HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites

Alix Chagué, Thibault Clérice, Laurent Romary

HTR-United : un écosystème pour une approche mutualisée de la transcription automatique des écritures manuscrites

Alix Chagué, ALMAnaCH, Inria, Paris, France

0000-0002-0136-4434

alix.chague@inria.fr

Thibault Clérice, LATTICE, École normale supérieure, Paris Sciences Lettres, Paris, France

0000-0003-1852-9204

Laurent Romary, ALMAnaCH, Inria, Paris, France

0000-0002-0756-0508

laurent.romary@inria.fr

Résumé

La reconnaissance des écritures manuscrites (HTR pour *Handwritten Text Recognition*) est un procédé informatique qui vise à obtenir un équivalent de texte numérique à partir de l'image d'un document physique comportant du texte manuscrit. S'appuyant sur Github, HTR-United invite la communauté des utilisateurs à décloisonner les données issues de différentes plateformes HTR afin de réduire leur coût de production. Cette solution propose un modèle opérationnel susceptible d'offrir un cadre pour la construction de *data papers* pour l'HTR, voire les prémices d'une standardisation pour ce genre de publication

Mots-clés

Reconnaissance des écritures manuscrites, reconnaissance optique de caractères, numérisation, mutualisation, Github, articles de données

Abstract

Handwritten Text Recognition (HTR) is a computer process that aims to obtain digital text equivalent to the content of the image of a physical handwritten document. Based on Github, HTR-United invites the community of users to decompartmentalize data sourced from different HTR platforms in order to reduce the costs of producing such data. This solution proposes an operational model that could offer a framework for the construction of data papers for HTR, and even the beginnings of a standardization for this type of publication.

Keywords

Handwritten text recognition, optical character recognition, digitization, sharing, Github, data papers.

Alix Chagué est doctorante en Humanités numérique à l'École Pratique des Hautes Études et à l'Université de Montréal, elle s'intéresse en particulier à l'utilisation de la transcription automatique pour des collections patrimoniales, elle est également co-coordinatrice pédagogique du Master 2 Documentation et Humanités Numériques à l'Ecole du Louvre.

Thibault Clérice est docteur en Lettres et Civilisation Antiques (Laboratoire HISOMA, Université Lyon 3). Ses recherches portent principalement sur le traitement automatique des langues anciennes à travers l'apprentissage profond, la mise à disposition de corpora la recherche en méthodes computationnelles appliquées aux humanités.

Laurent Romary était Directeur de recherche à Inria jusqu'à sa nomination en tant que directeur à la culture et à l'information scientifique, en janvier 2022. Il a mené pendant des années des recherches en informatique linguistique et en humanités numériques, au cours desquelles il a développé un intérêt marqué pour la normalisation et le partage de données linguistiques ouvertes, et identifié la nécessité de mettre en œuvre des infrastructures mutualisées au service de la science ouverte.

Introduction

Depuis quelques années, les projets en humanités numériques intègrent des tâches de transcription automatique d'écritures manuscrites pour l'acquisition des corpus, confirmant le transfert de cette technologie du domaine expérimental de la vision par ordinateur vers le grand public. En témoigne le développement de logiciels conviviaux, libres ou propriétaires, proposant des solutions quasi-clefs-en-main, tels que Transkribus [Kahle et al., 2017], eScriptorium [Stökl Ben Ezra, 2021] ou encore Arkindex [Teklia, 2021]. Parmi les projets ayant eu recours à ces logiciels, on peut citer HIMANIS [Stutzmann et al., 2017], Ffl [Massot et al., 2019], HORAE [Boillet et al., 2019], TIME US [Chagué et al., 2019], MaRITEM [Mariotti, 2020], LECTAUREP [Chagué et al., 2020]. On pourrait en déduire que n'importe qui peut désormais se lancer dans un projet de reconnaissance automatique d'écritures manuscrites, mais il reste en réalité de nombreux points de blocage. Ainsi, bien qu'elles soient à portée de main, les plateformes techniques implémentant des solutions de transcription automatique ne sont pas encore en mesure de traiter toutes les formes d'écritures manuscrites et nécessitent de grandes quantités de données pour cela. Produire ces données a un coût que la mutualisation des efforts peut atténuer.

Nous présentons dans ce papier un écosystème nommé HTR-United [HTR-United et al., 2020/2021] facilitant la mise en commun de la vérité de terrain. Cette solution propose un modèle opérationnel susceptible d'offrir un cadre pour la construction de *data papers* pour l'HTR, voire les prémices d'une standardisation pour ce genre de publication. Nous commençons par rappeler le fonctionnement de la transcription automatique ainsi que ses limites actuelles. Nous démontrons ensuite l'importance stratégique de décloisonner les données issues des activités préparatoires à l'HTR avant de présenter la solution mise en œuvre par l'intermédiaire de l'écosystème HTR-United. Cet écosystème est construit dans une logique minimaliste et s'appuie sur la plateforme Github ; nous en détaillons le fonctionnement et la structure. Enfin, nous revenons sur l'importance de mettre en place un contrôle qualité sur les données publiées et présentons les outils intégrés dans l'écosystème HTR-United pour aider à cela.

Principes de la transcription automatique

La reconnaissance des écritures manuscrites, que l'on appelle aussi HTR (*Handwritten Text Recognition*), est un procédé informatique qui vise à obtenir un équivalent de texte numérique à partir de l'image d'un document physique comportant du texte manuscrit. Ce traitement est décomposé en trois tâches (Figure 1) dont deux (1, 3) sont indispensables : on commence (1) par localiser l'emplacement du texte sur l'image de manière à produire en ensemble de coordonnées (segmentation) ; puis (2) en fonction des logiciels et des besoins, on peut déterminer automatiquement l'organisation logique de chaque segment par rapport aux autres et par rapport à la page (analyse de la mise en page) ; enfin, (3) on reconnaît les lettres et les mots tracés dans chaque portion de l'image définie par les coordonnées d'un segment (transcription).

Ces tâches relèvent du domaine de l'apprentissage profond, il est donc nécessaire d'entraîner, pour chacune d'entre elles, des modèles à partir de données d'exemple. Ce sont ces exemples que l'on appelle la vérité de terrain : des ensembles de données annotées de manière à fournir au modèle des paires composées d'une part d'une image ou d'une portion d'image (entrée) et d'autre part de l'annotation attendue (sortie). Celle-ci peut être des coordonnées dans le cas de la segmentation ou un ensemble de caractères dans celui de la transcription. Les performances des modèles dépendent de l'efficacité de l'architecture neuronale mise en place, mais aussi de la qualité et de la quantité de vérité de terrain fournies lors de l'apprentissage.

