HTR-United : Mutualisons la vérité de terrain !

Alix Chagué, Thibault Clérice, Laurent Romary

HTR-United : Mutualisons la vérité de terrain !

Alix Chagué, ALMAnaCH, Inria
Thibault Clérice, Centre Jean Mabillon, École nationale des chartes
Laurent Romary, ALMAnaCH, Inria

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, quoique disponibles, les plateformes techniques implémentant des solutions de transcription automatique nécessitent encore de grandes quantités de données. Produire ces données a un coût que la mutualisation des efforts peut atténuer. Nous présentons dans ce papier une solution nommée HTR-United facilitant la mise en commun de la vérité de terrain.

Principes de la transcription automatique

La reconnaissance des écritures manuscrites (REM), 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 :

1. localiser l'emplacement du texte sur l'image de manière à produire en ensemble de coordonnées (segmentation) ;
2. déterminer l'organisation logique de chaque segment par rapport aux autres et par rapport à la page (analyse de la mise en page) ;
3. à partir des coordonnées de chaque segment, reconnaître les lettres et les mots tracés dans la portion de l'image (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 et

corrigées 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), qui peut être des coordonnées dans le cas de la segmentation ou un ensemble de caractères pour la transcription. Les performances des modèles dépendent certes 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.

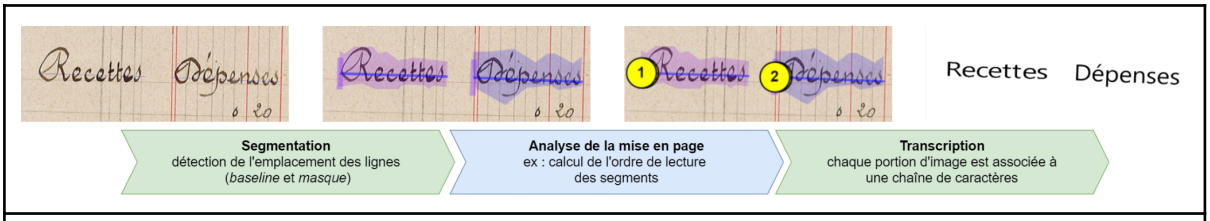

Figure 1 : Schématisation des étapes de traitement impliquées dans la REM, image produite par les auteur·rices.

De nombreux facteurs font que la tâche de transcription constitue encore un défi dans le domaine [Stokes et al., 2021]. On peut citer par exemple : la très grande variation dans la formation des lettres ; 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). A cela s'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. En dépit de cela, il est possible à l'heure actuelle d'obtenir des modèles produisant des transcriptions réussites à 95% [Pinche 2021].

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

Pour produire de tels modèles, on peut considérer qu'il existe deux approches (Figure 3) :

- une configuration qui s'apparente à un démarrage à froid : on part de zéro en générant un modèle à partir d'un jeu de vérité de terrain et d'une architecture neuronale ;
- ou au contraire, une configuration qui s'appuie sur l'affinage d'un modèle pré-entraîné : on ré-entraîne un modèle sur une architecture neuronale identique en ajoutant des données qui sont plus ou moins similaires à celles qui ont permis d'entraîner le modèle de départ.

La deuxième approche présente de nombreux 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. Au lieu de devoir transcrire une centaine de pages d'un corpus nouveau, on peut ainsi obtenir la même performance avec seulement 30 pages, si on affine un modèle [Reul et al., 2021].

Dans les deux cas toutefois, en matière de transcription automatique pour les écritures manuscrites, l'état de l'art est tel qu'on échappe rarement à la nécessité de produire sa propre vérité de terrain, à l'inverse de la transcription automatique de l'imprimé ou bien de la segmentation, où les modèles disponibles sont déjà suffisamment performants.

