Le projet *Gallic(orpor)a* s'est développé à partir de plusieurs projets précédents et il tire parti de divers domaines de recherche. Ses créateurs, en mettant en valeur leurs propres connaissances, ont visé à assembler un pipeline qui peut traiter tout document dans la base de données Gallica de la Bibliothèque nationale de France (BnF). Les chercheurs spécialisés en *l'Handwritten Text Recognition* (HTR), en le Traitement automatique des langues (TAL), en l'histoire, en la littérature, en la lexicographie et en la stylométrie se sont rassemblés pour réaliser ce pipeline. Le pipeline visait à prédire et analyser du ancien français et du français de l'Ancien Régime, ainsi que les manuscrits et les imprimés, à partir des pages numérisées des documents créés entre 1400 et la révolution française. Cependant, le vrai rêve du projet était de produire un prototype qui servirait d'exemple et pourrait être élaboré dans le but de traiter vraiment tout document source numérisé.

Les ambitions du *Gallic(orpor)a* se sont rendu possibles grâce aux recherches de plusieurs chercheurs et ingénieurs, tel que Laurent Romary, Philippe Gambette, Thibault Clérice, Pedro Suarez Ortiz, Claire Jahan, Caroline Corbières, et Alexandre Bartz. Mais les principaux qui se chargeaient de la surveillance du projet *Gallic(orpor)a* lors de mon stage en 2022 étaient Jean-Baptiste Camps, Simon Gabay, et Ariane Pinche, qui ont développé des modèles HTR pour extraire du texte des document numériques dans la base de données Gallica. Chez Inria, en tant que stagiaire, j'ai aussi travaillé en collaboration avec Benoît Sagot et Rachel Bawden, qui ont développé des outils d'analyse linguistique du texte extrait. Tous ensemble, ces chercheurs de divers spécialités ont contribué leurs connaissances pour produire un processus du traitement polyvalent.

1 Le contexte du projet

1.0.1 Bibliothèque nationale de France et le Data Lab

Le Data Lab s'est mis en place au sein de la Bibliothèque nationale de France (BnF) en 2021.[1] Lors de sa première année, le Data Lab a lancé son premier appel aux projets qui mettent en valeur les fonds et les ressources de l'institution phare patrimoniale. Le projet *Gallic(orpor)a* faisait partie des premiers projets acceptés en 2021, à côté des projets *AUREJ* (Accès Unifié aux REssources de la Jouabilité), *GALLICAENV*, *BUZZ-F*, et *AGODA* (Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale).[2] Ayant sa candidature retenu, *Gallic(orpora)* profitait du financement du Data Lab de la BnF. La plupart du travail sur le projet a eu lieu pendant la première moitié de 2022, suite au mis en place du stage et des vacations.

[1] Marie Carlin and Arnaud Laborderie. "Le BnF DataLab, Un Service Aux Chercheurs En Humanités Numériques". In: *Humanités numériques* 4 (Dec. 2021). URL: `https://hal-bnf.archives-ouvertes.fr/hal-03285816` (visited on 08/11/2022).

[2] Bibliothèque nationale de France. *Rapport d'activité 2021 de la Bibliothèque nationale de France*. Paris, France, July 1, 2022. URL: `https://www.bnf.fr/fr/bnf-rapport-dactivite-2021` (visited on 08/09/2022), p. 123.

1.0.2 Inria et l'équipe ALMAnaCH

Inria est l'Institut national de recherche en sciences et technologies du numérique et il compte plusieurs branches dans le monde. La branche parisienne encadre l'équipe ALMAnaCH dont le acronyme veut dire *Automatic Language Modelling and Analysis & Computational Humanities*. Au sein d'ALMAnaCH s'est développé le meilleur modèle TAL pour la langue française, CamemBERT.[3] L'équipe ALMAnaCH encadre les chercheurs, les ingénieurs, les doctorants, et les stagiaires attachés aux projets concernés soit par le traitement automatique des langues, soit par les humanités numériques. L'acronyme du nom fait référence à ces deux pôles de recherche : *Automatic Language Modelling and Analysis* est le traitement automatique des langues, et le *Computational Humanities* est l'humanités numériques. Le projet *Gallic(orpor)a* se situait entre les deux, impliquant l'extraction des données et l'édition des documents historiques ainsi que l'analyse linguistique du texte extrait.