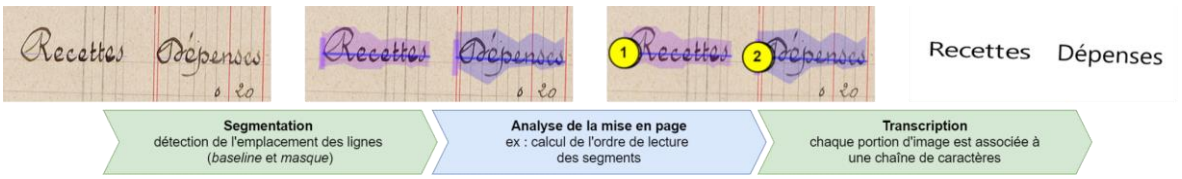


Figure 1 : Schématisation des étapes de traitement impliquées dans l'HTR, image produite par les auteur·rices.

De nombreux facteurs font que la tâche de transcription constitue encore un défi [Stokes et al., 2021]. On peut citer la très grande variation dans la formation des lettres (variation intra- et inter-classe), la forte présence de bruit et d'accidents sur les pages manuscrites, l'impossibilité de s'appuyer sur une segmentation à l'échelle des caractères ou encore la présence de graphèmes et de systèmes d'abréviations propres à chaque personne (Figure 2). S'y ajoute la difficulté pour les annotateurs et annotatrices de se mettre d'accord sur les pratiques de transcription, notamment la manière de traiter les variations graphétiques [Stutzmann, 2011] ou les abréviations.

Figure 2 : Exemples illustrant les principales difficultés rencontrées pour le traitement de textes manuscrits, image produite par les auteur·rices.

Malgré ces défis, il est déjà actuellement possible d'obtenir des modèles produisant des transcriptions à 95% réussies [Pinche 2021]. Pour produire de tels modèles, il existe deux approches (Figure 3). La première configuration s'apparente à un démarrage à froid : on part de zéro et on entraîne un modèle en fournissant de la vérité de terrain à un moteur de transcription et en définissant une architecture neuronale et des hyper paramètres. Dans l'autre cas de figure, on s'appuie sur un modèle préalable que l'on affine. C'est-à-dire que l'on fournit

au moteur de transcription un modèle de base plus ou moins performant et dont on reprend les paramètres pour lancer un nouvel entraînement basé sur des exemples plus ou moins similaires à ceux qui ont permis l'entraînement du modèle initial. Cette deuxième approche présente des avantages, parmi lesquels un important gain d'efficacité : en s'appuyant sur les acquis préalables d'un modèle, on a besoin d'une moindre quantité de vérité de terrain pour obtenir de bonnes, voire de meilleures performances. Cela signifie qu'au lieu de devoir transcrire manuellement une centaine de pages issues d'un corpus nouveau, on peut se contenter d'une trentaine de pages tout en parvenant *in fine* aux mêmes performances [Reul et al., 2021].

Dans les deux cas toutefois, en matière de transcription automatique pour les écritures manuscrites, il est rare de pouvoir se passer d'un entraînement sur des exemples correspondant au corpus à traiter, à l'inverse de la transcription automatique de l'imprimé ou bien de la segmentation, où les modèles disponibles sont déjà suffisamment performants. Cela signifie qu'il est presque toujours nécessaire de commencer par la transcription manuelle d'un échantillon du corpus.

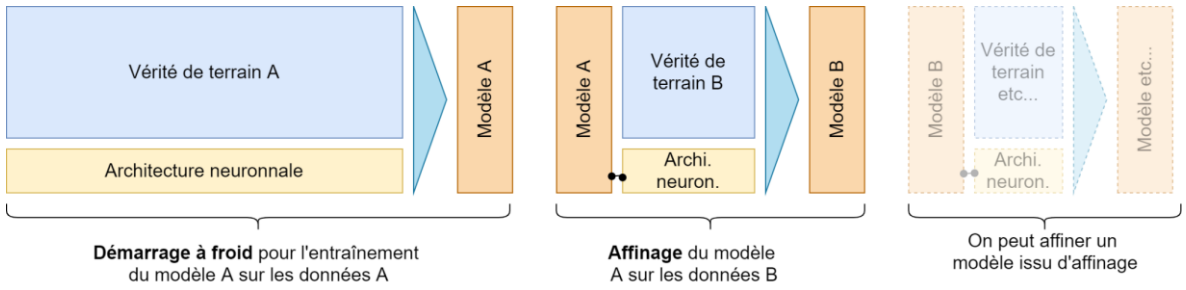


Figure 3 : Les deux configurations principales pour l'entraînement de modèles d'HTR ne demandent pas la même quantité de données, image produite par les auteur·rices.

Décloisonner les ressources, partager les données

Produire de la vérité de terrain de qualité suppose de posséder une connaissance minimale des environnements de transcription automatique, de mettre à plat l'ensemble des règles de transcription permettant d'obtenir une sortie de texte correspondant aux attentes, et surtout de posséder les moyens en temps et financiers de produire cette vérité de terrain. Il faut alors s'assurer de la disponibilité de personnes capables de lire les écrits du corpus, qu'elles disposent d'une compréhension suffisante des documents et des enjeux du projet, et qu'elles soient capables de contrôler la qualité de la transcription par rapport aux règles fixées en amont. Il est rare qu'un projet en humanités numériques souhaitant recourir à la transcription automatique possède toutes ces ressources.

Un écueil guette ces projets, celui d'échouer à obtenir un modèle de transcription efficace et d'abandonner alors l'ambition initiale d'automatiser cette tâche. Le risque est de dédier un temps considérable à produire une transcription qui, d'échantillon de vérité de terrain, finit par devenir le corpus final. Si on prend un peu de hauteur, c'est aussi une perte de ressources pour la

communauté des sciences humaines car à l'heure actuelle la vérité de terrain n'est souvent produite que pour servir les finalités d'un projet précis. Il s'opère donc un cloisonnement des ressources où chaque projet part de zéro et doit produire, ou tenter de produire, sa propre vérité de terrain comme si rien n'avait été transcrit avant.

Plusieurs facteurs expliquent le cloisonnement de ces données. En tout premier lieu, on peut considérer que les modèles de transcription (ou de segmentation) sont la partie la plus spectaculaire d'un projet de transcription automatique car ils permettent à n'importe qui, quelle que soit sa familiarité avec les principes de l'apprentissage profond, de produire immédiatement une sortie textuelle plus ou moins correcte. En plus, il est possible de communiquer les performances théoriques d'un modèle car elles peuvent être exprimées sous la forme de scores faciles à comprendre : on mesure notamment le taux d'erreur par caractère (CER pour *character error rate*) et le taux d'erreur par mot (WER pour *word error rate*). Comme les logiciels d'HTR permettent aux utilisateurs et utilisatrices de partager des modèles, il semble plus évident de baser sa communication sur le(s) modèle(s) entraîné(s) et employé(s) ainsi que sur la transcription finale obtenue (où les transcriptions manuelles sont souvent mêlées aux transcriptions obtenues automatiquement). A l'inverse, il est peut-être difficile de partager la vérité de terrain : outre parfois l'absence de fonctionnalités adéquates dans les logiciels, les droits sur les images peuvent restreindre les possibilités de partage, les transcriptions peuvent faire l'objet d'embargo et les compétences et capacités de calcul sont encore rares pour (re)produire des modèles à partir de ces données.

