Plus en plus d'institutions patrimoniales cherchent à numériser leurs ressources textuelles dans le but de démocratiser la recherche [1]. Cet objectif s'est ressenti fortement lors de la pandémie de Covid-19, quand des archives autour du monde ont fermé leurs portes physiques. Les portails des bases de données, tel que Gallica de la Bibliothèque nationale de France, sont fondamentaux pour la recherche. Mais l'extraction du texte des ressources numérisées n'est plus l'enjeu le plus important [2]. Il est tellement facile actuellement de numériser la page d'un document et d'extraire du texte brut et repérable qu'un individu possédant un portable avec un appareil photo et une application OCR (*Optical Character Recognition*) peut le faire. Les interfaces graphiques et libres, telle qu'*eScriptorium* [3], ainsi que les modèles OCR et HTR (*Handwritten Text Recognition*) publiés librement en ligne [4] ont révolutionné la recherche ainsi que l'archivage et la conservation du patrimoine. Le monde dispose désormais d'un nombre croissant de documents numérisés.

Le nouveau défi à relever aujourd'hui est de transformer ces numérisations en des ressources enrichies, qui augmentent le texte extrait et repérable avec de la métadonnée et de l'analyse. Le texte brut et non annoté ne suffit plus. De là vient l'impulsion pour le projet *Gallic(orpor)a*. Le projet envisage la mise en place d'un *pipeline* qui saisit un document numérisé depuis le portail Gallica et renvoie une ressource numérique très enrichie. En plus d'une transcription du texte repérable, la ressource présentera les données structurelles portant sur la mise en page, ainsi qu'une analyse linguistique du texte extrait et des métadonnées portant sur le document physique et le fac-similé numérique.

Le pipeline réalisé dans le cadre du projet Gallic(orpor)a a pour but de traiter automatiquement des collections de document aussi bien en ancien-français, moyen français que français classique, issus soit de manuscrits soit d'imprimés produits entre le xve siècle et le xviiie siècle. Cependant, le but sous-jacent du projet serait de parvenir à produire un prototype qui pourrait servir d'exemple pour mettre en place des chaines d'acquisition numérique pour des collections de documents issus des institutions patrimoniales.

J'ai créé l'application alto2tei afin de compléter le pipeline du projet Gallic(orpor)a, en allant des fichiers XML ALTO sortis des modèles HTR vers une version préliminaire de la ressource numérique sortie éventuellement du pipeline en format TEI. Pour résumer, la première étape du pipeline est la récupération des pages numérisées du document source. Ensuite le pipeline prédit le texte et la mise en page du fac-similé numérique en traitant chaque page avec des modèles HTR. Les données produites par les modèles sont en format XML ALTO. Mais le pipeline veut présenter les données enrichies par les métadonnées et par l'analyse linguistique en format XML TEI puisque le schème TEI est plus utilisé et il convient mieux à l'édition du texte que le schème ALTO. Mon application alto2tei a complété le pipeline en construisant un document TEI à partir des données des modèles HTR.

Le pipeline du projet Gallic(orpor)a se déroule dans cinq étapes. Dans un premier temps, il récupère les fac-similés numériques des documents sources sur Gallica. Dans un deuxième temps, il applique des modèles HTR aux fac-similés ainsi téléchargés afin de produire une prédiction du texte et une transcription de la mise en page. Ensuite, il crée un fichier TEI préliminaire qui réunit les données produites par les modèles HTR et les métadonnées récupérées de plusieurs sources en ligne. Le quatrième étape enrichit le fichier TEI avec une analyse linguistique du texte de la transcription. Et enfin, le pipeline export les données du fichier TEI

1. Depuis 2006, la Bibliothèque nationale de France s'engage à la numérisation et l'océrisation (Optical Character Recognition) en masse de ses documents pour afin qu'ils puissent être recherchés par le texte, au lieu de tout simplement la notice bibliographique, cf. **salahAdaptiveDetectionMissed2013**

2. La reconnaissance du texte à partir des manuscrits et des documents écrits dans un ancien état du français posent toujours plus de difficulté que les imprimés et que les documents en français moderne. Néanmoins, la technologie permettant l'amélioration de la reconnaissance du texte sur les manuscrits est déjà mise en place ; elle s'agit de l'entraînement des modèles supérieurs en s'appuyant sur la création d'encore meilleures données, cf. **gabayOCR17GroundTruth2020**.

3. gautierCompterenduJourneeEtude2022

4. **ModelsHuggingFace**.

en divers formats, y compris les données conformant aux schémas RDF, DTS, et IIIF.