Le directeur de recherches d'ALMAnaCH est Benoît Sagot, qui s'est chargé de l'encadrement du stage du projet *Gallic(orpor)a*. En tant que stagiaire, je faisais partie de l'équipe entre début avril et fin juillet 2022. Pendant le stage, Rachel Bawden a animé un groupe de lecture hebdomadaire et des séminaires mensuelles dont j'ai profité dans l'intérêt de me tenir au courant sur les nouvelles recherches et les nouveaux enjeux du TAL. Elle a aussi développé un modèle TAL pour le projet *Gallic(orpor)a* et m'a instruit dans sa mise en œuvre.[4] Disposée d'un bureau, j'ai travaillé en présentiel quatre jours par semaine, passant un jour toutes les semaines à l'École nationale des chartes pour travailler à côté de l'une des chefs du projet, Ariane Pinche. Chez Inria, j'ai profité de l'expertise de mes collègues de l'équipe ALMAnaCH, en particulier Alix Chagué et Hugo Scheithauer. L'équipe entière de *Gallic(orpor)a* a aussi profité des serveurs d'Inria, qui prennaient en charge une partie de la puissance de calcul et du stockage de données pour l'interface graphique HTR *eScriptorium*.

1.0.3 École nationale des chartes et l'université de Genève

En tant qu'une école, contrairement à une équipe de recherche comme ALMAnaCH, les rôles de l'École nationale des chartes et l'université de Genève dans le projet *Gallic(orpor)a* concernés l'encadrement des chercheurs qui y ont contribué leurs connaissances. L'École nationale des chartes (ENC) a aussi donné un lieu de travail, dont j'ai profité un jour par semaine. Ariane Pinche, qui

[3] Louis Martin et al. "CamemBERT: A Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Online: Association for Computational Linguistics, July 2020, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645. URL: https://aclanthology.org/2020.acl-main.645 (visited on 08/11/2022).

[4] Rachel Bawden et al. "Automatic Normalisation of Early Modern French". In: LREC 2022 - 13th Language Resources and Evaluation Conference. June 20, 2022. DOI: 10.5281/zenodo.5865428. URL: https://hal.inria.fr/hal-03540226 (visited on 08/11/2022); Simon Gabay. *FreEM-corpora/FreEMnorm: FreEM Norm Parallel Corpus*. Zenodo, Jan. 17, 2022. DOI: 10.5281/zenodo.5865428. URL: https://zenodo.org/record/5865428 (visited on 08/11/2022).

était post-doctorante à l'École nationale des chartes, et Simon Gabay, maître-assistant à l'université de Genève, ils ont géré la mise en place du stage et des vacations que la bourse du Data Lab de la BnF a financés pour 2022. L'ENC et l'université de Genève les ont soutenu lors de l'encadrement des vacations et du stage.

Gabay, Pinche, et deux autres chercheurs qui étaient attachés à l'École nationale des chartes pendant le stage, Jean-Baptiste Camps et Thibault Clérice, ont tous contribué au projet *Gallic(orpor)a*. Gabay et Pinche se sont occupés de l'harmonisation des vérités de terrain produites par l'équipe en reliant toute transcription faite dans l'interface graphique *eScriptorium*. Pinche et Clérice ont commencé à utilisé les vérités de terrain des documents médiévaux en entraînant des nouveaux modèles de l'HTR et de la segmentation.[5] Par rapport à la segmentation, Jean-Baptiste Camps, Pinche, et Gabay ont développé le syntaxe *SegmOnto* qui servait à harmoniser les vérités de terrain produites pour tout document dans le corpus d'entraînement, y compris les manuscrits et les imprimés.[6] Bien que chaque chercheur ait ses spécialités, ils ont tous collaboré et la division des aspects du projet n'étaient pas aussi fermes qu'ils n'ont pas profité des idées de l'un et de l'autre.

