Reconnaissance des écritures dans les imprimés

Simon Gabay, Thibault Clérice, Pauline Jacsont, Elina Leblanc, Marie Jeannot-Tirole, Sonia Solfrini, Sophie Dolto, Floriane Goy, Carmen Carrasco Luján, Maddalena Zaglio, et al.

Reconnaissance des écritures dans les imprimés. CATMuS *print* : un modèle générique, multilingue et diachronique

Simon Gabay[1], **Thibault Clérice**[2], **Pauline Jacsont**[1], **Elina Leblanc**[1], **Marie jeannot-Tirole**[3], **Sonia Solfrini**[1], **Sophie Dolto**[1], **Floriane Goy**[1], **Carmen Carrasco Luján**[1], **Maddalena Zaglio**[1], **Myriam Perregaux**[1], **Juliette Janès**[2], **Benoît Sagot**[2], **Rachel Bawden**[2], **Rasul Dent**[2], **Oriane Nédey**[2] et **Alix Chagué**[2,4,5]

[1]Université de Genève
{prenom.nom}@unige.ch
[2]Inria
{prenom.nom}@inria.fr
[3]Université de Strasbourg
{prenom.nom}@unistra.fr
[4]Université de Montréal
[5]École Pratique des Hautes Études, Paris

Résumé

La reconnaissance optique de caractères (OCR) a connu d'importants succès pour les documents manuscrits ou les imprimés anciens ces dernières années, mais ce type de document reste marginal dans la production textuelle aujourd'hui disponible. Afin d'offrir aux chercheur · se · s des modèles performants couvrant un plus grand large éventail de cas, nous avons conçu un nouveau modèle généraliste, capable de gérer au mieux des imprimés, anciens comme contemporains, écrits dans une pluralité de langues. Plusieurs architectures sont évaluées, afin de comparer leur efficacité respective en terme de taux d'erreur par caractère, mais aussi de temps d'inférence.

1 Introduction

Au début de l'année 2024, Gallica annonce avoir numérisé 863 729 livres et 20 352 titres de presse et revues, qui représentent 5 814 656 numéros [1]. Au vu de ces chiffres, il paraît évident que l'extraction du texte imprimé constitue un défi majeur pour la bibliothèque numérique française, comme pour toutes les autres à travers le monde.

Les dates des publications numérisées sont instructives : si un nombre important sont du XX^e^ siècle, environ 250 000 datent d'avant la Révolution française, témoignant de l'importance des documents historiques (tab. 1).

1. https://gallica.bnf.fr/GallicaEnChiffres.

Siècle	Livres	Presse Revues
XVI^e^	38 682	1
XVII^e^	73 307	54
XVII^e^	138 118	602
XIX^e^	356 548	8 123
XX^e^	241 385	11 339
XXI^e^	56 199	109

TABLEAU 1 – Nombre d'unités codicologiques par type et par siècle.

Si les fac-similés mis à disposition par la Bibliothèque nationale de France (BnF) sont bien évidemment majoritairement en français, des documents rédigés dans d'autres langues sont aussi présents dans les collections (cf. tab. 2), ne serait-ce que parce que beaucoup d'autres idiomes sont historiquement présents sur le territoire (corse, catalan, breton, allemand…) et que la France a connu nombre de publications par les communautés étrangères présentes sur son territoire (réfugiés espagnols, migrants italiens…). La situation est évidemment identique dans les autres bibliothèques francophones, souvent présentes dans des pays ayant plusieurs langues nationales (Suisse, Belgique, Canada…).

En regard de la composition des fonds des bibliothèques, il semble pertinent d'adopter des modèles multilingues et diachroniques, seuls à

Siècle	Livres	Presse Revues
Anglais	18 688	166
Allemand	55 514	297
Italien	12 305	137
Espagnol	3 103	113
Latin	65 707	0

Tableau 2 – Nombre d'unités codicologiques par type et par langue.

même d'acquérir le texte de la totalité des collections numérisées. De tels modèles étant encore inexistants, nous nous sommes attelés à la conception d'un premier prototype, que nous présentons ici.

2 État de l'art

La situation concernant l'OCRisation des imprimés récents est difficile à décrire. Si les principaux outils, comme Tesseract (Kay, 2007), OCRopy (Breuel, 2014), Pylaia (Puigcerver et Mocholí, 2018), HTR+ (Michael et al., 2018) ou encore Kraken (Kiessling, 2019) sont bien connus, les modèles qui sont proposés sont encore mal documentés.