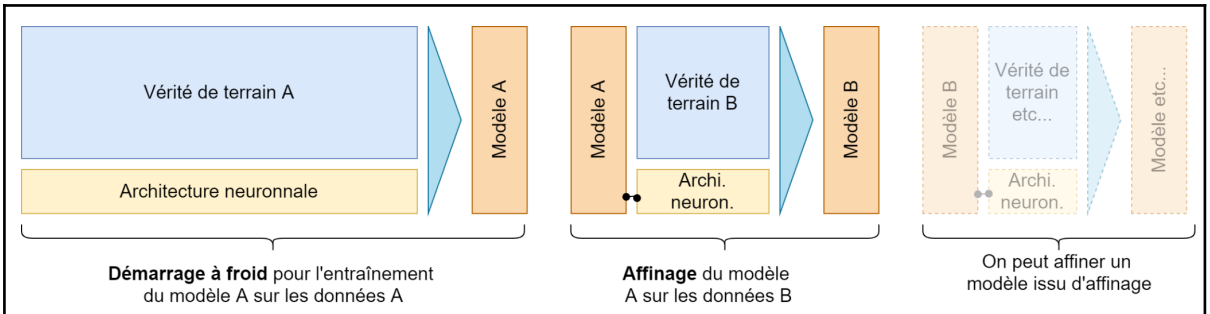

Figure 3 : Les deux configurations principales pour l'entraînement de modèles de REM ne demandent pas la même quantité de données, image produite par les auteur·rices.

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 REM possède toutes ces ressources.

Un écueil guette ces projets, celui d'échouer à obtenir un modèle de transcription efficace et d'abandonner l'ambition initiale d'automatiser cette tâche. On risque 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 la vérité de terrain n'est souvent produite que pour servir les finalités d'un projet précis. S'opère en effet 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 le manque d'attention portée à la vérité de terrain elle-même par rapport à celle portée aux modèles. Les logiciels permettent aux utilisateurs et utilisatrices de partager des modèles, qui ont l'avantage d'être immédiatement capables de produire une transcription en plus de fournir des informations sur leurs performances théoriques. A l'inverse, il est plus difficile de partager la vérité de terrain : outre parfois l'absence de fonctionnalités adéquates dans les logiciels, se posent des problèmes de droits sur les images, d'embargo sur les transcriptions, et de compétences et capacités de calcul pour entraîner de nouveaux modèles sur ces données.

Autre difficulté découlant de la première : trouver, en ligne, des jeux de vérité de terrain déjà constitués. En passant par des moteurs de recherche ou des plateformes de dépôt comme Zenodo, on peut trouver ce type de données, mais on n'a jamais la certitude que le jeu est complet, à jour, dans un format standard, ou compatible avec notre projet. Par exemple, même un corpus, pour l'imprimé, aussi récent et bien documenté que OCR17 [Gabay et al., 2020], recense des données dans un format appauvri incompatible avec la version actuelle de Kraken (3.0.5) [Kiessling, 2015/2021][1].

Pourtant s'appuyer sur la vérité de terrain plutôt que sur un modèle a du sens. Alors qu'il est impossible de transposer un modèle d'un moteur de transcription à un autre, les données s'avèrent plus souples. La raison en est simple : il n'y pas de pratique standardisée pour l'enregistrement des modèles de transcription puisque les architectures neuronales varient d'un système à l'autre et certains logiciels ne permettent tout simplement pas de télécharger les modèles. En revanche, il existe des standards pour la représentation des données et la plupart des logiciels de transcription les intègrent : XML ALTO [ALTO 4.2, 2020] et XML PAGE [Pletschacher & Antonacopoulos, 2010].

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, c'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[2].

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 présente de nombreux avantages et 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 de ce constat. Il faut mettre en commun la vérité de terrain pour permettre à chacun d'en bénéficier. Pourtant cela pose de nombreuses questions méthodologiques que nous pouvons rappeler.

---

[1] Le dataset OCR17+ [Jahan & Gabay, 2021] résout cela en proposant désormais des paires d'images et de fichiers XML ALTO.

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

En premier lieu, le signalement, la documentation et les métadonnées. Parmi les métadonnées qui nous semblent devoir accompagner un dépôt de vérité de terrain, outre celles qui permettent de l'identifier et de le citer[3], nous avons recensé les suivantes : la licence ; la langue ; le système d'écriture (ou alphabet) ; le nombre de mains ou de polices et leur proportion[4] ; la période couverte ; ou encore, l'importance matérielle (le volume). Ces éléments de description sont cruciaux pour faciliter le filtrage, par une personne porteuse d'un projet, entre plusieurs jeux de données. Ce filtrage se fait alors en fonction de critères permettant de composer une vérité de terrain nouvelle, susceptible de compléter les données que le projet possède déjà ou d'aboutir à un modèle idoine.