2 Les prédécesseurs du projet

Comme expliqué avant, *Gallic(orpor)a* a rassemblé les recherches de plusieurs projets précédents. Il a profité des glossaires codicologiques, des progrès dans la prédiction et la segmentation des documents, des progrès dans l'analyse linguistique, et des catalogues des données. Le glossaire *SegmOnto* se servait à harmoniser les vérités de terrain pour les manuscrits et les imprimés transcrits. Les modèles HTR étaient à la fois un support à la production des vérités de terrain, en faisant en premier temps une transcription préliminaire, et le but du projet, étant de produire les modèles entraînés sur le vocabulaire *SegmOnto*. le catalogue HTR-United, spécifiquement son outil `HTRVX`, a surveillé l'harmonisation des transcriptions produites comme des vérités de terrain.[7] En outre, le catalogue HTR-United a publié gratuitement les vérités de terrain du projet *Gallic(orpor)a* pour que d'autres projets et d'autres modèles HTR puissent en profiter.[8] Pour terminer, l'analyse linguistique était aussi un objectif du

[5] Thibault Clérice. *YALTAi: Segmonto Manuscript and Early Printed Book Dataset*. Zenodo, July 10, 2022. DOI: 10.5281/zenodo.6814770. URL: https://zenodo.org/record/6814770 (visited on 08/12/2022).

[6] Simon Gabay et al. "SegmOnto: Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More)". In: *1st International Workshop on Computational Paleography (IWCP@ICDAR 2021)*. Lausanne, Switzerland, Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03336528 (visited on 08/09/2022).

[7] Thibault Clérice and Ariane Pinche. *HTRVX, HTR Validation with XSD*. version 0.0.1. Sept. 2021. DOI: 10.5281/zenodo.5359963. URL: https://github.com/HTR-United/HTRVX (visited on 08/12/2022).

[8] Alix Chagué and Thibault Clérice. "Sharing HTR Datasets with Standardized Metadata: The HTR-United Initiative". In: Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites. June 23, 2022. URL: https://hal.inria.fr/hal-03703989 (visited

projet *Gallic(orpor)a*, et la mise en œuvre de cet aspect comptait sur les modèles de la langue française construits par les chercheurs de l'équipe ALMAnaCH.

2.0.1 Le glossaire codicologique

Le projet *Gallic(orpro)a* a harmonisé ses données selon le vocabulaire *SegmOnto*, qui est expliqué en détail dans le chapitre **??**. Ayant géré les données produites selon cette codicologie, j'ai aussi contribué à l'élaboration et le perfectionnement du vocabulaire. La décision d'utiliser le syntaxe descriptif des lignes et des zones de *SegmOnto* a été prise bien en avance de la mise en œuvre du projet *Gallic(orpor)a*. La généralité du vocabulaire était déterminée de bien conformer à la diversité ciblée des documents traités dans le cadre du projet. Puisque *Gallic(orpor)a* visait à livrer un prototype d'un pipeline qui pourrait traiter tout document source numérisé, la généralisation des étiquettes appliquées aux lignes et aux zones était impérative, et le vocabulaire de *SegmOnto* était jugé la meilleure solution.

2.0.2 La segmentation et la prédiction du texte

Les progrès de la reconnaissance du texte sont expliqués en détail dans le chapitre **??**. Un logiciel HTR commence par la segmentation de la page, et dès qu'il sait où se trouvent les caractères, les mots, et les lignes du texte il le prédit à partir de l'écriture. Ces deux tâches se font selon les compétences qu'il a appris lors de son entraînement. Dans le but de produire les vérités de terrain pouvant entraîner les modèles HTR pour les manuscrits, les incunables, et les imprimés dans la base de données Gallica, le projet *Gallic(orpor)a* a profité de l'expertise de Simon Gabay et d'Ariane Pinche, qui s'occupaient de la relecture des vérités de terrain et la gestion des corpus d'or.

Les progrès dans la prédiction du texte sur les imprimés de l'Ancien Régime ainsi que son analyse ont aidé le projet *Gallic(orpor)a*. Par rapport aux progrès dans l'OCR des imprimés françaises de l'Ancien Régime, Gabay a entraîné les modèles sur les vérités de terrain des imprimés du XVIe au XVIIIe siècle.[9] En collaboration avec d'autres chercheurs, il a travaillé sur le jeu de données OCR17+ qui a fourni des vérités de terrain des imprimés du XVIIe siècle.[10] Pour tester le pipeline, dans l'attente des modèles nouvellement entraînés sur les données produites dans le cadre de *Gallic(orpor)a*, j'ai utilisé un modèle de segmentation et un modèle d'HTR que Gabay a développé dans le cadre de son projet *E-ditiones*.[11]

---

on 08/12/2022).