Concernant les imprimés, une grande gamme de modèles par langues sont proposés pour Tesseract[2], et les tentatives de conception de modèles multilingues ne sont pas nouvelles (Smith et al., 2009 ; Etter et al., 2023). Transkribus propose lui aussi des modèles multilingues (16 langues) et diachroniques (XVI^{e}-XXI^{e}) entraînés avec Pylaia et affichant un taux d'erreur par caractère (CER, *Character Error Rate*) de 2,2% (Trankribus, 2022), et depuis peu des « Super modèles » comme *Text Titan* (Trankribus, 2023) s'appuyant sur des transformeurs (Vaswani et al., 2017).

De tels modèles, bien qu'extrêmement utiles, posent plusieurs problèmes. Aucune documentation ne décrit le contenu des données d'entraînement, les règles de transcription, les principaux types d'erreur produites... Sans compter que les modèles entraînés sur une plateforme comme Transkribus ne sont accessibles qu'à travers elle, limitant les métadonnées disponibles pour recomposer les conditions d'entraînement de ces modèles.

En ce sens, la démarche de Pinche et al. (2023a) nous paraît plus prometteuse, en décrivant précisément le modèle, distribué gratuitement en ligne (Pinche et al., 2023b), et adossé à une documentation solide concernant les choix de transcription (Pinche, 2022).

3 Données

Grâce au catalogue HTR-United (Chagué et al., 2021 ; Chagué et Clérice, 2023), il est possible de rassembler des données en garantissant une qualité et une documentation minimales des fichiers au moyen d'outils d'intégration continue (Clérice et al., 2023).

Le volume total du corpus rassemblé par l'intermédiaire d'HTR-United s'élève à près de 7 000 pages, pour presque 300 000 lignes (cf. tab. 3). Les données sont majoritairement en français, mais des imprimés en latin, anglais, italien, allemand, catalan, portugais, corse et surtout en espagnol sont présents en quantité variable. Il en va de même pour les siècles traités. Le choix des langues et des périodes couvertes par les jeux de données accessibles dépend uniquement de l'intérêt des chercheur · se · s : le déséquilibre du corpus est donc inhérent à la nature des projets susceptibles de publier des données d'entraînement pour l'OCR.

Pour le cas du français, une attention particulière a néanmoins été portée à garantir une représentation de la totalité de la diachronie, du XVI^{e} au XX^{e}s. Les documents qui ne sont pas en écriture latine classique, comme la gothique (Pouspin, 2016) ou les caractères de civilité (Jimenes, 2011), ne sont pas retenus. Le fait que des imprimés français de tous les siècles soient présents dans nos données d'entraînement devrait permettre un traitement correct pour les langues qui ne sont pas représentées pour tous les siècles (anglais, italien, latin…).

4 Recommandations pour la transcription

Aucun guide de transcription n'existe pour les imprimés, si ce n'est quelques premières propositions publiées par Gabay et al. (2023)

2. https://tesseract-ocr.github.io/tessdoc/Data-Files.html.

Projet	Langue	Siècle	Pages	Lignes	Dépôt Github
SETAF	frm	XVI^{e} s.	87	2 402	HTR-Varia-Malingre
FoNDUE	frm	XVI^{e} s.	223	5 936	FONDUE-LA-PRINT-16
FoNDUE	la	XVI^{e} s.	930	17 817	FONDUE-FR-PRINT-16
Gallic(orpor)a	frm	XVI^{e} s.	180	4 918	HTR-imprime-16e-siecle
CREMMA	frm	XVI^{e}-XVII^{e} s.	98	2 603	cremma-16-17-print
Gallic(orpor)a	fr	XVII^{e} s.	327	8 950	HTR-imprime-17e-siecle
FoNDUE	fr	XVII^{e} s.	69	1 899	FONDUE-FR-PRINT-17
Gallic(orpor)a	fr	XVIII^{e} s.	160	4 500	HTR-imprime-18e-siecle
Sous-total			2 074	49 025	
FoNDUE	es, ca	XIX^{e} s.	2 896	179 339	FONDUE-ES-CORDEL-19
FoNDUE	es	XIX^{e} s.	48	1 668	FONDUE-ES-PRINT-19
FoNDUE	it†	XX^{e} s.	28	1 150	FONDUE-IT-PRINT-20
HTR-United	fr†	XX^{e} s.	150	4 115	tapuscorpus
FoNDUE	en	XX^{e} s.	30	1 728	FONDUE-EN-PRINT-20
FoNDUE	fr	XX^{e} s.	55	1 604	FONDUE-FR-PRINT-20
FoNDUE	fr, de	XX^{e} s.	215	5 664	FONDUE-MLT-ART
FoNDUE	fr, it, pt	XX^{e} s.	1 381	43 114	FONDUE-MLT-CAT
	fr, co	XX^{e} s.	47	1 681	HN2021-OCR-Poesie-Corse
Sous-total			4 850	240 063	
Total			6 924	289 088	

