Projet HN-2023-Nevers :

1 Introduction

Le projet HN-2023-Nevers représente un travail collectif de quatre élèves du master Humanités Numériques de l'École nationale des chartes qui vise à créer des vérités de terrain pour l'entraînement de modèles de reconnaissance automatique des écritures manuscrites (Handwritten Text Recognition, HTR). Notre démarche s'appuie sur l'utilisation de deux outils : d'une part, Git et GitHub pour la gestion collaborative des données, et d'autre part la plateforme eScriptorium pour l'entraînement de modèles de reconnaissance automatique des écritures manuscrites.

1.0.1 Présentation de la source

1.0.2 1) Contenu et acteurs

Le manuscrit retenu, daté entre 1401 et 1500, constitue une mise en prose du "Roman de la Violette" de Gerbert de Montreuil. Il a été attribué à l'auteur prétendu Jean de Wavrin, tandis que l'enluminure a été réalisée par Loyset Liédet, un enlumineur actif au XVe siècle. Les contributeurs historiques incluent Guiot d'Augerans en tant que copiste, ainsi que plusieurs personnalités de renom ayant possédé le manuscrit au fil du temps, telles que Roger de Gaignières, Philippe III (duc de Bourgogne), et Charles-Quint (empereur germanique). Il est également mentionné dans les collections de la Librairie des ducs de Bourgogne, ainsi que par des figures notables telles que Louis-Jean Gaignat, Louis-César de La Baume Le Blanc, et Louis de Gand de Mérode (prince d'Isenghien).

1.0.3 2) Lieu de conservation et disponibilité

Il est conservé à la Bibliothèque nationale de France sous la côte Français 24378. Il a été numérisé et rendu consultable en ligne depuis le 28 mars 2016.

1.0.4 3) Scriptorium et scripto

Copié dans la région des Flandres, à Bruges, ce manuscrit présente une écriture bâtarde bourguignonne typique de cette période : L'écriture bâtarde, nommée ainsi en raison de son caractère hybride, fusionne des éléments de l'écriture gothique mais avec des traits plus cursifs et moins anguleux. De fait, elle est reconnaissable par l'utilisation de lettres cursives et de ligatures,

où certaines lettres sont liées pour former des ensembles plus fluides (formes arrondies, traçage plus doux).

1.0.5 4),Description matérielle du manuscrit

Le manuscrit se compose de 348 pages en parchemin, avec un feuillet de garde en parchemin précédant et suivant, ainsi que deux feuilles de garde en papier moderne. Les dimensions sont de 270 × 185/190 mm, avec une justification d'environ 170 × 107 mm. Il est structuré en 22 cahiers : 21 cahiers de 8 feuillets chacun, numérotés de 1 à 16, 17 à 32, et ainsi de suite jusqu'à 321 à 336, et un cahier de 6 feuillets pour les pages 337 à 348. Chaque cahier est identifié par une lettre de l'alphabet, de "a" à "y", suivie du numéro de bifeuillet de "I" à "IIII". Des réclames verticales sont présentes. La réglure est effectuée à l'encre rose et violette, avec 20 longues lignes.

1.1 Choix du modèle

Après une phase préliminaire de recherche, notre choix s'est porté sur un modèle de reconnaissance : le modèle Generic CREMMA Model for Medieval Manuscripts (DOI 10.5281/zenodo.7631619). Ce choix a été guidé par la pertinence de ce modèle pour notre corpus : c'est un modèle spécifiquement conçu pour traiter des documents médiévaux en français. Il a été développé par le laboratoire CREMMALab et est adapté pour la période allant de 1100 à 1499. Notre manuscrit correspond aux deux critères énoncés ci-dessus : sa datatation se trouve dans cette borne chronologique et il contient un texte en langue française. Ce modèle a été formé sur un ensemble de données comprenant environ 21 656 documents médiévaux en français, avec un total de 579 368 mots.

