Compte-rendu de projet Git :
Le projet Notre-Dame

Soline Doat, Elsa Falcoz, Margaux Faure, Anaïs Mazoué, Ariane Menu

Janvier 2022

Le projet Notre-Dame vise, à partir d'un extrait des journaux quotidiens des travaux de restauration de la cathédrale Notre-Dame de Paris, à créer des données utilisables pour l'entraînement de modèles HTR. Il en découle une mise à disposition d'une transcription de la source à la communauté scientifique. Pour ce faire, le projet s'appuie sur la numérisation des dits journaux réalisée par la Médiathèque de l'Architecture et du Patrimoine (E/80/14/10).

Après une brève présentation des données étudiées, nous verrons quelle fut la méthode et les choix mis en œuvre pour mener à bien le projet. Enfin, nous aborderons les difficultés et réussites rencontrées lors de l'expérience.

1 Présentation des sources

Laissée à l'abandon depuis quelques réfections réalisées en 1804 pour le sacre de Napoléon Ier, Notre-Dame est finalement restaurée à partir de 1844 à l'initiative du ministre des Cultes de l'époque, Nicolas Martin du Nord. D'un montant total de plus de douze millions de francs, les travaux concernent une vaste partie de la cathédrale et s'étendent de 1844 à 1865 sous la direction d'Eugène Viollet-le-Duc et Jean-Baptiste Lassus.

Un suivi quotidien de l'avancement du chantier est réalisé par écrit dans un carnet regroupant l'intégralité des années des travaux. Notre projet se concentre sur l'année 1860 (ff. 330-341), rédigée par Maurice Ouradou (1822-1884), architecte et gendre de Viollet-le-Duc, alors inspecteur en chef des travaux. Comprenant douze pages, sa longueur nous a permis de traiter l'intégralité d'une année de travaux ainsi que le début de l'année 1861 se trouvant sur la partie basse du dernier folio considéré. Ces pages rendent majoritairement compte des travaux effectués sur la charpente de l'édifice et des restaurations mises en œuvre au niveau des transepts nord et sud. Ce sont avant tout les détails architecturaux et techniques qui sont précisés, les journaux ne s'attachent pas à rendre compte de la vie quotidienne sur le chantier. Quelques accidents ou morts sont toutefois indiqués.

Chaque page se décompose en deux zones principales : une zone centrale où est noté l'essentiel de l'information et une zone de marge sur la gauche où sont indiqués les années, mois et jours permettant de suivre la chronologie des travaux. Certaines marges portent également des dessins architecturaux, plus ou moins détaillés, tracés à main levée. L'écriture est une cursive fine et serrée, provenant de la même main pour les pages sélectionnées.

2 Méthode de travail

Une fois le sujet sélectionné, nous avons pu passer à la réalisation du projet en lui-même. La première étape a été de se répartir le travail à raison de deux ou trois pages par personne. Nous avons ensuite créé une branche distincte pour chacune d'entre nous afin de travailler sur nos pages respectives. Pour que le repository reste harmonieux et structuré proprement, nous avons établi des normes de nommage des branches et des fichiers. Ainsi, la branche principale se nomme "projetND" et les branches secondaires prennent la forme de Partie_[PRENOM]. De plus, chaque branche doit suivre une organisation semblable : un dossier par folio transcrit, nommé NDP_page[NUMERO DE PAGE] contenant un fichier xml, un fichier txt et une image png. Pour l'accessibilité du projet, il nous a semblé bon de rendre disponible une version de la transcription sous forme de fichier texte, ce format permettant une consultation directe de l'information textuelle sans utilisation de logiciels tiers. Ainsi, les différents fichiers sont respectivement appelés :

— NDP_page[NUMERO DE PAGE].xml,
— NDP_page[NUMERO DE PAGE].txt,
— NDP_page[NUMERO DE PAGE].png.

Concernant le travail sur la source, nous avons utilisé l'outil de segmentation automatique et de reconnaissance de caractères eScriptorium qui permet d'appliquer un processus d'HTR sur le contenu d'un document. Nous avons donc obtenu une segmentation des lignes automatiquement. Cependant, il nous a fallu déterminer manuellement les zones tel que suit :

— une zone principale qui constitue le corps du texte : "main",
— une zone marge : "marge",
— une zone pour encadrer les illustrations : "illustrations",
— une zone pour délimiter le numéro de folio : "folio",
— une zone pour délimiter les tableaux : "table".