TABLEAU 3 – Détails des données utilisées pour l'entraînement. Le signe † indique la présence de tapuscrits. Les données sont séparées entre données modernes (XVI-XVIII^{e}) et contemporaines (XIX-XX^{e}).

pour les documents francophones des XVI^{e}-XVIII^{e} siècles, à l'inverse du Moyen Âge (Pinche, 2022) ou des imprimés gothiques de la Renaissance (Solfrini et al., 2023). En gardant constamment à l'esprit l'idée de conserver autant que possible une compatibilité maximum avec les recommandations des guides pour les autres époques, et notamment l'idée d'une représentations graphématique du texte[3], nous présentons ici quelques grands choix qui ont été faits au cours de notre travail de transcription (cf. tab. 4).

— Nous utilisons ‹¬› (`U+00AC`) pour le tiret de fin de ligne, afin de le différencier du trait d'union ;
— nous ne conservons pas ce qui concerne l'emphase (italique, gras…), qui doit être traité par un autre modèle (Scius Bertrand et al., 2021), à l'exception des petites majuscules qui sont rendues avec des majuscules ;
— nous maintenons le *s* long (‹ſ›, `U+017F`), mais aucune autre particularité de la typographie ancienne (ligatures disparues…) ;
— les différentes formes de tiret long pouvant être dures à reconnaître, surtout en diachronie (cadratin, demi-cadratin…), nous recommandons en cas de doute d'opter pour le demi-cadratin (`U+2013`) ;
— la gestion des espaces typographiques est rationalisée au maximum. En cas d'espace insécable ou fine (recommandée avant certains signes en français, comme le point-virgule ou les deux-points), nous recommandons de ne pas mettre d'espace, sauf quand cela est nécessaire (après le point, la virgule, avant les guillemets fermants…) ;
— la partie suscrite des abréviations (comme M^{lle}) est précédées de ‹ˆ›. Les caractères suscrits ont été envisagés, mais l'unicode ne couvre pas l'entièreté des possibilités.

3. Le concept de « graphématique » a été présenté par Stutzmann et al. (2017) : la transcription graphématique est une transcription qui réduit chaque forme à son sens dans le système alphabétique actuel et préserve la suite des lettres.

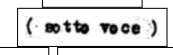
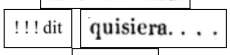
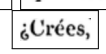
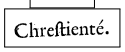
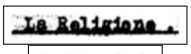
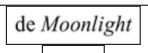
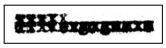
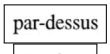
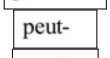
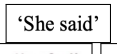
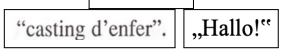
Catégorie	Cas	Traitement	exemple
Espace	Avant virgule, deux-points, point…	Pas d'espace	
	Après et avant guillemets, parenthèses…	Pas d'espace	
	Itération de signes	Pas d'espace	
	Signe inversé, deux signes forts de suite	Pas d'espace	
Ligatures	Conservées en langue cont.	Conservation	
	Disparues en langue cont.	Pas de ligature	
Emphase	Soulignement	Pas de soulignement	
	Italique	Pas d'italique	
	Petites majuscules	Majuscules	
Rature	Texte lisible	Entre double-crochets ⟦⟧ (avec le texte entre)	
	Texte illisible	Entre double-crochets ⟦⟧ (un point par lettre)	
Tiret	Court (trait-d'union)	Trait d'union	
	Court (fin de ligne)	Not-sign (‹¬›)	
	Long	Demi-cadratin (‹–›)	
	Double	Demi-cadratin (‹–›)	
Signes auxiliaires	Guillemet simple	Respect du système	
	Guillemets doubles	Respect du système	
	Apostrophe	Apostrophe droite	
Abréviation	Suscrit	Faire précéder de ‹^›	