1.1.1 Mise en œuvre

Après avoir sélectionné notre manuscrit, nous avons établi un plan d'exécution pour le projet : il s'agissait de déterminer les tâches à réaliser et leur répartition. Pour ce qui est de la tâche de transcription du texte du manuscrit, nous avons mis en place une section « Répartition des pages à transcrire » dans le Read.me : elle se présente sous la forme d'un tableau dans lequel chaque collaborateur du projet s'est vu attribuer des folios précis à transcrire. Une autre colonne a été ajoutée afin d'informer les autres collaborateurs de l'état d'avancement de la transcription. En ce qui concerne les tâches périphériques (création du document sur les normes de transcription, la bibliographie e.c.t…), nous avons créé un fichier Organisation.md sur la branche

main de notre dépôt Github. La répartition de ces tâches se présente également sous la forme d'un tableau qui comprend les tâches, les personnes qui y ont été assignées et leur évolution.

1.1.2 Fonctionnement de Github

-fonctionnement plutôt général -comment nous s'est approprié l'outil et on a fonctionné

1.1.3 Fonctionnement de E-scriptorium

–fonctionnement plutôt général -comment nous s'est approprié l'outil et on a fonctionné : parler notamment de la segmentation, des zones (souligner que l'on a suivi la documentation Segmonto), de la transcription(utilisation du modèle Cremma et après correction manuelle). -Réserver une section pour expliquer le clavier virtuel

1.1.4 Détermination des normes de transcription

Les normes de transcription que nous avons adoptées et qui sont détaillées dans le fichier normesTranscription.md suivent celles définies dans le cadre du projet CREMMA-Médiéval(mettre référence). Une autre norme a été rajoutée, en application particulière à notre manuscrit : ce dernier comportait parfois ce que l'on nomme des « tirets de renvoi » en fin de ligne. Le « tiret de renvoi » qui prend généralement la forme d'un tiret ou d'un tilde horizontal, signale une liaison forte entre le mot en fin de ligne et le premier mot de la ligne qui suit. Nous avons choisi de le transcrire par le symbole « -«. La transcription n'a pas soulevé de grandes problématiques si ce n'est : - cas des points médians : les folios transcrits comportaient parfois des points médians, généralement placés en fin de ligne. Parfois, le point médian était comme effacé (l'encre était beaucoup plus claire que le texte) : pour respecter les normes de transcription, nous avons inclus ces points médians dans notre transcription. - cas de l'hyphénisation : nous avons relevé quelques phénomènes d'hyphénisation, bien que rares au sein de nos folios. Conformément aux normes définies dans le cadre du projet Cremma, l'hyphénisation a été transcrite par une barre oblique « / ». Parfois, des barres obliques, dessinées à l'encre très claire, étaient tracées entre les mots mais elles résultaient d'une intention typographique et non celle de signaler une hyphénisation. Il fallait donc réfléchir à la fonction des barres obliques : marquaient-elles des pauses majeures ou étaient-elles seulement tracées pour des soucis typographiques ? -cas des lettres jumelées « i » et

« j » : L'application des principes de la transcription graphématique a rencontré quelque difficulté à propos des caractères i/j. En effet, pour un mot mot, « dijon », les caractères « i » et « j » étaient distingués graphiquement. Néanmoins, afin de suivre les normes de transcription établies dans le cadre du projet Cremmalab, nous avons pris la décision de ne pas les distinguer graphiquement. « Dijon » a donc été transcrit par « diion ». Avantages et inconvénients : L'adoption des normes de transcription du projet CREMMA-Médiéval assure une cohérence et une conformité avec les standards reconnus dans le domaine. De plus, l'ajout d'une norme spécifique pour les "tirets de renvoi" permet une transcription plus précise et fidèle au texte original. L'inclusion des points médians dans la transcription assure une représentation exhaustive du texte, même dans les cas où ils sont partiellement effacés. La transcription des phénomènes d'hyphénisation avec une barre oblique respecte les normes établies tout en tenant compte des spécificités typographiques du manuscrit. Cependant, la présence de barres obliques tracées pour des raisons typographiques pose la question de leur interprétation, nécessitant une analyse approfondie pour déterminer leur fonction. De plus, la décision de ne pas distinguer graphiquement les caractères "i" et "j" peut entraîner des confusions dans certains mots, comme "Dijon" transcrit en "diion", compromettant la lisibilité du texte dans certains cas.

1.2 Bilan : sur Github/ sur E-scriptorium et le projet en général