Lorsqu'elles sont publiées en tant que vérité de terrain, les données d'entraînement sont peu visibles car il n'existe aujourd'hui, au niveau national comme international, ni entrepôt, ni modèle de description dédiés. Pour les personnes susceptibles de réutiliser ces données, il n'est alors pas possible de filtrer rapidement les jeux correspondant à leurs projets. Par exemple, sur un entrepôt généraliste comme Zenodo, on peut chercher des jeux de données liés à l'HTR grâce aux mots-clefs associés aux dépôts, mais il n'est pas garanti que leur description suffise pour évaluer leur qualité. Les autres entrepôts généralistes, y compris Nakala, souffrent de ce même problème : les données HTR y côtoient des données diverses (modèles 3d, éditions de textes, données géographiques, transcriptions d'entretiens, etc.) et, à ce titre, leur recherche est difficile. Quelques tentatives d'énumération de jeux de données pour l'OCR et/ou l'HTR existent[1] mais elles sont imparfaites à plusieurs titres : aucune ne propose des descriptions suffisamment complètes, la cohérence des jeux de données n'est pas garantie, les critères de sélection ne sont pas explicites ; et elles reposent sur une veille de la part de leurs créateurs.

Pourtant s'appuyer sur la vérité de terrain plutôt que sur un modèle présente plusieurs avantages. Ils sont liés en tout premier lieu au fait qu'il est impossible de transposer un modèle d'un moteur de transcription à un autre. En effet, chaque solution d'HTR est basée sur un

[1] Par exemple, la liste des jeux de données d'OCR recensés par Awesome-OCR (Konstantin Baierer) sur Github (https://github.com/kba/awesome-ocr#datasets) ou encore le travail de Clemens Neudecker (https://cneud.github.io/ocr-gt/) (liens consultés le 12/02/2022)

écosystème de développement pour l'apprentissage machine[2] qui vient avec ses propres spécifications, et l'export et l'enregistrement des modèles dépend des choix de développement propre à chaque moteur. Certains logiciels grand public ne permettent même pas à l'utilisateur ou à l'utilisatrice d'exporter le modèle entraîné : celui-ci n'est alors disponible que par l'intermédiaire du logiciel qui l'a produit et du serveur qui l'héberge. Cette captivité des modèles rend les utilisateurs et utilisatrices, ainsi que leurs projets, vulnérables à l'arrêt des développements des logiciels ou de leurs dépendances et plus généralement face aux aléas informatiques. A l'inverse, les données d'entraînement s'avèrent plus souples, notamment du fait qu'il existe des standards ouverts comme XML ALTO [ALTO 4.2, 2020] et XML PAGE [Pletschacher & Antonacopoulos, 2010], que la plupart des moteurs d'HTR implémentent, pour enregistrer le résultat des différentes étapes de transcription. Un fichier XML étant essentiellement un fichier textuel, il est aisé de le lire ou de le modifier. Il est donc possible d'exporter les données produites à l'aide d'un logiciel de transcription, de les modifier au besoin, et de les réinjecter dans un autre logiciel de transcription.

Dans le cadre de la Science Ouverte, l'enjeu de la publication des transcriptions finales est certes compris, de même que celui de publier les modèles lorsque cela est possible, mais il manque un réflexe de publier la vérité de terrain en tant que vérité de terrain et non pas en tant que transcription. En fait, cela est d'autant plus dommageable que le fait de pouvoir accéder à la vérité de terrain permet de comprendre quelles ont été les pratiques de transcription conduisant à un modèle, d'en comprendre les résultats et même de reproduire l'entraînement du modèle[3].

Outre ces aspects de portabilité des données, il nous faut mentionner la plasticité de la vérité de terrain : il est impossible de fusionner des modèles de transcription, alors qu'on peut assembler, diviser, croiser différents jeux de vérité de terrain pour en recomposer un nouveau. De même, on peut modifier les exemples de transcription qu'ils contiennent de manière à rendre des jeux compatibles entre eux, ou pour obtenir un modèle dont la sortie correspond à nos besoins. On comprend alors qu'accéder à la vérité de terrain d'autres projets permet à coup sûr d'éviter un démarrage à froid : grâce à ces données, on peut créer son propre modèle pré-entraîné pour basculer dans un scénario d'affinage, ou bien augmenter rapidement l'importance matérielle de sa vérité de terrain de manière à réduire le temps passé à transcrire manuellement son corpus pour entraîner un premier modèle.

HTR-United : questions méthodologiques

Le projet HTR-United est né du constat qu'il faut mettre en commun la vérité de terrain pour permettre à chacun et chacune d'en bénéficier. Cela pose cependant de nombreuses questions méthodologiques que nous pouvons rappeler.

---

[2] Pour les systèmes basés sur Python, on peut citer PyTorch, Tensorflow ou encore DyNet.

[3] A condition que l'ensemble des paramètres de l'entraînement ait été documenté.

En premier lieu, le signalement, la documentation et les métadonnées. La description d'un jeu de données est cruciale pour permettre sa réutilisation par d'autres. En effet, on veut généralement savoir, par exemple, quelle est la langue utilisée dans les documents ou encore à quelle époque ils ont été rédigés. Ces informations permettent d'opérer un tri entre des lots de données pertinents pour composer une nouvelle vérité de terrain et ceux qui ne le sont pas. Au fil de nos réflexions sur l'écosystème HTR-United, nous avons élaboré un modèle de description[4] des jeux de données qui reprend notamment les champs suivants : la licence ; la langue ; le système d'écriture (ou alphabet) ; le nombre de mains[5] ou de polices et leur proportion ; la période couverte ; ou encore, l'importance matérielle (c'est-à-dire le volume). A ces éléments s'ajoutent les informations qui permettent d'identifier et de citer un jeu de données[6]. Le modèle de données fournit des indications sur la manière de remplir les champs correspondants en proposant, lorsque c'est possible, des listes fermées de valeurs. Cela permet de définir des pistes pour aboutir à une uniformisation des descriptions pour les cas complexes les plus courants. Par exemple, nous proposons une liste fermée pour décrire la manière dont les deux principaux états du texte sont représentés au sein d'un jeu de données (« *only-manuscript* », « *only-typed* », « *mainly-manuscript* », « *mainly-typed* » ou encore « *evenly-mixed* »), ou bien pour quantifier le nombre de mains représentées. HTR-United propose d'ajuster la quantification du nombre de mains à l'échelle des fichiers ou des dossiers avec des valeurs comme « *one-per-file* » (une main par fichier) ou « *one-per-folder* » (une main par dossier). Considérant d'une part que ce n'est pas le nombre exact de mains mais l'importance de la variation des écritures qui importe, et que d'autre part il n'est parfois pas possible de quantifier avec précision cette variation, nous proposons trois options : « *one* », « *few* » (10 mains ou moins) ou « *many* » (plus de 10 mains). La précision de cette quantification est indiquée par un autre champ dont les valeurs peuvent être « *exact* » ou « *estimated* ».

