(diapo 1)

Le projet Lectaurep consiste à appliquer les technologies de reconnaissance de caractères aux pages manuscrites et partiellement imprimées des répertoires de notaires de Paris conservés aux Archives nationales.

Un répertoire est un registre où les notaires consignent par ordre chronologique les actes notariés qu'ils ont établis et conservent.

Les répertoires des Archives nationales, qui vont aujourd'hui du dernier quart du XVe siècle aux années 1940, sont systématiquement numérisés et accessibles en ligne. Ils servent de clé d'accès aux actes notariés qu'ils résument, et qui sont aujourd'hui conservés – pour les minutes notariales – au département du Minutier central des notaires de Paris.

(diapo 2)

Les répertoires contiennent, par nature, une masse de données nominatives intéressant la vie, la mort et les biens de personnes physiques ou morales. Ils constituent donc des corpus de recherche en soi, où il est possible d'isoler des sous-corpus chronologiques, mais aussi géographiques, topographiques ou prosopographiques, intéressant l'histoire économique et sociale.

Le but du Minutier est d'offrir à ses usagers un service de lecture augmentée au sein de ces répertoires. Nous souhaitons, avec le projet Lectaurep, faciliter et massifier l'accès aux contenus de ces documents, pour permettre à nos lecteurs

> de gagner du temps dans l'identification des répertoires de notaires utiles à leurs recherches,
>
> de gagner en efficacité dans la constitution de corpus de recherche aussi cohérents, complets et maîtrisés que possible,
>
> et de manipuler ces corpus en filtrant leurs données.

Notre objectif est aussi de partager le résultat de nos travaux et nos retours d'expérience à des fins d'enrichissement mutuel et collectif, mais aussi pour favoriser l'interopérabilité des données produites en faisant converger nos pratiques autant que faire se peut, autour de communautés réunissant archives, bibliothèques, et musées, chercheurs et généalogistes, prestataires de solutions d'HTR ou de logiciels de gestion d'archives ou de bibliothèques.

C'est entre autres pour cette raison qu'en 2019, nous avons choisi d'utiliser le logiciel libre Kraken et son interface eScriptorium développés par PSL dans le cadre du projet eScripta.

Notre corpus cible est considérable. Il est aussi très diversifié, car nous avons fait le choix de prendre en considération un patrimoine numérique de plus de 15 ans, résultant aussi bien d'une numérisation rétrospective en NB de microfilms, que d'une numérisation directe en couleurs

d'après originaux, suivant des cahiers des charges de prise de vue et de qualité d'image qui ont évolué dans le temps. Mais surtout, nous avons potentiellement affaire à des **milliers d'écritures**.

(diapo 3)

Pour aborder le corpus des XIXe et XXe siècles, objet du projet, nous avons effectué un échantillonnage de ce qui était disponible, et nous nous sommes concentrés sur cet échantillon, dont nous n'avons pu traiter qu'une partie. Nous avons également ouvert deux chantiers plus homogènes, avec un notaire du XVIIIe siècle, Me Bronod, et des registres d'enregistrements de contrats de mariage, séparations et divorces de commerçants.

Sur l'ensemble de ces lots, représentant **150 à 200 mains**, environ **2000 pages ont été transcrites (soit quelques dizaines de mains)**, dont moins d'un tiers a été relu, ce qui n'a pas empêché d'élaborer des modèles d'HTR satisfaisants, qu'il a bien sûr été nécessaire d'affiner sur la base d'une vérité terrain de qualité (la vérité « vraie ») afin d'obtenir des CER inférieurs à 10, voire à 5 %.

(diapo 4)

Sur un plan méthodologique, nous avons d'abord envisagé de créer des modèles d'HTR spécifiques, propres à une écriture donnée, en pensant regrouper ces modèles par grandes familles d'écritures affiliées entre elles à la manière d'arbres généalogiques ou de stemmas. Mais nous nous sommes heurtées à une difficulté matérielle et diplomatique, à savoir l'impossibilité de prédire de manière systématique le taux de rotation de différentes écritures au sein d'un registre donné, ce taux pouvant varier selon les époques et selon les pratiques notariales.

(diapo 5)

Nous avons donc opté pour la constitution de modèles génériques « taille unique », à adapter éventuellement à des écritures spécifiques, tout en permettant des recherches floues. Ces modèles sont d'ores et déjà en mesure de reconnaître des mots entiers dans des corpus étrangers à Lectaurep, avec, bien sûr, beaucoup d'erreurs aussi.

(diapo 6)

L'état d'avancement du projet sera présenté à la BnF le 10 décembre prochain et rediffusé simultanément. Pour le résumer, les technologies d'HTR ont fait leurs preuves sur un corpus contemporain d'écritures administratives, mais leur déploiement à l'échelle du corpus cible demande aujourd'hui des infrastructures dédiées et une logistique participative impliquant l'animation de bénévoles. Il faut par ailleurs envisager d'affiner non seulement des modèles d'HTR, mais aussi des modèles de segmentation, dont aucune métrique vraiment précise ne permet d'évaluer la qualité, à la différence de l'HTR.

Enfin, une manipulation visuelle des données doit pouvoir s'appuyer sur la reconnaissance des entités nommées, qui repose ici sur des données d'HTR brutes. Ces données peuvent être précorrigées avec des outils de traitement automatique des langues, pour lesquels le projet Lectaurep met à disposition un sous-corpus linguistique du langage naturel, qui est le langage

administratif écrit, parfois fortement abrégé et télégraphique des notaires parisiens contemporains.

(diapo 7)

Le projet Lectaurep, qui s'achève ce mois-ci, s'est déroulé de 2018 à 2021 grâce à la convention cadre Culture - Inria, et nous remercions chaleureusement l'ex-DIN (SNUM) du MIC, en particulier Bertrand Sajus et Nicolas Orsini, pour leur confiance, Laurent Romary et Benoît Sagot, de l'équipe-projet AlMANAch d'Inria, pour avoir retenu notre proposition, ainsi que les artisans de Kraken-eScriptorium à PSL, Daniel Stökl, Peter Stokes, Ben Kiessling, Robin Tissot, et Marc Bui.

Nos remerciements vont également à la TGIR Huma-Num pour la mise à disposition d'un espace ShareDocs.

Et nous remercions, enfin, tout particulièrement Lionel Tadjou et Yves Tadjo-Tapianki à Inria, Marie-Laurence Bonhomme, Lucas Terriel, Hugo Scheithauer, stagiaires du master TNAH de l'ENC, Alix Chagué, chef de projet côté ingénierie de 2019 à 2021 à Inria, Gaetano Piraino et Frédéric Zamarreno à la DINUC, et toute l'équipe projet au DMC, sans oublier nos étudiants stagiaires.

Je passe à présent la parole à Hugo.