Outre la question de savoir quelles informations doivent être fournies, se pose également celle de savoir comment les renseigner. Si certains champs posent peu de problèmes, d'autres qui semblent au premier abord aller de soi s'avèrent difficiles à évaluer. Par exemple : le nombre de mains. Lorsque quantifier le nombre de mains représentées dans un corpus disparate s'avère impossible, on doit se contenter d'en donner au mieux une estimation. Il est en fait préférable d'exprimer cette information en suivant deux modalités : soit une quantification précise lorsque l'on connaît l'identité des scripteurs, soit des mot-clefs comme "few" ou "many" pour exprimer un ordre de grandeur pertinent[5]. En effet, ce qui importe vraiment ce n'est pas le nombre exact, mais de savoir si un jeu de vérité de terrain ne contient qu'une seule main, s'il existe un peu de variation dans ce lot ou bien si au contraire les écritures y sont très disparates, en grand nombre et parfois difficiles à distinguer.

Parmi les autres aspects méthodologiques à considérer : les standards. On a mentionné PAGE et ALTO. Il en existe donc au moins deux, mais qui 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 desktop proposait d'exporter des données au format XML ALTO 2 et au format XML PAGE. Soudainement, le logiciel est passé à la version 4.2 d'ALTO, pour l'export puis pour l'import, sans rétrocompatibilité. Cela rend-il caduque les nombreux jeux de données déposés et publiés sur Zenodo avant mars 2021 ? On pourrait arguer que les modèles sont de ce point de vue plus robustes que les données mais il n'en est rien. En janvier 2020, lorsque Kraken est passé à sa version 3, permettant l'entraînement de modèles de segmentation, ce qui n'était pas possible avant, tous les modèles produits avec des versions antérieures du logiciel ont cessé d'être compatibles avec les versions ultérieures.

Enfin, troisième aspect méthodologique important : le contrôle de la qualité d'un jeu de données. Ce qui définit la qualité d'une vérité de terrain dépend des objectifs de la personne par rapport au modèle qu'elle en tire. On peut toutefois s'accorder à prendre en compte plusieurs critères :

---

[3] 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*" et "*support*".

[4] HTR-United propose plusieurs valeurs telles que "*only-manuscript*", "*only-typed*", "*mainly-manuscript*", "*mainly-typed*" ou encore "*evenly-mixed*".

[5] 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).

- l'homogénéité des règles de transcription suivies pour produire ces 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 l'auteur de l'article qui rend compte de ce travail, mais constitue de fait une transcription qui diffère de ce que l'image originale contient et peut ne pas être adaptée à d'autres projets où le modèle doit distinguer alphabet grec et alphabet latin. Pour autant, cette vérité de terrain potentielle est-elle de mauvaise qualité ? Ce qui importe en fait, c'est qu'a minima l'information concernant la pratique 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).

Un projet adossé à Github

Le projet HTR-United s'est 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. L'organisation Github HTR-United est donc composée de plusieurs répertoires (Figure 4) dont les principaux sont :

- un répertoire central [Chagué et al., 2021] contenant :
    - un catalogue sous la forme d'un fichier YAML unique et généré automatiquement grâce au moissonnage des métadonnées fournies par chaque sous-projet ;
    - un sous-dossier de signalement nommé "catalog" rassemblant les métadonnées de projets hébergés en dehors de Github[6], par exemple sur Zenodo, mais signalés dans HTR-United ;
- une série de répertoires satellites permettant le dépôt des jeux de données directement sur Github s'ils n'ont pas été déposés ailleurs.

Pour être complet, un dépôt de données devrait comporter :

- des transcriptions alignées avec des images, dans un format standard comme XML PAGE ou XML ALTO ;
- des images soit directement dans le dépôt, soit sous la forme de liens vers des ressources hébergées ailleurs, par exemple sur un serveur IIIF ;
- un fichier de présentation du corpus et de son contexte de production sous la forme d'un fichier README, donnant autant d'informations que possible, y compris sur l'architecture du dossier de dépôt, à la manière d'un data paper ;
- et enfin un fichier YAML de métadonnées correspondant aux champs identifiés plus haut et requis par HTR-United. Ce fichier se nomme "htr-united.yml" dans les dépôts réalisés directement sous l'organisation Github, ou bien porte le nom du corpus lorsqu'il se situe dans le sous-dossier "catalog".