Évoquons en deuxième lieu les standards qui sont un autre aspect méthodologique à prendre en compte. PAGE et ALTO constituent au moins deux exemples de standards, mais il faut noter qu'ils se déclinent chacun en plusieurs versions. Faut-il s'en tenir à un standard et une version uniques, et si oui lesquels ? Est-il seulement possible de répondre à cette question alors que les logiciels continuent d'évoluer ? Par exemple, jusqu'à la publication de la version 1.5.0 de Transkribus en mars 2021, l'application de bureau (*desktop*) proposait d'exporter des données au format XML ALTO 2 et au format XML PAGE. Avec la version 1.5.0, le logiciel est soudainement passé à la version 4.2 d'ALTO, pour l'export et l'import des données, sans assurer

---

[4] HTR-United. (2021). HTR-United Schema (v. 15-10-2021). Alix Chagué & Thibault Clérice (éds.). URL : https://htr-united.github.io/schema/2021-10-15/schema.json (consulté le 10/02/2022)

[5] Une « main » correspond à l'écriture d'un individu, donc à une variation d'écriture.

[6] Nous proposons pour cela de fournir les informations (nom, prénom, rôle) permettant de citer l'ensemble des personnes ayant contribué à la création d'un jeu de vérité de terrain, notamment à travers les rôles « *transcriber* », « *aligner* », « *project-manager* » ou « *support* ».

de rétrocompatibilité avec ALTO 2[7]. On pourrait être tenté de penser que les modèles sont de ce point de vue plus robustes que les données, mais il faut noter qu'en janvier 2020, lorsque Kraken est passé à sa version 3, en permettant alors d'entraîner des modèles de segmentation en plus des modèles de transcription, tous les modèles produits avec les versions 2.x du logiciel ont cessé d'être compatibles avec les versions ultérieures.

Enfin, un troisième aspect méthodologique important : le contrôle de la qualité d'un jeu de données. Ce sont en général les objectifs du projet pour lequel un modèle est entraîné qui définisse la qualité attendue pour la vérité de terrain. Des invariants permettent toutefois d'établir plusieurs critères : l'homogénéité des règles suivies pour la production des données ; la fidélité de la transcription par rapport à l'image ; et sa capacité à s'adapter aux objectifs d'un autre projet. Dans un corpus de transcription comme celui créé à l'occasion de l'édition des journaux d'Eugène Wilhelm [Schlagdenhauffen, 2020] des passages rédigés en alphabet grec ont été transcrits en alphabet latin. Cela est justifié par le porteur de ce projet mais constitue néanmoins une transcription qui diffère de ce que l'image originale contient : cela peut poser un problème pour une réutilisation dans le cadre d'un projet prévoyant de produire un modèle capable de distinguer alphabet grec et alphabet latin. Cette vérité de terrain potentielle est-elle pour autant de mauvaise qualité ? Non. En fait, ce qui importe, c'est qu'a minima l'information concernant la pratique de transcription suivie soit documentée afin qu'elle puisse être prise en compte par une personne ré-utilisant de telles données. Idéalement, ce genre de problématique est pris en charge dans le cadre d'un Plan de Gestion des Données (PGD) qui décrit le contexte de production des données et les règles de transcription établies avant la campagne de transcription, ou bien actualisées durant sa conduite.

Un projet adossé à Github

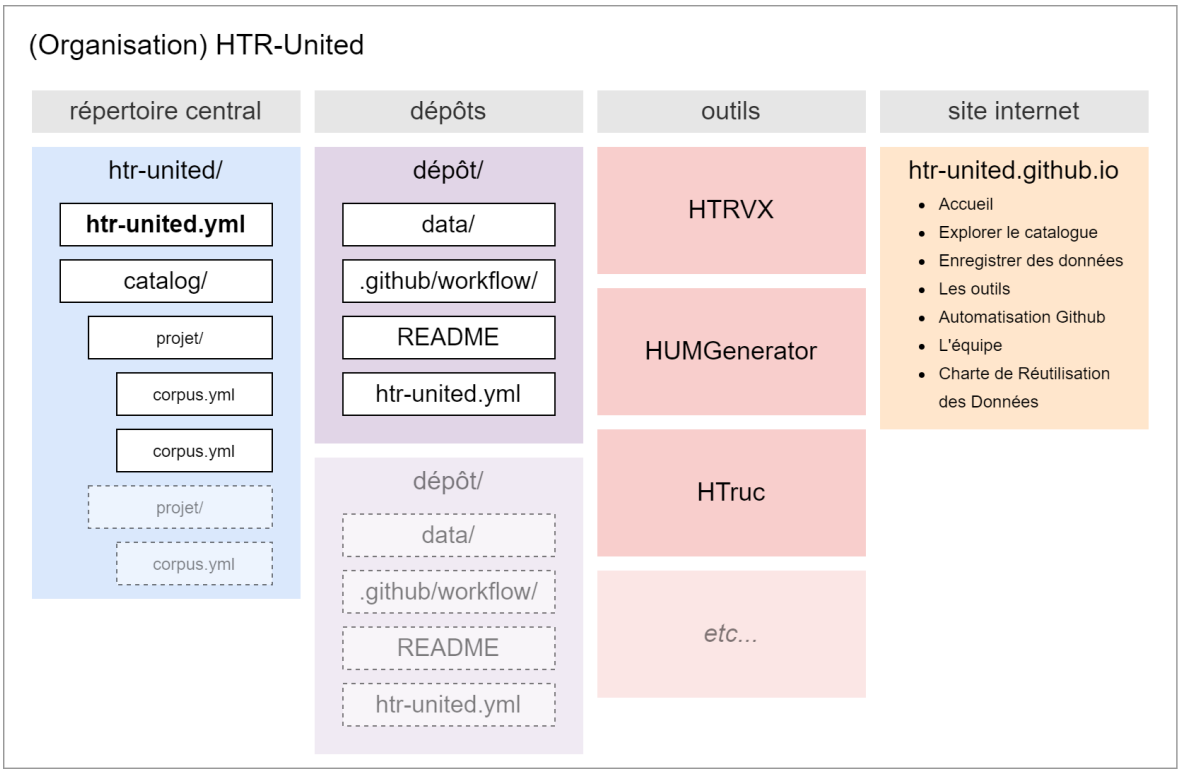
Le projet HTR-United a été mis en place sous la forme d'une organisation Github en octobre 2020. Il s'agit d'une entité propre à la plateforme permettant à plusieurs utilisateurs et utilisatrices de se rassembler autour d'un projet commun se déployant sur différents répertoires de travail. Le fait de s'appuyer sur une plateforme comme Github présente plusieurs avantages, dont la facilité d'y mettre en place un travail collaboratif dépassant les limites d'un projet donné ainsi que la possibilité de gérer finement plusieurs versions de travail et d'en faire coexister plusieurs simultanément.