TABLEAU 4 – Principales règles de transcription

5 Entraînements

L'entraînement est fait avec Kraken (Kiessling, 2019) (v. 4.3.13), qui est compatible avec l'application eScriptorium (Kiessling et al., 2019 ; Stokes et al., 2021). Ces deux outils forment une suite complète, ouverte, déjà utilisée dans de nombreuses institutions de recherche comme à l'université de Genève (Gabay et al., 2021-). Kraken offrant une souplesse au niveau des architectures de modèles, nous avons entraînés trois modèles basés sur trois architectures différentes :

Tiny `[1,36,0,1 Ct3,3,16 Mp3,3 Lfys48 Lfx96 Lrx96 Lfx128]` ;

Small `[1,48,0,1 Ct3,3,16 Mp3,3 Lfys64 Lfx96 Lrx96 Lfx192]` ;

Large `[1,120,0,1 Cr4,2,32,4,2 Gn32 Cr4,2,64,1,1 Gn32 Mp4,2,4,2 Cr3,3,128,1,1 Gn32 Mp1,2,1,2 S1(1x0)1,3 Lbx256 Do0.5 Lbx256 Do0.5 Lbx256 Do0.5]`.

La dernière architecture, *large*, correspond à la configuration de base de Kraken. Les configurations *tiny* et *small* sont inspirées de celles utilisées par Tesseract (v. 5.0) : celle de *tiny* correspond à celle du modèle *fra* (français), tandis que celle de *small* reprend, entre autres, celle des modèles *ara* (Arabe) et *grc* (Grec).

La principale différence entre les modèles se trouve dans la taille d'image utilisée en entrée (en rouge *supra*), avec *small* utilisant des lignes 25% plus grandes que *tiny*, et dans le nombre de paramètres, qui est de 287 000 pour *tiny*, 423 000 pour *small* et 5 700 000 pour *large* – ce qui a un impact direct sur le poids de ces modèles, qui est respectivement de 1,2 Mo, 1,7 Mo et 22 Mo. Théoriquement, des modèles plus petits utilisant des images plus petites devraient réduire l'exactitude du modèle, mais diminuer aussi le temps d'inférence, ce qui est un enjeu crucial pour les numérisations de masses à venir.

Les entraînements sont effectués avec les mêmes hyperparamètres (précision 16, $1e^{-4}$ de taux d'apprentissage, patience de 10 pour l'arrêt après plateau, normalisation unicode NFKD) en dehors des tailles de *batch* (64 pour

Modèle	Accuracy		Temps d'inférence par ligne (hors-domaine)		
	Médiane	Moyenne (Macro)	Moyen	Médian	Ratio (Moy.)
Tiny	96.78	97.39	0.48s	0.39s	1.00
Small	97.53	98.00	0.85s	0.71s	1.79
Large	99.24	98.49	3.19s	2.69s	6.74

TABLEAU 5 – Résultats des tests hors-domaine pour les trois modèles, avec l'exactitude et le temps d'inférence pour chacun.

les petits modèles, 32 pour le plus gros). Les données du corpus sont réparties comme suit : 90% sont conservées pour l'entraînement, 5% pour l'évaluation et 5% pour le test.

6 Évaluation

6.1 Évaluation en-domaine

Les trois modèles ont été évalués sur les 5% de données réservées pour le test, issue des mêmes documents que ceux utilisés pour l'entraînement (cf. tab. 6). Les résultats montrent un impact de la configuration du modèle, mais la progression de l'exactitude (*accuracy*) par caractère n'est pas linéaire avec le nombre de paramètres. Compte tenu de la variation interne du jeu de données (genre, période, etc.), ces scores suggèrent une forte utilisabilité pour les trois modèles.

Modèle	*Accuracy*
Tiny	98.01%
Small	98.27%
Large	98.56%

TABLEAU 6 – Résultat de l'évaluation en domaine.

6.2 Évaluation hors-domaine

Des données hors-domaine – c'est-à-dire ne comprenant aucune donnée commune à celles vues pendant l'entraînement – ont été compilées en suivant deux principaux axes :

— un axe chronologique, qui cherche à produire, par siècle, 10 images de documents aléatoires, issues de la BnF ou des numérisations de la bibliothèque en ligne Persée (Fargier, 2014) ;

— un axe linguistique, qui évalue la résistance du modèle aux variations linguistiques, avec des langues minoritaires de France, le créole louisianais et des langues de pays frontaliers.