[6] Ainsi HTR-United n'impose pas qu'images et transcriptions soient déposées sur Github.

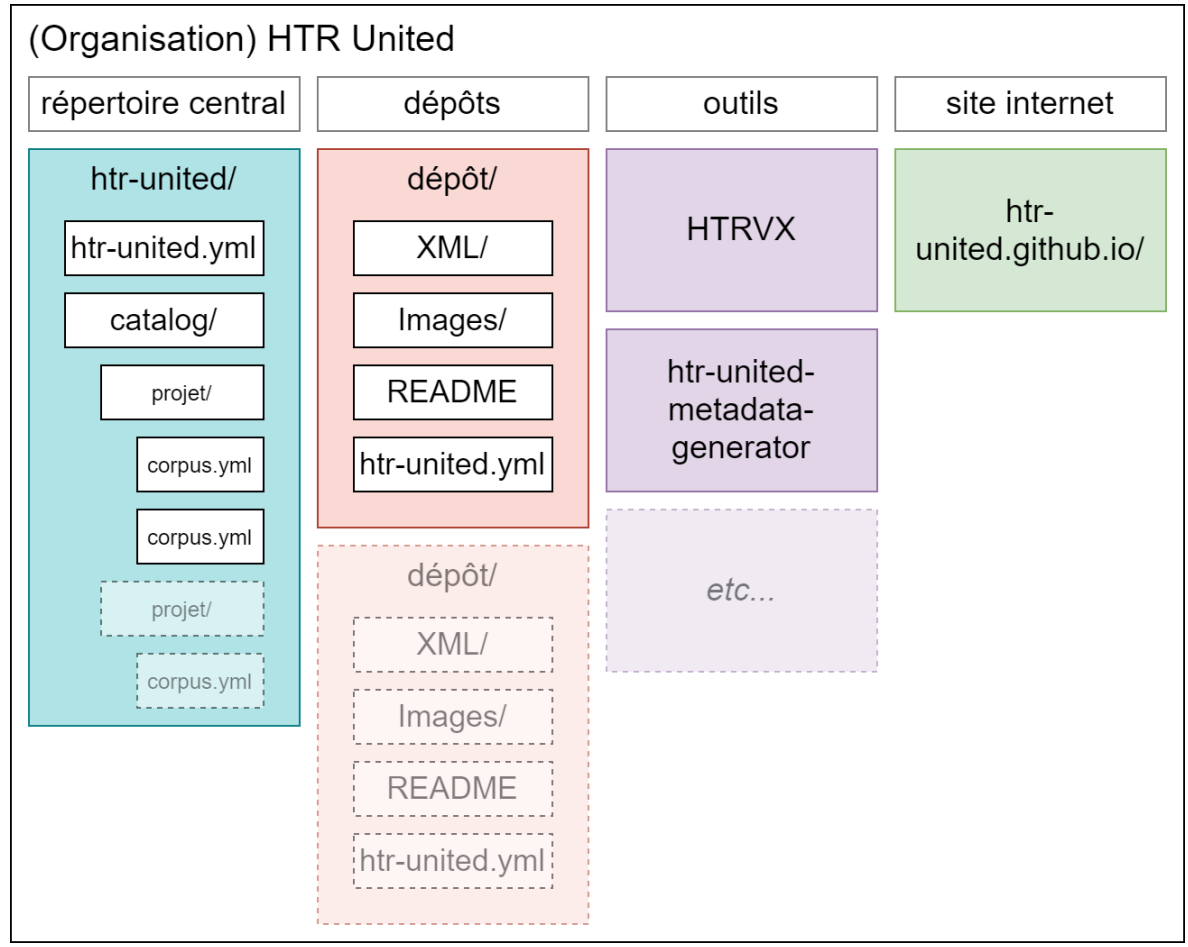

Figure 4 : Schématisation des répertoires rassemblés dans l'organisation HTR-United, image produite par les auteur·rices.