Nous avons ensuite essayé d'appliquer le modèle HTR "Manuscrit 19e Lectaurep", disponible sur eScriptorium pour obtenir une première transcription automatique. Le résultat étant peu satisfaisant, nous avons décidé de nous baser sur des transcriptions originales. Nous avons opté pour une reproduction textuelle qui soit la plus fidèle possible au document original. Dans cette optique, les choix que nous avons fait s'inscrivent dans une démarche de transcription et non d'édition. Ainsi, nous n'avons rétabli ni l'accentuation ni l'orthographe contemporaine, ni les abréviations et avons respecté au mieux les particularités de l'écriture de l'auteur. Pour chaque mot illisible ou incertain, il a été décidé d'ouvrir une issue sur Github afin de pouvoir soumettre les difficultés à l'ensemble du groupe, à raison d'une issue par page.

Enfin, pour que notre travail soit le plus accessible possible et ainsi soit réutilisable par une personne extérieure au projet, ces normes ont été développées dans un fichier markdown readme disponible sur Github. De plus, c'est dans cette optique que nous avons décidé de rassembler les mots qui nous ont posé problème dans un lexique également disponible sur Github.

Une fois le travail de transcription et d'alignement terminé pour chacune d'entre nous, nous avons décidé d'organiser une relecture des fichiers afin de corriger les éventuelles erreurs de lecture du premier transcripteur. Pour cela, nous sommes passées par un système de *fork* et de *pull request*, afin que chaque propriétaire d'une branche puisse valider les modifications.

3 Retour d'expérience

3.1 Difficultés de transcription et techniques

L'une des premières choses dont nous avons débattu était les règles de transcription à suivre. Pour cela nous en avons discuté ensemble sur place, puis nous avons créé des issues sur le sujet qui ont abouti à la création d'un readme sur le dépôt. Toutes les spécificités liées à notre projet n'ont pas été déterminées dès le début, et les soucis ont été résolus au fur et à mesure de leur rencontre, toujours via des issues, dans un souci de collectivité. L'écriture manuscrite nous a posé également quelques difficultés de transcription, surtout pour certains mots tirés du vocabulaire architectural, d'où la création à la fin de notre projet d'une table de vocabulaire mise à disposition.

3.2 Maîtrise d'eScriptorium

Un autre souci rencontré a été le manque d'efficacité du modèle HTR "Manuscrit 19e Lectaurep" quant à nos données. Nous avons dû corriger la grande majorité (pour ne pas dire la totalité) des mots, voire réécrire les phrases entièrement. Peut-être aurions nous dû passer plus de temps à la recherche d'un modèle plus efficace, mais dans la mesure où nous n'avions que deux ou trois pages chacune à transcrire, nous avons décidé de gagner du temps en transcrivant manuellement pour nous concentrer sur l'apprentissage et l'assimilation de Git.

La prise en main d'eScriptorium n'a pas été évidente non plus. L'outil étant en version beta, nous avons été confrontées à de nombreux bugs : mots et lignes initialement transcrits qui disparaissent, bugs d'affichage des zones, problèmes dans les fichiers d'export XML, problèmes de numérotation des lignes... Mais tous ces soucis ont été résolus grâce à des issues multiples, comme par exemple pour le problème de la numérotation des lignes. La solution a été découverte par l'une d'entre nous (de manière hasardeuse) qui a ensuite créé un tutoriel déposé sur le repository, non seulement pour nous toutes, mais également dans un soucis de pérennité du projet, dans l'hypothèse où quelqu'un d'extérieur souhaiterait contribuer et rencontrerait le même problème.

3.3 Maîtrise de Git

Au début du projet, nous avons toutes eu du mal à conceptualiser le mode de fonctionnement de Git. Nous n'arrivions pas à visualiser ni à intégrer la logique du passage d'un contenu en local vers un serveur distant, simplement par des lignes de commandes tapées dans un terminal...Devant notre (grande) détresse, nous sommes allées interroger des camarades qui semblaient beaucoup mieux maîtriser le sujet que nous, et nous avons bien fait : grâce à eux et en très peu de temps, cet obstacle a été dépassé ! Nous avons donc pu commencer à faire nos premiers push test sur le repository, créer des branches, modifier nos dossiers... le projet était lancé !