L'organisation Github HTR-United est composée de plusieurs répertoires (Figure 4) dont le principal, nommé « htr-united » [HTR-United, 2020/2022], contient un catalogue prenant la forme d'un fichier de texte YAML (htr-united.yml) dont le contenu est généré automatiquement grâce au moissonnage des métadonnées fournies par les contributeurs et contributrices. Les contributions sont ajoutées selon deux modalités : soit sous la forme d'un fichier de métadonnées créé, dans le répertoire principal, dans un dossier nommé « catalog », soit sous la forme d'un répertoire à part, créé au sein de l'organisation Github HTR-United.

[7] Il est heureusement possible de mettre à niveau les jeux de données publiés avant mars 2021, à l'aide de script XSLT par exemple, sous réserve que toutes les informations attendues par ALTO 4 soient présentes.

Dans le dossier « catalog », un dossier est attribué à chaque projet contributeur, dans lequel on crée un fichier YAML par jeu de données. Par exemple, le projet e-ditiones a créé le corpus de vérité de terrain « OCR17+ » [Gabay et al., 2020 ; Jahan & Gabay, 2021], il existe donc dans le dossier « catalog », un dossier nommé « e-ditiones » contenant un fichier de métadonnées nommé « ocr17plus.yml ».

Dans les répertoires satellites, plusieurs éléments sont attendus : des transcriptions alignées avec des images, enregistrés dans un format standard comme XML PAGE ou XML ALTO ; les images ou bien les informations permettant d'accéder aux images (par exemple par un lien vers un manifeste IIIF) ; un document de présentation du corpus et de son contexte de production sous la forme d'un fichier « README.md », donnant autant d'informations que possible, y compris sur l'architecture du dossier de dépôt; et enfin un fichier de métadonnées intitulé « htr-united.yml »[8]. Notons que dans le cadre de la publication d'un *data paper* décrivant un jeu de données pour l'entraînement de modèles HTR, notre structure propose ainsi un modèle opérationnel pouvant servir de base à la structuration des informations[9]. Nous définissons un ensemble d'éléments de documentation qui sont cruciaux pour comprendre et réutiliser ces données.



[8] Nous proposons un modèle de répertoire de dépôt de vérité de terrain accessible via https://github.com/HTR-United/template-htr-united-datarepo (consulté le 16/03/2022).

[9] On peut se référer à l'exemple du gabarit proposé par le *Journal of Digital Humanities* pour les *data papers* [Hengchen & Pedrazzini, 2022]. Il propose des items de descriptions similaires à ceux demandés par HTR-United.

Figure 4 : Schématisation des répertoires rassemblés dans l'organisation HTR-United et de leurs contenus, image produite par les auteur·rices.

Les fichiers de métadonnées sont conformes au modèle de description évoqué plus haut. Un formulaire accessible depuis le site Internet du projet[10] aide à leur génération. Ce formulaire, ainsi que l'utilisation d'un format léger comme YAML, entend faciliter la création des descriptions par des personnes ne disposant pas des connaissances suffisantes pour produire des fichiers à structure plus complexe comme XML. YAML est en outre un format facile à analyser automatiquement : il permet de générer le catalogue principal recensant l'ensemble des dépôts, de contrôler la validité du contenu de certains champs, et surtout d'alimenter la page d'exploration du catalogue proposée sur le site Internet du projet[11].

Pour alimenter ces fichiers de métadonnées, il est possible d'automatiser le calcul des valeurs des champs liés à l'importance matérielle. C'est l'ambition de HTR-United Metadata Generator (HUM Generator) [Clérice & Chagué, 2021], un processus qui analyse les fichiers XML déposés afin de relever le nombre de pages, de lignes et de caractères constituant un lot de vérité de terrain. Combiner ces métriques est important car lorsqu'elles sont exprimées individuellement, elles ont peu de signification : le nombre de caractères dans une ligne et le nombre de lignes dans une page sont très variables en fonction des types de documents. Pour renseigner sur la taille réelle d'un lot de vérité de terrain, il faut donc les associer.

La transparence permise par Github, signifie que lorsqu'une personne publie sa vérité de terrain et la signale dans le catalogue HTR-United, sous réserve qu'elle soit libre de droit, une autre personne peut l'utiliser pour son projet et la citer, ou bien la mettre à jour ou la convertir pour la rendre compatible avec son logiciel et la re-publier comme une version alternative du jeu de données initial[12]. A l'échelle d'un dépôt, cela permet aussi de mettre en place un mécanisme de publication progressive de la vérité de terrain : il n'est pas nécessaire d'attendre la forme la plus aboutie du jeu de données pour le publier car on peut versionner le répertoire et mettre à jour progressivement le contenu des données ou bien sa documentation.

Un soutien au contrôle qualité

La production de vérité de terrain en HTR repose sur trois piliers : une production d'annotations –le texte, la segmentation–, sa formalisation –en XML, avec différents jeux de caractères–, et son intégration dans un réseau de production –à travers des ontologies de segmentation, des schémas et des choix d'encodage (*cf.* Figure 5). Si la première section relève principalement de données à vérification qualitative, l'ensemble des autres informations correspond à des

[10] Le formulaire est accessible via l'URL suivante : https://htr-united.github.io/document-your-data.html

[11] Cette page est accessible via l'URL suivante : https://htr-united.github.io/catalog.html (consulté le 16/03/2022).

[12] Pour comprendre le fonctionnement des forks dans Github : https://docs.github.com/en/get-started/quickstart/fork-a-repo [Github Inc., 2021]

éléments dont la validation peut être prise en charge par la machine. Dans ce cadre, afin d'assurer à la fois la qualité des données et réduire le temps passé à leur vérification formelle, HTR-United et le projet CREMMA travaillent à la mise à disposition de divers outils dits « d'intégration continue ».