Un formulaire[7] aide à la génération automatique des fichiers YAML en permettant notamment de normaliser les valeurs attribuées à certains des champs mentionnés plus tôt. Ce formulaire, ainsi que l'utilisation d'un format léger comme YAML, entend faciliter leur création 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 : cela permet de générer le catalogue central recensant l'ensemble des dépôts, de contrôler la validité du contenu de certains champs, voire d'en générer automatiquement les valeurs.

On peut ainsi automatiser le calcul des valeurs des champs liés à l'importance matérielle. C'est l'ambition de HTR-United Metadata Generator (HUM) [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. Ces métriques sont importantes car lorsqu'elles sont exprimées individuellement, elles ont peu de signification. Pour renseigner sur la taille réelle d'un lot de vérité de terrain, il faut les croiser, le nombre de caractères dans une ligne et le nombre de lignes dans un page étant extrêmement variables en fonction des types de document.

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

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 de recherche donné ; la gestion fine des versions ; et la possibilité d'en faire coexister plusieurs simultanément[8].

Cela 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, la mettre à jour ou la convertir pour la rendre compatible avec son logiciel et la re-publier comme un *fork* du jeu de données initial. A l'échelle d'un dépôt, cela veut aussi dire qu'une personne déposant un lot de vérités de terrain peut, plus tard, augmenter ce lot, mettre à jour la documentation et ainsi en publier une nouvelle version.

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 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 de 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[9] de tests voire de compilations[10] au moment de la synchronisation d'un dépôt tel que ceux de Github[11] : 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).

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

[9] 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-e des contributeurs-ices.

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

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

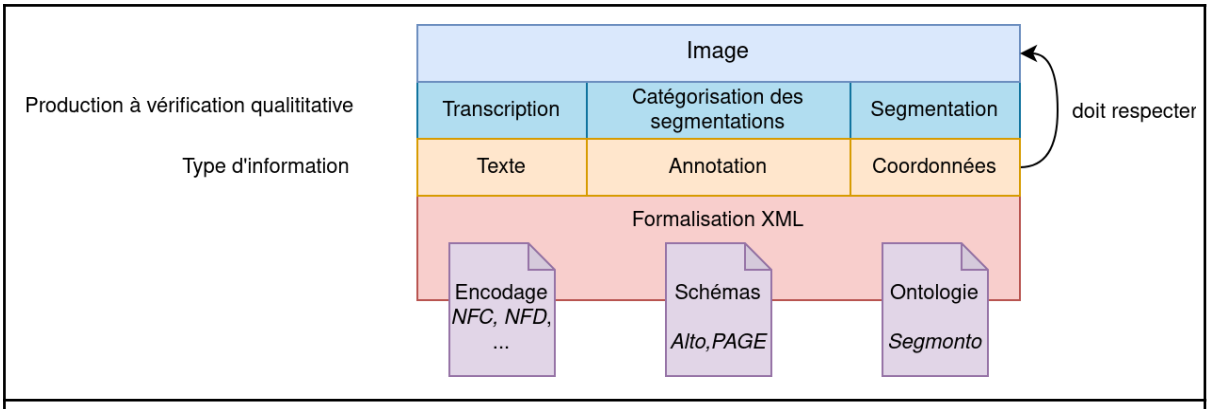

Figure 5 : Typologie des informations et leurs relations dans le cadre de vérité de terrain, image produite par les auteur·rices.

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" (ꝛ [et], [con]) ou non (macrons, barres obliques, etc.), de ligatures et de caractères (s connaît au moins trois variations principales : ß, ſ,). 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[12]. 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] et la spécification ALTO 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 annotateur·rices, 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.

[12] 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 Pléïade 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.

