École nationale des chartes

Séminaire «Fondamentaux Informatiques: Git»

Master 1, 2022-2023

Prof. Jean-Victor Boby

Projet de transcription collaboratif du roman du géant Fierabras en utilisant GitHub

FIGURE 1 – Fierabras, 27 novembre, 2013 [eBook 44301], Project Gutenberg. Domaine public.

Fengyi Chen - Apolline Parmentelat - Yaëlle Zribi - Reda Tamtam

École nationale des Chartes - PSL

Résumé

Nous discutons dans cet article de notre stratégie collaborative pour travailler collectivement au sein d'une équipe de projet d'humanité numérique. Nous décrirons nos choix d'outils, de software, ainsi que de méthode. Puis nous expliquerons les obstacles que nous avons rencontrés et les stratégies que nous avons déployées pour y faire face, ainsi que les avantages d'un travail d'équipe.

Mots-clés : Fierabras, eScriptorium, transcription

Introduction

Nous avons effectué une transcription d'une monographie imprimée du *Roman du géant Fierabras*, de Jean Bagnyon, éditée par Adam Steinschaber. Cette monographie est conservée à la Bibliothèque nationale de France, dans le département Réserve des livres rares [1]. Elle a été écrite entre 1465 et 1470 et a été commandée par Henri Bolomier, chanoine de Lausanne. Le *Roman du géant Fierabras* est une œuvre en trois parties.

La première partie est constituée d'un récit abrégé de l'histoire des rois de France jusqu'à Clovis ainsi que d'un récit du règne de Charlemagne adapté du « Speculum historiale » de Vincent de Beauvais. La deuxième partie est une mise en prose et une adaptation de la chanson de geste du XIIe siècle « Fierabras ». La troisième partie est une traduction de la « Chronique de Turpindu » qui est le récit de la guerre de Charlemagne en Espagne.

Pour notre transcription, nous nous sommes servis de l'HTR (Handwritten Text Recognition), qui consiste en la prédiction d'un contenu textuel à partir d'une image de la source par une intelligence artificielle.

1. cf. notice de la BnF : `http://ark.bnf.fr/ark:/12148/cb162261683`

Celle-ci est entraînée par des transcriptions antérieures. [2] Comme nous l'expliquerons plus loin, notre objectif est de produire une transcription homogène du point de vue de la segmentation des mots, de l'utilisation de caractères spéciaux pour représenter les abréviations, des signes diacritiques ou encore des marques de correction. Cet enjeu est essentiel à tout projet HTR, afin que le modèle mis en place pour la transcription d'un certain jeu de données puisse être opérationnel pour d'autres. [3] On peut citer quelques autres projets HTR de la même période. Par exemple, le projet d'édition numérique de la *Mer des histoires*, un incunable de 1488 de Frédéric Duval. [4] L'HTR permet d'accélérer la transcription de corpus importants, comme en témoigne le travail de Lucien Dugaz concernant l'édition de l'Énéide française d'Octovien de Saint-Gelais d'après un manuscrit du XVIe siècle dans le cadre de son post doctorat à l'Ecole nationale des chartes [5], ou de Benedetta Salvati, doctorante en philologie romane à l'université de Lausanne et à l'École nationale des chartes qui utilise l'HTR pour créer une édition numérique de la Chronique rimée de Nicaise Ladam (XVIe siècle également). [6]

Les choix en termes d'ontologie et de transcription

Le document à annoter est mis en ligne sur la base de données Gallica élaborée par la BnF, avec un manifeste associé. Cela permet d'importer le fichier par l'adresse IIIF.

2. Ariane Pinche, *Compte-rendu de la séance n°1 du séminaire : « Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le Xe-XIVe siècle »*, École nationale des chartes ; Centre Jean Mabillon, 2021.

3. Ariane Pinche. Guide de transcription pour les manuscrits du Xe au XVe siècle. 2022. ffhal-03697382f

4. Ariane Pinche, Compte-rendu de la séance n°1 du séminaire : *« Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le Xe-XIVe siècle »*, École nationale des chartes ; Centre Jean Mabillon, 2021.

5. ibid

6. ibid

Au niveau de l'ontologie, nous avons défini les types suivants pour les régions : *« Main », « Commentary », « Illustration », « Title »*. Nous avons également ajouté la catégorie *« StampZone »* pour les tampons. La zone *« Main »* est la plus utilisée pour définir la partie de texte principale. Pour les pages que nous avons segmentées, il n'y a ni titre ni enluminure, donc nous n'avons pas eu l'occasion de tracer ces autres régions.

Au niveau des lignes, nous avons principalement utilisé les types *« default »* pour les lignes de texte et *« DropCapitalLine »* pour les lettrines situées au début du paragraphe. Le modèle de segmentation par défaut de Kraken, « blla.mlmodel », a été appliqué aux lignes et aux régions, verticalement de droite à gauche. Toutefois, le modèle pose problème. D'une part, les régions ne sont pas correctement délimitées et il y a de nombreux points de délimitation inutiles. Nous avons donc supprimé ces points superflus pour dessiner une région rectangulaire qui occupe toute la zone « Main ». D'autre part, le tampon n'est pas correctement identifié. Nous avons donc tracé la zone correspondante à la main.

Néanmoins, la ligne de base qui définit le sens de la ligne est correctement déterminée. Elle se trouve à droite au début de la ligne. La numérotation est globalement correcte, nous n'avons donc apporté que quelques modifications mineures.