L'intégration continue consiste au lancement automatisé et externalisé[13] de tests voire de compilations[14] au moment de la synchronisation d'un dépôt tel que ceux de Github[15] : elle permet par son caractère décentralisé de produire une vérification publique de la qualité des données et du code à chaque modification. Cette pratique reste encore assez rare dans le domaine des données en humanités numériques, mais connaît une progression sur les dernières années [Almas & Clérice, 2017 ; Ferger & Hedeland, 2020].

HTR-United propose l'utilisation de trois outils ayant chacun des objectifs partagés :

- ChocoMufin [Clérice & Pinche, 2021b],
- HTRVX [Clérice & Pinche, 2021a]
- Et *HTR United Metadata Generator (HUM Generator)* présenté plus haut (cf. Figure 6).

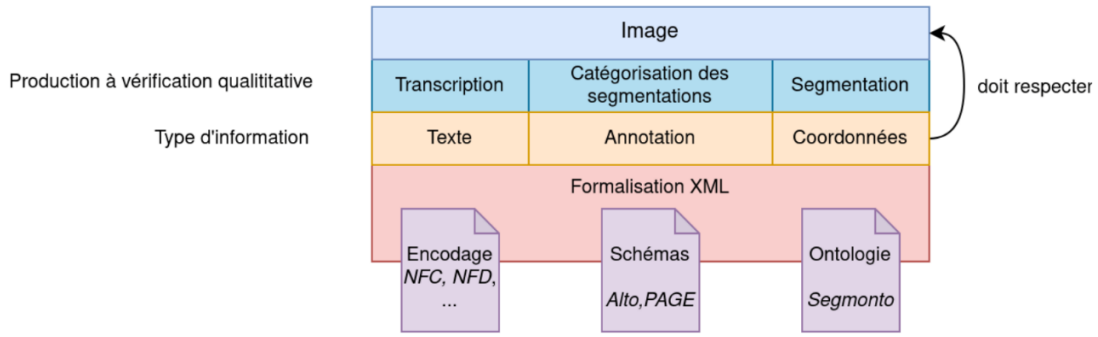


Figure 5 : Typologie des informations et leurs relations dans le cadre de vérité de terrain, image produite par les auteur·rices.

[13] Ces tests sont nécessairement lancés sur des machines vierges, permettant ainsi un test « objectif » des données : le même code est lancé indépendamment de toute particularité de paramétrage des ordinateurs de chacun et chacune.

[14] Dans notre cas, la phase compilation peut prendre la forme d'une normalisation automatisée ou de versionnage automatique du corpus.

[15] Github propose son propre service d'Intégration Continue, *Github Actions*, mais d'autres existent : TravisCI, CircleCI, etc.

ChocoMufin a été développé originellement dans le cadre du corpus CREMMA Médiéval [Pinche & Clérice, 2021] afin de traquer la variation dans l'encodage des caractères médiévaux. En effet, les manuscrits médiévaux présentent une très grande diversité d'abréviations utilisant différents signes additionnels, qu'ils soient « nouveaux » (7 [et], 9 [con]) ou non (macrons, barres obliques, etc.), de ligatures et de caractères (s connaît au moins trois variations principales : ß, ſ, ꝛ). Or, maintenir de la constance dans la transcription pour choisir le « bon » caractère peut s'avérer difficile. En outre, les pratiques diffèrent d'un projet à l'autre[16]. Afin de rendre ces pratiques « interopérables », ChocoMufin vérifie chacune des lignes transcrites en fonction d'une table de valeurs autorisées propre à chaque dépôt. Cette vérification s'accompagne d'une table des nouveaux caractères apparus, qui peuvent alors être inclus à la table des caractères validés ou au contraire corrigés. Par ailleurs, cette table des caractères contient aussi une valeur de remplacement, permettant aux utilisateur-rice-s de proposer des « simplifications » de leurs transcriptions, afin d'uniformiser les pratiques entre dépôts et projets.

Le deuxième outil est un simple outil de vérification des schémas, au format XSD. *HTRVX* s'appuie sur un schéma –nous fournissons uniquement un schéma pour l'ontologie SegmOnto [Gabay et al., 2021] pour le moment– et permet alors la vérification de la validité du fichier en fonction des catégories de segmentation proposées par *SegmOnto*. Nos schémas incluent aussi une vérification d'absence de lignes vides, qui auraient pu échapper à l'œil des annotateurs et annotatrices, soit parce que la ligne était difficile à percevoir et à l'origine d'une erreur de segmentation, soit parce qu'elle a tout simplement été oubliée. Chaque fichier fait alors l'objet d'un rapport individualisé avec un regroupement lisible de l'ensemble des erreurs rencontrées.

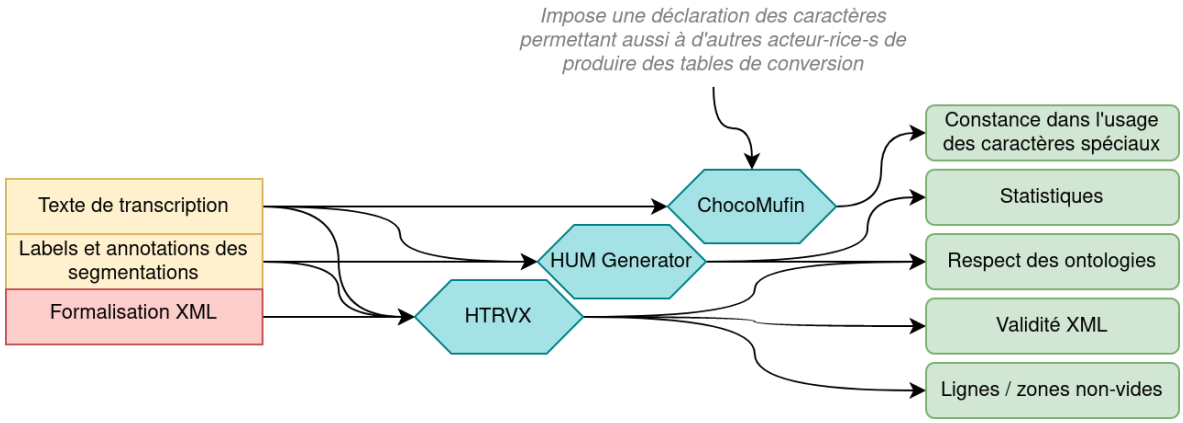


Figure 6 : Informations, outils de contrôle et résultat de ces contrôles, image produite par

[16] Ce problème dépasse la période médiévale : d'une part, il est encore commun de trouver des abréviations à la période moderne dans les documents manuscrits, mais on trouve encore après cette période des variations graphiques tels S Long / S ou des ligatures (les éditions de la Pléiade présentent encore des st en ligature). Les abréviations type numéro, les guillemets, etc. peuvent aussi faire l'objet de variations d'un annotateur à une autre.

les auteur·rices.