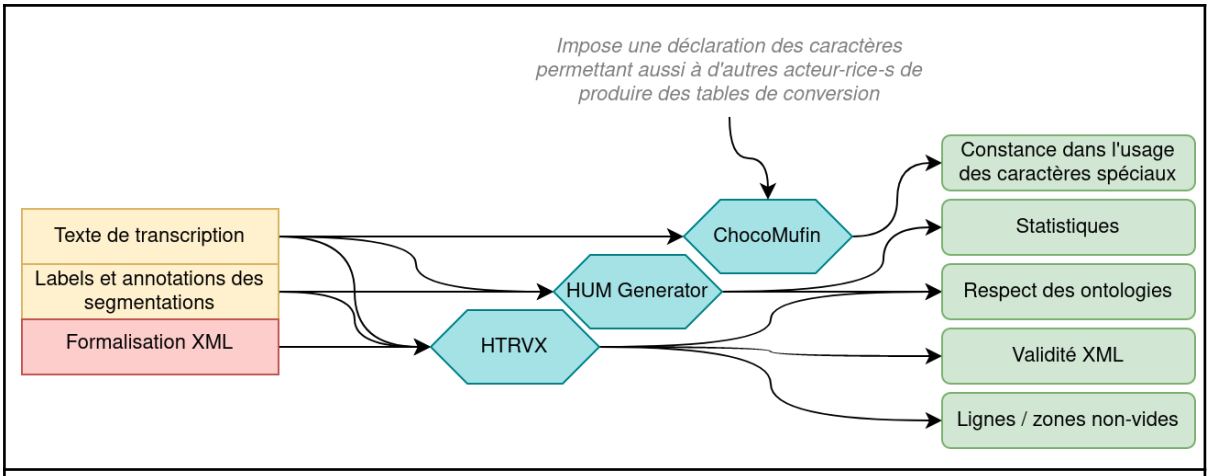

Figure 6 : Informations, outils de contrôle et résultat de ces contrôles, image produite par les auteur·rices.

Ainsi, les annotateur·rices et 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).

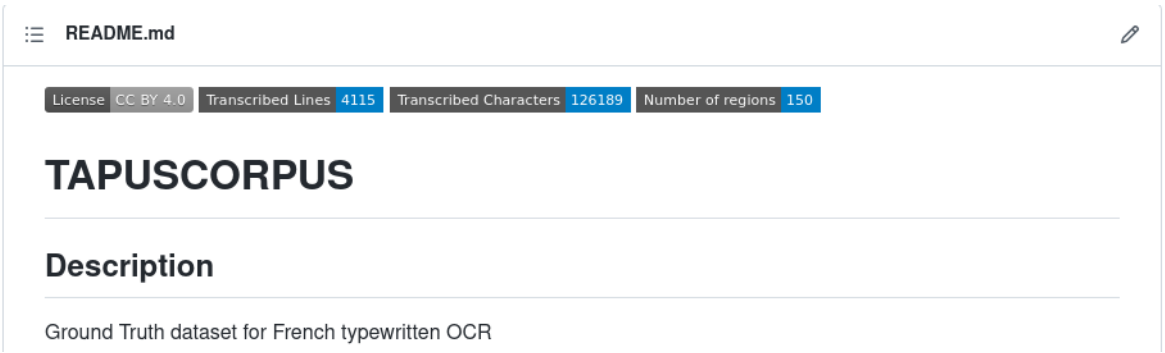

Figure 7 : Badges statistiques automatiquement mis à jour pour un corpus HTR United, image produite par les auteur·rices.

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. Outre la disponibilité de données de qualité, correctement documentées et recensées, cette mutualisation des ressources peut pousser l'ensemble des utilisateur·rices à établir progressivement des pratiques homogènes. Cela concerne autant les règles de transcription que les métadonnées permettant ce partage. On peut noter des initiatives comme SegmOnto, dont l'objectif 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.

Un pot commun s’appuyant sur une plateforme initialement orientée vers le versionnage telle que Github présente de nombreux avantages, mais l’on peut envisager que des institutions patrimoniales s’emparent progressivement de la tâche de 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. 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]. En cherchant à 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.

Bibliographie

Almas, B., & Clérice, T. (2017). Continuous Integration and Unit Testing of Digital Editions. *Digital Humanities Quarterly*, *11*(4).