Le résultat est un corpus de 100 images (Gabay et al., 2024d), dont 60 représentent 6 siècles d'impression en français, tandis que 40 représentent le créole louisianais (XIX^{e} s., issues de numérisation de gazette), l'alsacien (XX^{e} s.)[4], le picard (XX^{e} s.), ainsi que l'occitan, le catalan, l'allemand, l'espagnol et l'italien (XIX^{e}-XX^{e} s.). Les documents utilisés contiennent seulement en partie des textes littéraires.

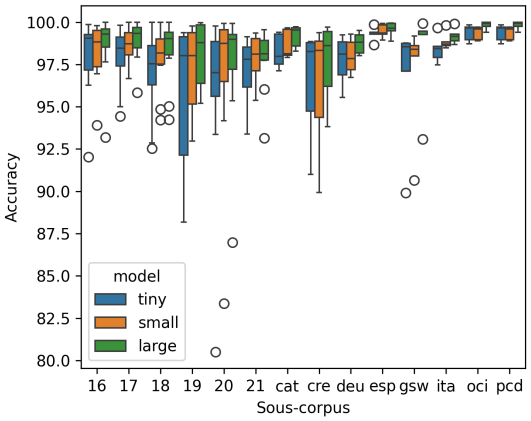


FIGURE 1 – Dispersion des scores par sous-corpus du hors domaine.

Les scores montrent une assez bonne résistance au hors domaine (cf. tab. 5)[5]. La variation est plus forte à l'intérieur des corpus « chronologiques », pour le créole, le XX^{e} s. et l'alsacien (cf. fig. 1). Ces variations sont, d'après nous, le résultat d'une qualité de numérisation plus faible (créole), du faible nombre de données pour l'entraînement (allemand), ou

4. Notons ici que l'Alsacien pose un problème de taille, une partie très importante de la production étant imprimée en *Frakturschrift*.

5. Les moyennes sont évaluées au niveau macro, pour ne pas faire peser trop les documents aux nombreuses lignes.

		Base		Affinage 5 époques			Affinage basé sur le temps			
		Inférence	*Acc.*	*Acc.*	Amélio.	Temps	*Acc.*	Amélio.	Temps	Époques
Français	Tiny	**7,0s**	97,45	98,02	+0,6	12,7s	98,22	+0,8	55,3s	27
	Small	9,7s	97,51	98,15	+0,6	18,0s	**98,31**	+0,8	54,4s	17
	Large	33s	98,46	**98,92**	+0,5	55,4s	-	-	-	-
Catalan	Tiny	**5,7s**	91,44	98,80	+8,0	24,8s	98,90	+8,1	119,9s	27
	Small	7,3s	93,19	98,92	+6,2	36,6s	**98,93**	+6,2	115,1s	17
	Large	21,3s	94,57	**99,19**	+4,9	116,6s	-	-	-	-

TABLEAU 7 – Cette expérience est réalisée sur des *arrows* pré-extraites. Le temps d'inférence et d'affinage est calculé à l'aide de la commande *time unix* et de la métrique « réelle ». Pour le *fine-tuning* basé sur le temps, nous faisons correspondre le temps de *fine-tuning* à 5 époques du modèle *base*. Les améliorations sont calculées en pourcent (+0,6 = 100,6% du score avant affinage).

des textes et fontes très différentes du corpus d'entraînement (XX^{e} s.).

La taille du modèle a un impact sur sa rapidité et sur son exactitude (cf. tab. 5) : chaque ajout de complexité apporte des gains de performance (+0,75 points dexactitude de *tiny* à *small* en médiane, +1,71 points de *small* à *large*) mais aussi une importante multiplication du temps de calcul (*small* est 1,77 fois plus rapide que *tiny* et 6,64 fois que *large*). La mise à disponibilité de ces modèles appelle donc à différents usages, où le temps d'inférence est plus important que l'exactitude obtenue.

6.3 Capacités d'affinage

Les trois modèles sont évalués sur leur capacité d'affinage (*fine-tuning*) sur deux corpus. L'un est un roman français de Jean-Pierre Bastid publié en 1997, le second est un article médical en catalan publiée en 1935. L'objectif est de simuler le gain, en exactitude (*accuracy*) de caractère, avec une page d'entraînement sur neuf pages de test. Pour le premier texte, cela représente 20 lignes pour l'affinage en entraînement, et 5 lignes en évaluation ; pour le second 38 et 5.