Pour la transcription du document, nous avons commencé par utiliser le modèle GalliCorpora+ (French Early Modern Print) : DOI 10.5281/zenodo.7410359. Ce modèle est conçu pour les imprimés français du XVIe au XIXe siècle. Notre document date de 1478, à la fin du XVe siècle, mais il s'agit d'une monographie imprimée. L'Unicode utilisé est MUFI (The Medieval Unicode Font Initiative) pour retranscrire les signes particuliers tels que le et tironien, qui a pour code UTF-8 « U+204A ». eScriptorium propose une série de caractères en UTF-8, donc nous pouvons ajouter des caractères spéciaux dans le clavier d'eScriptorium, et les insérer dans

la transcription. La correction de celle-ci s'appuie sur l'eBook produit par le Projet Gutenberg1. Nous avons corrigé la transcription en suivant le protocole établi par Ariane Pinche dans son article « Guide de transcription pour les manuscrits du Xe au XVe siècle ». L'objectif était de trouver un modèle de transcription général en utilisant les formes simples qui conviennent au plus grand nombre. De ce fait, nous avons choisi de conserver une seule forme pour les caractères qui ont des variantes, comme les lettres « r », « s » et « q ». Nous avons donc supprimé les « U+017F » (s longs), les « U+A75B » (r rond) et les « U+E8B3 » (q). Nous avons également supprimé les accents sur les voyelles et les tilde ou tirets sur les consonnes. Enfin, nous avons regroupé les « i » et les « j », ainsi que les « u » et les « v », pour ne garder que les « i » et les « u ». Après discussion, nous avons choisi de conserver les lettres « ∂ » (d) et « U+A762 » (z) car se sont les seules formes présentes dans la partie du texte que nous avons transcrites.

Pour finir, nous avons exporté les fichiers ALTO une fois la segmentation et la transcription corrigées. Ce type de fichier nous permet de faire une annotation plus précise du texte en XML et d'entraîner plus tard des modèles plus adaptés avec Kraken.

Notre organisation, les difficultés rencontrées et nos choix face à ces difficultés

Organisation

J'ai commencé par créer un groupe WhatsApp, parce que le groupe de la classe est sur cette plateforme et que je pouvais donc rajouter toutes les personnes de mon groupe. J'ai créé un repository GitHub, que je leur ai partagé sur WhatsApp. Dans ce répertoire j'ai créé un *README.md* pour répartir deux pages par personne. Par la suite nous avons utilisé en

priorité GitHub pour rassembler les informations utiles, mais WhatsApp était pratique comme outil de premier contact et de messagerie.
Chen a créé un projet eScriptorium et nous a invité·e·s.

Nous avons convenu d'un rendez-vous où nous avons discuté des normes de transcription que nous voulions utiliser et nous nous sommes réparti les différentes parties du devoir. Nous avions lu les ressources sur eScriptorium, et Chen a répondu à nos questions sur la plateforme. Nous avons ensuite édité nos pages et écrit nos parties du dossier chacun.e de notre côté.

Nous avons ensuite fait un deuxième rendez-vous pour mettre en commun ce que nous avions écrit. Après avoir partagé les cas qui nous avaient posé problème, nous avons rédigé une notice définitive sur les normes de transcription que nous avons choisies. Nous avons procédé aux corrections communes, en corrigeant deux pages par personne. Nous avons également rassemblé les liens utiles dans *liens_utiles.md*.

Nous avons ensuite envoyé nos parties à Reda, qui s'occupe de la mise en page LaTeX.

Choix GitHub

Nous avons fait une certaine partie du travail collaboratif sur eScriptorium, parce que c'est une plateforme qui s'y prête bien, notamment pour les corrections communes.

Nous avons eu l'occasion de nous voir en présentiel, ce qui nous a permis de décider de certaines choses de vive voix plutôt que par GitHub, ce qui explique pourquoi nous avons peu utilisé les *issues*. Je pense en particulier aux règles de transcription que nous avons choisies.

Nous avons mis sur GitHub toutes les informations utiles à une personne qui rentrerait dans le projet :

— le

Liens utiles/ Bibliographie :

- **Lien gallica de la source**
  https://gallica.bnf.fr/ark:/12148/btv1b8600180x/f19.item

- **Lien spreadsheet groupes**
  https://docs.google.com/spreadsheets/d/1u4xDDqEqggfRDb6rrPi6Gg3Qj_ewYYYmBssuyLthu7E/edit#gid=0

- **Lien consignes (google docs)**
  https://docs.google.com/document/d/1JIpAUSxpJg3DwjQrpogxDnAdIeMin-UDr1-U9E edit#

- **Lien e_scriptorium**
  https://traces6.paris.inria.fr/document/2668/images/

- **Documentation eScriptorium : ici, et la**
  https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptoriu
  https://ephenum.hypotheses.org/1412

- **Transcription du texte**
  https://www.gutenberg.org/cache/epub/44301/pg44301-images.html#l2p1c1

- **Medieval Unicode Font Initiative**
  https://mufi.info/m.php?p=mufichar

- **Articles :**

  - **Article de Ariane Pinche**
    https://hal.science/hal-03697382/

— **Article de Benoît Sagot, Laurent Romary, Rachel Bawden, Pedro Javier Ortiz Suárez, Kelly Christensen, Simon Gabay, Ariane Pinche, Jean-Baptiste Camps**
`https://hal.science/hal-03930542/`