de faire tourner Kraken, en particulier lors des entraînements. Certaines institutions ou infrastructures de recherche proposent d'ouvrir des comptes sur leur serveur eScriptorium, mais il est difficile de les recenser toutes. Heureusement, il est aisé de déplacer ses données d'une instance à une autre puisque tout peut être téléchargé.

À l'heure actuelle, eScriptorium propose un modèle de segmentation par défaut efficace mais n'en propose pas pour la transcription : il faut en créer un soi-même ou trouver sur Internet ceux que d'autres utilisateurs de Kraken/eScriptorium ont créés. Certains sont déposés sur Zenodo[3] et des initiatives comme HTR-United[4] permettent de trouver des données à partir desquelles générer ces modèles. La production de modèles, qu'ils soient spécialistes d'une écriture, d'un type de document ou bien généralistes, est l'un des principaux défis à relever pour faire progresser l'implémentation de l'HTR dans les institutions patrimoniales. L'avantage de l'écosystème ouvert de Kraken/eScriptorium réside justement dans le fait qu'il permet aux utilisateurs de créer en autonomie et en toute transparence ces données et ces modèles.

Alix Chagué

Doctorante en humanités numériques au sein de l'équipe ALMnaCH (Inria – Paris) et du GREN (Université de Montréal)

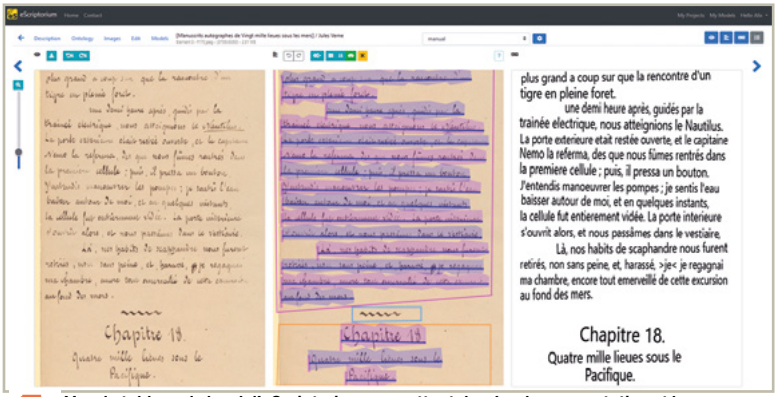
alix.chague@inria.fr

[1] SCRIPTA PSL : https://scripta.psl.eu

[2] Kraken : https://kraken.re

[3] Zenodo : https://zenodo.org/communities/ocr_models

[4] HTR-United : https://htr-united.github.io



Vue du tableau de bord d'eScriptorium permettant de gérer la segmentation et la transcription d'un document manuscrit.