Les modèles sont comparés sur deux bases : d'une part en termes de nombre d'époques (nombre de fois où les données sont vues pour l'affinage) et d'autre part en termes de temps, où le temps maximum d'entraînement correspond au temps nécessaire pour affiner le plus grand modèle. Les expériences sont faites sous format déterministique, sur un PC portable sous Ubuntu 22.04 muni d'un CPU Ryzen PRO 7840U. Les hyperparamètres sont les mêmes pour les trois expériences : $1e^{-4}$ de taux d'apprentissage, normalisation NFKD, nombre fixe d'époques, taille de batch 1. Le dernier modèle entraîné est toujours celui sélectionné pour tester.

Pour les deux documents, les modèles semblent atteindre un seuil équivalent, aux alentours de 98 à 99%. Le modèle *base* propose constamment les meilleurs scores d'exactitude, y compris en prenant le même temps d'affinage (cf. tab. 7). Pour le document français, l'affinage a finalement peu d'impact : il atteint déjà les 97,45% de précision avec son plus faible modèle, les modèles obtenant au mieux un gain de 0,64 points de précision. Pour le document catalan cependant, l'affinage a beaucoup plus d'impact, avec des gains atteignant jusqu'à 7,36 points de pourcent. Les modèles sont encore plus serrés en termes de score sur ce jeu de données, avec 0,39 points de pourcent séparant le plus mauvais modèle du meilleur. L'affinage au temps n'apporte au final que très peu d'avantage, pour les deux documents. À temps équivalent, et pour des résultats visant le minage de données, le modèle *tiny* apparaît comme une très bonne alternative sur le document catalan, avec une inférence 4,7 fois plus rapide que le modèle *large*.

7 Prochains travaux

Si les résultats sont satisfaisants, notre modèle est nettement orienté vers quelques-unes des langues de l'Europe occidentale. Si l'on peut penser que nous obtiendrons de bons résultats avec d'autres langues utilisant l'alphabet latin, les caractères spécifiques aux langues scandinaves (‹å›, ‹ø›…), slaves (‹ǎ›, ‹ȁ›…), gaéliques (‹ṡ›, ‹ȝ›)… ne sont pas présents dans notre jeu d'entraînement. C'est là une des principales améliorations à apporter à notre modèle, même si nous avons vu que l'affinage de-

vrait pallier de nombreux problèmes, au moins dans un avenir proche.

Modèles

Les trois modèles, *tiny* (Gabay et al., 2024c), *small* (Gabay et al., 2024b) and *large* (Gabay et al., 2024a) sont disponibles sur Zenodo.

Bibliographie

Thomas M. Breuel. 2014. Ocropy : Python-based tools for document analysis and OCR.

Alix Chagué et Thibault Clérice. 2023. "I'm here to fight for ground truth" : HTR-United, a solution towards a common for HTR training data. In *Digital Humanities 2023 : Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

Alix Chagué, Thibault Clérice, et Laurent Romary. 2021. 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.

Thibault Clérice, Alix Chagué, et Hugo Scheithauer. 2023. Workshop HTR-United : metadata, quality control and sharing process for HTR training data. In *DH 2023 - Digital Humanities Conference : Collaboration as Opportunity*, Graz, Austria. Alliance of Digital Humanities Organizations and University of Graz.

David Etter, Cameron Carpenter, et Nolan King. 2023. A hybrid model for multilingual OCR. In *Document Analysis and Recognition - ICDAR 2023. ICDAR 2023. Lecture Notes in Computer Science.*

Nathalie Fargier. 2014. Persée, une bibliothèque numérique par et pour les chercheurs. In *La francesistica italiana à l'ère du numérique*, Gênes, Italy. Univeristé de Gênes, SUSLLF (Società Universitaria per gli Studi di Lingua e Letteratura Francese), Institut français d'Italie.

Simon Gabay, Robin Champenois, Pierre Kuenzli, Jean-Luc Falcone, et Christophe Charpilloz. 2021-. Formes numerisées et détection unifiée des Écritures (FoNDUE).

Simon Gabay, Thibault Clérice, et Christian Reul. 2023. OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more). *Journal of Data Mining and Digital Humanities*, 2023.