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

Boillet, M., Bonhomme, M.-L., Stutzmann, D., & Kermorvant, C. (2019). HORAE : An annotated dataset of books of hours. *the 5th International Workshop on Historical Document Imaging and Processing*, 7‑12. https://doi.org/10.1145/3352631.3352633

Chagué, A., Clérice, T., & Chiffoleau, F. (2021). *HTR United, a centralization effort of HTR and OCR ground-truth repositories for French languages*. https://github.com/HTR-United/htr-united (Original work published 2020)

Chagué, A., Le Fourner, V., Martini, M., & de la Clergerie, É. (2019, octobre 16). *Deux siècles de sources disparates sur l'industrie textile en France : Comment automatiser les traitements d'un corpus non-uniforme ?* Colloque DHNord 2019 « Corpus et archives numériques ». https://hal.inria.fr/hal-02448921

Chagué, A., Terriel, L., & Romary, L. (2020, novembre). *Des images au texte : LECTAUREP, un projet de reconnaissance automatique d'écriture*. https://hal.archives-ouvertes.fr/hal-03008579

Clérice, T., & Chagué, A. (2021). *HUM Generator, the HTR United Metadata Generator* (v0.0.1) [Python]. https://doi.org/10.5281/zenodo.5363307

Clérice, T., & Pinche, A. (2021a). *HTR-United/HTRVX : 0.* (v0.0.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5359963

Clérice, T., & Pinche, A. (2021b). *PonteIneptique/choco-mufin : 0.0.4* (v0.0.4) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5356154

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

Ferger, A., & Hedeland, H. (2020). Towards Continuous Quality Control for Spoken Language Corpora. *International Journal of Digital Curation*, *15*(1), Article 1. https://doi.org/10.2218/ijdc.v15i1.601

Gabay, S., Camps, J.-B., Pinche, A., & Jahan, C. (2021, septembre). *SegmOnto : Common vocabulary and practices for analysing the layout of manuscripts (and more)*. 16th International Conference on Document Analysis and Recognition (ICDAR 2021), Lausanne, Switzerland. https://hal.archives-ouvertes.fr/hal-03336528

Gabay, S., Clérice, T., & Reul, C. (2020). *OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more)*. https://hal.archives-ouvertes.fr/hal-02577236

Github Inc. (2021). *Fork a repo* [Documentation]. GitHub Docs. https://docs.github.com/en/get-started/quickstart/fork-a-repo

Jahan, C., & Gabay, S. (2021). *OCR17 +* (1.0) [Python]. https://doi.org/none (Original work published 2021)

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

Kiessling, B. (2021). *mittagessen/kraken : 3.0.5* (v3.0.5) [Python]. https://github.com/mittagessen/kraken (Original work published 2015)

Mariotti, V. (2020, octobre 19). Transcription automatique des feuillets du Manuscrit du Roi [Billet]. *ANR Maritem*. https://maritem.hypotheses.org/193

Massot, M.-L., Sforzini, A., & Ventresque, V. (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, A. (2021). *CREMMA Medieval, an Old French dataset for HTR and segmentation* (1.0.1 Bicerin (DOI)) [XSLT]. https://doi.org/10.5281/zenodo.5235186

Pinche, A., & Clérice, T. (2021). *HTR-United/cremma-medieval : 1.0.1 Bicerin (DOI)* (1.0.1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.5235186

Pletschacher, S., & Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. *2010 20th International Conference on Pattern Recognition*, 257‑260. https://doi.org/10.1109/ICPR.2010.72

Reul, C., Wick, C., Nöth, M., Büttner, A., Wehner, M., & Springmann, U. (2021). Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning. *Proceedings of ACM Conference (HIP'21 (submitted to))*, 6. http://arxiv.org/abs/2106.07881

Schlagdenhauffen, R. (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, P. A., Kiessling, B., Stökl Ben Ezra, D., Tissot, R., & Gargem, E. H. (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, D. (2021, mars 24). *L'infrastructure eScriptorium de reconnaissance automatique d'écriture manuscrite (HTR)*. Rendez-vous IIIF360. https://projet.biblissima.fr/fr/infrastructure-escriptorium-reconnaissance-automatique-ecriture-manuscrite-htr

Stutzmann, D. (2011). *Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?* 247.

Stutzmann, D., Moufflet, J.-F., & Hamel, S. (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* (v0.15.4) [Computer software]. https://teklia.com/solutions/arkindex/releases/0-15-4/