Cependant, ce projet étant notre premier essai, nous avons fait l'erreur de ne pas créer d'organisation comme base au projet, mais de créer un dépôt directement sur le compte de Soline. Cela nous a porté préjudice pendant la relecture : nous avions décidé de *fork* le dépôt afin de pouvoir corriger les fichiers des unes et des autres puis de proposer la modification par *pull request*. Cependant, Soline ne pouvait pas effectuer un *fork* sur son propre dépôt et a donc dû passer directement par un *push* sur la branche des fichiers relus. Nous avions également fait l'erreur de travailler directement sur la branche main, en y envoyant nos fichiers, mais cela a été rattrapé par la suite en les supprimant et en les insérant sur nos branches respectives.

Le merge final a également posé problème : seules deux des cinq branches de travail acceptaient le *merge* vers la branche principale. Il nous a donc fallu "forcer" le merge pour les trois autres branches en acceptant la fusion d'historiques différents, à l'aide d'une commande *(git merge [nomProjet]/[nomBranche] –allow-unrelated-histories)* trouvée sur le forum d'aide *stack overflow*. Aussi, sur github, seules deux branches sont bien affichées comme merged.

Dans l'ensemble, nous maîtrisons beaucoup mieux Git qu'au début du projet. Il nous faudrait cependant encore un peu de temps et de pratique pour être parfaitement à l'aise avec l'outil (qui nous fait encore parfois des frayeurs...).

Conclusion :

En conclusion, ce projet a été pour nous l'occasion d'expérimenter le travail de versionning avec git et Github et de s'essayer à la technologie HTR en plein développement via l'interface d'eScriptorium. La continuation de ce travail pourrait permettre de couvrir l'ensemble des journaux et d'accumuler plus de données afin d'entraîner un modèle HTR sur une écriture manuscrite du XIXe siècle. La chaîne de traitement des données pourrait s'étendre à un encodage en XML-TEI et une transformation XSLT afin de produire à terme une édition électronique. L'ensemble pourrait former un projet d'actualité au vu des travaux actuels de Notre-Dame suite à son incendie.

Projets de transcription similaires

Le projet de recherche MSS-Abbadie

— Mathilde ALAIN, 2021,« Noter, classer, utiliser : les carnets de voyage d'Antoine d'Abbadie en Éthiopie », Materials and Fieldwork in African Studies n° 3 : 137-188, `https://www.sources-journal.org/560`

— Olivier JACQUOT, 2021, « Nouveau transcrithon des carnets scientifiques d'Antoine d'Abbadie en Éthiopie (1840-1852) », BnF, 13.12.2021, https ://bnf.hypotheses.org/10944 ?fbclid=IwAR0xBf3OpOPKjwju5an1Xl

— Projet de transcription collaborative : `https://transcrire.huma-num.fr/scripto/13/item?fbclid=IwAR2JdFYS`

Le projet Transcribe Bentham

— Gauthier HERBILLE, Jeremy MAZET, Axel PETIT, « Transcribe Bentham : recherche historique et crowdsourcing », Histoire et Humanités Numériques, 28.12.2016, `https://ahl.hypotheses.org/344`

Bibliographie indicative

— Françoise BERCÉ, *Viollet Le Duc*, Paris, Editions du patrimoine, 2013.

— Christine NOUGARET, Elisabeth PARINET, *L'édition critique des textes contemporains XIXe-XXIe siècle*, Paris, Les manuels de l'Ecole des chartes, 2015.

— Jean-Marie PÉROUSE DE MONTCLOS, *Architecture, méthode et vocabulaire, Paris, Éditions du patrimoine (4e édition)*, 2002.

— Daniel RAMÉE, *Dictionnaire général des termes d'architecture en français, allemand, anglais et italien*, Paris, C. Reinwald, 1868.

— Eugène VIOLLET-LE-DUC, *Dictionnaire raisonné de l'architecture française du XIème au XVIème siècle (10 tomes)*, Paris, B. Bance, A. Morel, 1854 à 1868.