Simon Gabay, Thibault Clérice, Pauline Jacsont, Elina Leblanc, Marie Jeannot-Tirole, Sonia Solfrini, Sophie Dolto, Floriane Goy, Carmen Carrasco Luján, Maddalena Zaglio, Myriam Perregaux, Juliette Janès, Benoît Sagot, Rachel Bawden, Rasul Dent, Oriane Nédey, et Alix Chagué. 2024a. Catmus-print [large].

Simon Gabay, Thibault Clérice, Pauline Jacsont, Elina Leblanc, Marie Jeannot-Tirole, Sonia Solfrini, Sophie Dolto, Floriane Goy, Carmen Carrasco Luján, Maddalena Zaglio, Myriam Perregaux, Juliette Janès, Benoît Sagot, Rachel Bawden, Rasul Dent, Oriane Nédey, et Alix Chagué. 2024b. Catmus-print [small].

Simon Gabay, Thibault Clérice, Pauline Jacsont, Elina Leblanc, Marie Jeannot-Tirole, Sonia Solfrini, Sophie Dolto, Floriane Goy, Carmen Carrasco Luján, Maddalena Zaglio, Myriam Perregaux, Juliette Janès, Benoît Sagot, Rachel Bawden, Rasul Dent, Oriane Nédey, et Alix Chagué. 2024c. Catmus-print [tiny].

Simon Gabay, Thibault Clérice, Benoît Sagot, Rachel Bawden, Juliette Janès, et Rasul Dent. 2024d. FONDUE-MLT-PRINT-TEST.

Rémi Jimenes. 2011. *Les caractères de civilité. Typographie et calligraphie sous l'Ancien Régime.* Atelier Perrousseaux, Gap.

Anthony Kay. 2007. Tesseract : An open-source optical character recognition engine. *Linux J.*, 2007(159) :2.

Benjamin Kiessling. 2019. Kraken - an Universal Text Recognizer for the Humanities. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, The Netherlands. ADHO.

Benjamin Kiessling, Robin Tissot, Peter Stokes, et Daniel Stökl Ben Ezra. 2019. eScriptorium : an open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19. IEEE.

Johannes Michael, Max Weidemann, et Roger Labahn. 2018. HTR engine based on neural networks P3. optimizing speed and performance – HTR+. Rapport technique, Transkribus.

Ariane Pinche. 2022. Guide de transcription pour les manuscrits du Xe au XVe siècle. Document de travail.

Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'Connor, Wouter Haverals, Mike Kestemont, et Caroline Vandyck. 2023a. CATMuS-Medieval : Consistent Approaches to Transcribing ManuScripts. Document de travail.

Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Maria Vlachou-Efstathiou, Marc Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Anne Boutreux, Alec Manton, et Simon Gabay. 2023b. CATMuS Medieval. v. 1.0.0.

Marion Pouspin. 2016. *Publier la nouvelle : Les pièces gothiques, histoire d'un nouveau média (XVe-XVIe siècles).* Éditions de la Sorbonne.

Joan Puigcerver et Carlos Mocholí. 2018. Pylaia. https://github.com/jpuigcerver/PyLaia.

Anna Scius Bertrand, Simon Gabay, Ljudmila Petkovic, Juliette Janes, Caroline Corbières, et Thibault Clérice. 2021. The BIR database – Identifying typographic emphasis in list-like historical documents. In *HIP@ICDAR21 - The 6th International Workshop on Historical Document Imaging and Processing*, Lausanne, Switzerland.

Ray Smith, Daria Antonova, et Dar-Shyang Lee. 2009. Adapting the tesseract open source OCR engine for multilingual OCR. In *Proceedings of the International Workshop on Multilingual OCR*, MOCR '09, New York, NY, USA. Association for Computing Machinery.

Sonia Solfrini, Simon Gabay, Geneviève Gross, Pierre-Olivier Beaulnes, Aurélia Marques Oliveira, et Daniela Solfaroli Camillocci. 2023. Guide de transcription pour les imprimés français du XVIe siècle en caractères gothiques : Version A. Document de travail.

Peter A. Stokes, Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, et El Hassane Gargem. 2021. The eScriptorium VRE for Manuscript Cultures – Classics@ Journal. *Classics@ Journal*, 18(1).

Dominique Stutzmann, 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*, 73 :67–96.

Equipe Trankribus. 2022. Transkribus print M1. Modèle No 39995.

Equipe Trankribus. 2023. Introducing transkribus super models – get access to 'the text titan i'.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, et Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.