Ainsi, les personnes chargées de l'annotation et les gestionnaires de corpus de vérité de terrain réduisent le temps de maintenance et de recherche d'erreurs, en ne se concentrant que sur les logiques globales du corpus (normes de transcriptions, transcriptions) et en s'appuyant sur ces outils. Par ailleurs, les détails statistiques fournis par HUM Generator permettent de suivre la progression de la production de contenus voire de fournir des badges publics informant les personnes découvrant les corpus de l'état de ceux-ci au moment de leur visite (cf. Figure 7).

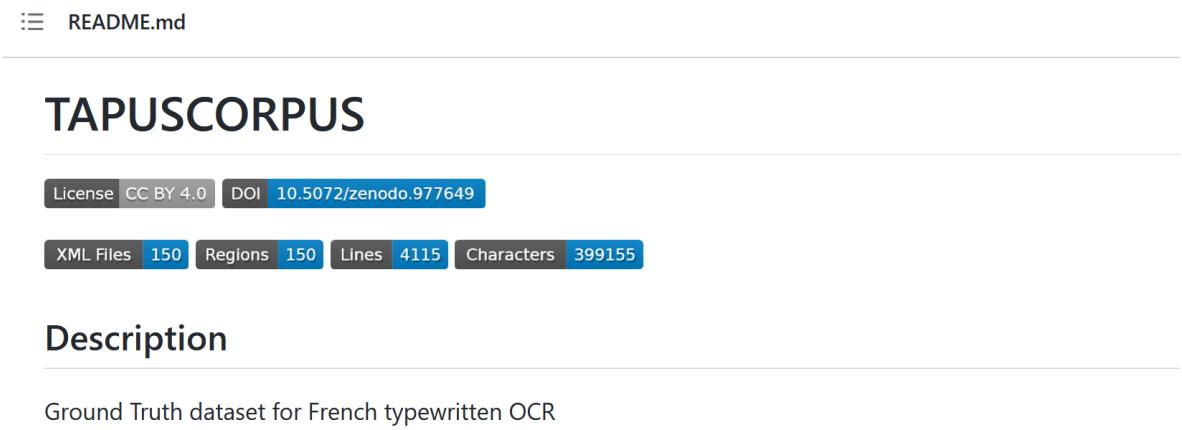


Figure 7 : Badges statistiques automatiquement mis à jour pour un corpus HTR United, image produite par les auteur·rices.

Conclusion

Mettre en commun la vérité de terrain est crucial pour permettre à la recherche d'avancer vers une meilleure intégration de la reconnaissance des écritures manuscrites dans les projets en humanités numériques. Une meilleure disponibilité des données passe par l'amélioration de la documentation sur leur contexte de production : on pense notamment aux règles de transcription mais aussi aux objectifs du projet ainsi qu'aux éventuelles omissions volontaires. En plus de la documentation présente dans les dépôts de données d'entraînement, le format de publication des *data papers* constitue un excellent moyen de rendre publiques les stratégies d'échantillonnage, les pratiques de transcriptions et les éventuelles spécificités de ces jeux de données. Les efforts de structuration de cette documentation, opérés au fil de l'élaboration de l'environnement HTR-United, encouragent à l'établissement de pratiques homogènes pour la construction et la description de ces données. Ces échanges –publics et ouverts à toute contribution de la part de la communauté des utilisateurs et utilisatrices de ces technologies–

pourraient finalement produire un modèle pour la rédaction de *data papers* spécifique aux données d'entraînement pour l'HTR.

Nous avons montré l'importance de mettre en place un recensement correct des jeux de données, passant notamment par l'établissement d'un modèle de métadonnées et d'outils facilitant son appropriation par des chercheurs et chercheuses aux profils variés. L'uniformisation des descriptions de la vérité de terrain permet d'interroger la communauté sur les critères permettant de filtrer efficacement plusieurs ensembles de données. Des initiatives similaires existent à d'autres niveaux : on peut mentionner à nouveau le projet SegmOnto, dont l'ambition est de produire des modèles de segmentation et d'analyse de mise en page basés sur des données très diversifiées mais suivant les mêmes règles d'annotation sémantique. L'objectif d'un tel projet est double : produire des modèles prêts à l'emploi et adaptés aux documents patrimoniaux manuscrits et imprimés, et partager des données permettant d'aboutir à ces modèles.

L'initiative que nous avons décrit s'appuie très fortement sur des mécanismes d'intégration continue et de versionnage qu'une plateforme telle que Github rend particulièrement propices. En plus d'inciter à la transparence des processus de constitution de la vérité de terrain, le paradigme adopté permet d'alléger la tâche de contrôle qualité en fixant des critères qu'il est possible d'adapter (absence ou tolérance des lignes ou des zones vides, conformité ou non au modèle sémantique SegmOnto, etc.) et en automatisant ces contrôles. S'appuyer sur une infrastructure privée n'est pas sans poser question, mais il existe fort heureusement des mécanismes d'archivage pérennes pour les répertoires Github, comme la Software Foundation, qui permettent à HTR-United d'offrir aux projets contributeurs des moyens de répondre aux objectifs d'accessibilité des principes FAIR. D'une manière générale, HTR-United contribue à la découvrabilité des jeux de données d'entraînement pour la transcription automatique, incite à l'emploi de standards garantissant l'interopérabilité et met en place les conditions de leur réutilisation par le biais du modèle de description implémenté. A mesure que la quantité de données signalées dans le catalogue grossit, on peut envisager que des institutions patrimoniales s'empareront de la question du recensement et de collecte de la vérité de terrain afin d'en pérenniser l'enregistrement.

Il est temps d'encourager la publication de ces données pour ce qu'elles sont : des données d'entraînement et pas seulement des transcriptions. Cela peut passer par la mise en place de chartes incitant au dépôt de la vérité de terrain en contrepartie de l'utilisation de ressources librement mises à disposition, comme ce sera par exemple le cas du serveur CREMMA, financé par le DIM MAP [DIM MAP, 2021]. Puisque nous visons de garantir une simplicité d'utilisation, HTR-United peut également être intégré dans les cursus universitaires qui forment à la transcription ou aux outils de versionnage. En effet, en réalisant une simple tâche d'alignement entre transcription et image, ou en mettant à jour un corpus, n'importe qui peut contribuer à cette initiative.

Entrepôt des données

Nous proposons un modèle de répertoire de dépôt de vérité de terrain accessible via https://github.com/HTR-United/template-htr-united-datarepo

Bibliographie

Almas, Bridget, et Thibault Clérice. 2017. « Continuous Integration and Unit Testing of Digital Editions ». *Digital Humanities Quarterly* 11 (4).

Analyzed Layout and Text Object (ALTO) (version v4.2). 2020. https://www.loc.gov/standards/alto/news.html#4-2-released.