[9] Simon Gabay et al. "Standardizing Linguistic Data: Method and Tools for Annotating (Pre-Orthographic) French". In: *Proceedings of the 2nd International Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisia, Oct. 2020. DOI: 10.1145/3423603.3423996. URL: https://hal.archives-ouvertes.fr/hal-03018381 (visited on 08/12/2022).

[10] Simon Gabay, Thibault Clérice, and Christian Reul. *OCR17: Ground Truth and Models for 17th c. French Prints (and Hopefully More)*. May 2020. URL: https://hal.archives-ouvertes.fr/hal-02577236 (visited on 08/12/2022).

[11] Simon Gabay. *E-Ditiones, 17th c. French Sources*. Nov. 2018. URL: https://hal.archives-ouvertes.fr/hal-02388415 (visited on 08/10/2022).

Pour la reconnaissance du texte sur les documents médiévaux, le projet CREMMALab est clef. Géré dans le cadre des études postdoctorales d'Ariane Pinche, le Consortium Reconnaissance d'Écriture Manuscrite des Matériaux Anciens, ou CREMMA, est un dépôt des images et leurs transcriptions corrigées à la main, c'est-à-dire des vérités de terrain. Afin d'entraîner un modèle HTR, il faut un jeu des vérités de terrain.[12] Le projet CREMMALab fournit un jeu des vérités de terrain de 13 manuscrits médiévaux qui se composent de 21 656 lignes de texte transcrites.[13] Sur les données du projet, Pinche a entraîné un modèle HTR qui est désormais disponible sur Zenodo.[14] Le projet *Gallic(orpor)a* en a profité pour aider à la création des vérités de terrain pour les manuscrits médiévaux de son propre jeu de données.

2.0.3 L'harmonisation et la partage des données

Le projet HTR-United mis en commun les vérités de terrain générées par tout projet *open source*.[15] Sa base de données, gratuitement mise en ligne par GitHub, contient les images et leurs transcriptions faites par plusieurs projets de recherche, et elle porte sur les documents de plusieurs périodes historiques et écritures. Un modèle HTR peut être entraîné sur ces jeux de données. Par exemple, Alix Chagué a entraîné un modèle HTR sur les vérités de terrain du *LECTAUREP Project*, soutenu par Inria et les Archives Nationales, qui sont mis en commun sur la base de données HTR-United.[16]

Dans l'esprit de la science ouverte, le projet *Gallic(orpor)a* a transféré toute vérité de terrain de ses dépôts GitHub vers le catalogue HTR-United. À la fin du stage, en juillet 2022, Ariane Pinche et Thibault Clérice ont entraîné un modèle pour les manuscrits médiévaux en utilisant les transcriptions que l'équipe du projet *Gallic(orpor)a* ont produites. Ces vérités de terrain sont désormais mises en commun sur HTR-United et Pinche et Clérice ont lié le premier modèle publié du projet *Gallic(orpor)a* avec le catalogue HTR-United et le dépot du projet CREMMA (Consortium Reconnaissance d'Écriture Manuscrite des Matériaux Anciens).[17] La partage des données du projet est l'un de ses objectifs.

[12] Alix Chagué, Thibault Clérice, and Laurent Romary. "HTR-United : Mutualisons La Vérité de Terrain !" In: *DHNord2021 - Publier, Partager, Réutiliser Les Données de La Recherche : Les Data Papers et Leurs Enjeux*. Lille, France: MESHS, Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03398740 (visited on 08/10/2022).

[13] Ariane Pinche and Jean-Baptiste Camps. "CremmaLab Project: Transcription Guidelines and HTR Models for French Medieval Manuscripts". In: *Colloque "Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites"*. Paris, France, June 2022. URL: https://hal.archives-ouvertes.fr/hal-03716526 (visited on 08/10/2022).

[14] Ariane Pinche. "HTR model Cremma Medieval". In: (June 21, 2022). DOI: 10.5281/zenodo.6669508. URL: https://zenodo.org/record/6669508 (visited on 08/10/2022).

[15] Alix Chagué et al. *HTR-United/Htr-United: V0.1.28*. Zenodo, Aug. 10, 2022. DOI: 10.5281/zenodo.6979746. URL: https://zenodo.org/record/6979746 (visited on 08/12/2022).

[16] Alix Chagué. *LECTAUREP Contemporary French Model (Administration)*. Zenodo, May 12, 2022. URL: https://zenodo.org/record/6542744 (visited on 08/12/2022).

[17] Ariane Pinche and Thibault Clérice. *HTR-United/Cremma-Medieval: Cortado 2.0.0*. Zenodo, July 11, 2022. DOI: 10.5281/zenodo.6818057. URL: https://zenodo.org/record/6818057 (visited on 08/12/2022).

Ainsi qu'à contribuer au catalogue, le projet *Gallic(orpor)a* a aussi profité des outils de HTR-United. Le dernier met en commun des outils qui ont pour but d'harmoniser les données ajoutés à son catalogue. Ces outils peuvent être intégrés dans un *workflow* de GitHub, ce que l'équipe de *Gallic(orpor)a* a fait. L'un de ces outils est `HTRVX`, qui se prononce comme le personnage Asterix, et il a rendu possible à l'équipe du projet nettoyer les transcriptions sorties de *eScriptorium*.[18] En exemple, `HTRVX` relit les transcriptions et les cherche pour les erreurs communes. L'existence d'un tel outil et sa disponibilité gratuite grâce au projet HTR-United a beaucoup aidé le projet *Gallic(orpor)a*.

2.0.4 L'analyse linguistique

Après l'extraction et le nettoyage des données des documents source de Gallica, le projet *Gallic(orpor)a* a envisagé à analyser le texte. Dans cet objectif, il a profité des progrès dans l'analyse linguistique des anciens états de la langue française. L'analyse linguistique du français de l'Ancien Régime, tel que ce qui se voit dans les écrits de Molière, est un domaine de recherche actuellement en plein développement. Depuis une dizaine d'années, les chercheurs dans la linguistique computationnelle ont élaboré des outils pour analyser le français autre que le français contemporaine, dont les recherches sont déjà animées par l'application commerciale et les jeux de données plus nombreuses.

L'histoire de l'analyse linguistique et du Traitement automatique des langues (TAL) est hors de ce mémoire. Néanmoins, les projets qui ont précédés *Gallic(orpor)a* et sur lesquels il a compté méritent de discussion. Achim Stein a bordé le sujet de l'analyse linguistique du français du Moyen Âge dans son article de 2013.[19] Stein a montré que, pour les chercheurs qui ont débuté d'appliquer les progrès dans l'analyse linguistique à l'étude du français des anciens états, il faut faire attention aux propriétés syntaxiques et morphologiques propres à la langue. Du coup, une architecture qui pourrait parvenir aux résultats souhaités pour l'anglais du Moyen Âge n'aura pas forcement le même taux de réussite avec le français du Moyen Âge à cause de différences syntactiques et morphologiques dans la langue.

Achim Stein et Sophie Prévost ont créé un *treebank* pour l'ancien français, le *Syntactic Reference Corpus of Medieval French* (SRCMF), qui avance toujours l'analyse linguistique des anciens états du français.[20] En 2014, Prévost est des autres chercheurs, Gael Guibon, Isabelle Tellier, Matthieu Constant, et Kim Gerdes, ont ajouté aux conclusions de Stein que la variation lexique de l'ancien français pose aussi un défi à l'analyse linguistique.[21] En 2019, Mathilde Reg-

[18] Clérice and Pinche, *HTRVX, HTR Validation with XSD*.

[19] Achim Stein and Sophie Prévost. *Syntactic Annotation of Medieval Texts: The Syntactic Reference Corpus of Medieval French (SRCMF)*. Narr Verlag, 2013, p. 275. ISBN: 978-3-8233-6760-4. URL: `https://halshs.archives-ouvertes.fr/halshs-01122079` (visited on 08/10/2022).

[20] Achim Stein and Sophie Prévost. *Syntactic Reference Corpus of Medieval French (SRCMF)*. Stuttgart: ILR University of Stuttgart, 2013. ISBN: 899-492-963-833-3. URL: `http://srcmf.org`.

[21] Gaël Guibon et al. "Parsing Poorly Standardized Language Dependency on Old French".