Boillet, Mélodie, Marie-Laurence Bonhomme, Dominique Stutzmann, et Christopher Kermorvant. 2019. « HORAE: an annotated dataset of books of hours ». In *The 5th International Workshop on Historical Document Imaging and Processing*, 7-12. Sydney, Australia: ACM Press. https://doi.org/10.1145/3352631.3352633.

Chagué, Alix, Victoria Le Fourner, Manuela Martini, et Éric Villemonte de la Clergerie. 2019. « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? ». Présenté à DHNord 2019 « Corpus et archives numériques ». Lille. https://hal.inria.fr/hal-02448921.

Chagué, Alix, Lucas Terriel, et Laurent Romary. 2020. « Des images au texte : LECTAUREP, un projet de reconnaissance automatique d'écriture (poster) ». Présenté à DHnord 2020: The Measurement of Images. Computational Approaches in the History and Theory of the Arts, Lille. https://hal.archives-ouvertes.fr/hal-03008579.

Clérice, Thibault, et Alix Chagué. 2021. *HUM Generator, the HTR United Metadata Generator* (version 0.0.1). Python. https://doi.org/10.5281/zenodo.5363307.

Clérice, Thibault, et Ariane Pinche. 2021a. *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects* (version 0.0.4). Python. https://doi.org/10.5281/zenodo.5356154.

———. 2021b. *HTRVX, HTR Validation with XSD* (version 0.0.1). Python. https://doi.org/10.5281/zenodo.5359963.

DIM Matériaux anciens et patrimoniaux - PPSM (CNRS, ENS, Paris-Saclay). 2021. « Projets soutenus/CREMMA ». DIM MAP. 2021. https://www.dim-map.fr/projets-soutenus/cremma/.

Ferger, Anne, et Hanna Hedeland. 2020. « Towards Continuous Quality Control for Spoken Language Corpora ». *International Journal of Digital Curation* 15 (1). https://doi.org/10.2218/ijdc.v15i1.601.

Gabay, Simon, Jean-Baptiste Camps, Ariane Pinche, et Claire Jahan. 2021. « 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. https://hal.archives-ouvertes.fr/hal-03336528.

Gabay, Simon, Thibault Clérice, et Christian Reul. 2020. « OCR17: Ground Truth and Models for 17th c. French Prints (and hopefully more) ». Preprint. https://hal.archives-ouvertes.fr/hal-02577236.

Github Inc. 2021. « Fork a Repo ». Documentation. GitHub Docs. 2021. https://docs.github.com/en/get-started/quickstart/fork-a-repo.

Hengchen, Simon, et Nilo Pedrazzini. « JOHD Data Paper Template », 2022. https://www.overleaf.com/latex/templates/johd-data-paper-template/mqcypcbntsds.

HTR-United. (2020) 2022. *HTR-United Catalog*. YAML. https://github.com/HTR-United/htr-united/blob/026e680323b47f6206a6d6007cb96d6cc756fab5/htr-united.yml.

HTR-United, Alix Chagué, et Thibault Clérice. (2020) 2021. *HTR-United: Ground Truth Resources for the HTR of patrimonial documents*. https://github.com/HTR-United/htr-united.

Jahan, Claire, et Simon Gabay. (2021) 2021. *OCR17 +* (version 1.0). https://github.com/e-ditiones/OCR17plus.

Kahle, P., S. Colutto, G. Hackl, et G. Mühlberger. 2017. « Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents ». In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 04:19-24. https://doi.org/10.1109/ICDAR.2017.307.

Kiessling, Benjamin. (2015) 2021. *mittagessen/kraken: 3.0.5* (version v3.0.5). Python. https://github.com/mittagessen/kraken.

Mariotti, Viola. 2020. « Transcription automatique des feuillets du Manuscrit du Roi ». Billet. *ANR Maritem* (blog). 19 octobre 2020. https://maritem.hypotheses.org/193.

Massot, Marie-Laure, Arianna Sforzini, et Vincent Ventresque. 2019. « Transcribing Foucault's handwriting with Transkribus ». *Journal of Data Mining and Digital Humanities* Atelier Digit_Hum. https://hal.archives-ouvertes.fr/hal-01913435.

Pinche, Ariane. 2021. *CREMMA Medieval, an Old French dataset for HTR and segmentation* (version 1.0.1 Bicerin). https://doi.org/10.5281/zenodo.5235186.

Pinche, Ariane, et Thibault Clérice. 2021. *HTR-United/cremma-medieval: 1.0.1 Bicerin (DOI)* (version 1.0.1). Zenodo. https://doi.org/10.5281/ZENODO.5235186.

Pletschacher, Stefan, et Apostolos Antonacopoulos. 2010. « The PAGE (Page Analysis and Ground-Truth Elements) Format Framework ». In *2010 20th International Conference on Pattern Recognition*, 257-60. https://doi.org/10.1109/ICPR.2010.72.

Reul, Christian, Christoph Wick, Maximilian Nöth, Andreas Büttner, Maximilian Wehner, et Uwe Springmann. 2021. « Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning ». In *Proceedings of ACM Conference (HIP'21 (submitted to))*, 6. New York, NY, USA: ACM. http://arxiv.org/abs/2106.07881.

Schlagdenhauffen, Régis. 2020. « Optical Recognition Assisted Transcription with Transkribus: The Experiment Concerning Eugène Wilhelm's Personal Diary (1885-1951) ». *Journal of Data Mining and Digital Humanities* Atelier Digit_Hum. https://hal.archives-ouvertes.fr/hal-02520508.

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, et El Hassane Gargem. 2021. « The eScriptorium VRE for Manuscript Cultures ». *Classics@ Journal* 18 (1). https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.

Stökl Ben Ezra, Daniel. 2021. « L'infrastructure eScriptorium de reconnaissance automatique d'écriture manuscrite (HTR) ». Présenté à Rendez-vous IIIF360 2021. https://projet.biblissima.fr/fr/infrastructure-escriptorium-reconnaissance-automatique-ecriture-manuscrite-htr.

Stutzmann, Dominique. 2011. « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? » In *Kodikologie und Paläographie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age*, édité par Franz Fischer, Christiane Fritze, et Georg Vogeler, BoD, 2:247-77. Schriften des Instituts für Dokumentologie und Editorik. BoD. https://halshs.archives-ouvertes.fr/halshs-00596970.

Stutzmann, Dominique, Jean-François Moufflet, et Sébastien Hamel. 2017. « La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique ». *Médiévales. Langues, Textes, Histoire* 73 (73): 67-96. https://doi.org/10.4000/medievales.8198.

Teklia. 2021. *Teklia/Arkindex: 0.15.4* (version v0.15.4). https://teklia.com/solutions/arkindex/releases/0-15-4/.