Erreurs OCR et biais d'indexation : impact sur les usages

Guillaume Chiron[*,***], Jean-Philippe Moreux[*,****]
Antoine Doucet[**,♯], Mickaël Coustaty[**,♯♯], Muriel Visani[**,♯♯♯]

[*]Bibliothèque nationale de France, Service numération, Paris
[**]L3i, Université de la Rochelle, Avenue Michel Crépeau, 17042 La Rochelle
[***]guillaume.chiron@bnf.fr, [****]jean-philippe.moreux@bnf.fr
[♯]antoine.doucet@univ-lr.fr, [♯♯]mickael.coustaty@univ-lr.fr, [♯♯♯]muriel.visani@univ-lr.fr

Résumé. Les méthodes d'analyse classiquement appliquées dans le contexte du *Big Data*, provoquent souvent un phénomène de « boîte noire » où la qualité de numérisation des documents peut être un paramètre négligé. En dépit des bonnes pratiques en vigueur inhérentes au métier de *data-journalist*, se pose la problématique des biais statistiques induits par ce manque de transparence sur la fiabilité des sources. S'inscrivant dans le cadre du projet AméliOCR, cet article vise à estimer ces potentiels biais sur l'indexation et la recherche. Cette étude s'appuie sur un corpus de documents OCéRisés associés à leur vérité terrain, ainsi que sur des historiques de recherche sur Gallica.

1 Introduction

L'amélioration des technologies de numérisation – toutes sources confondues (p. ex. documents papiers, archives audio/vidéo) – associée à des méthodes de traitement toujours plus performantes (p. ex. OCR/reconnaissance de textes/images, transcription automatique) génère une quantité croissante d'informations. Pour les *data-journalists*, cela constitue un terrain de jeu au potentiel sans précédent, mais dans lequel il est recommandé de s'aventurer avec précaution. Les méthodes d'analyses (p. ex. filtrage, croisement de données) classiquement appliquées dans le contexte du Big Data, négligent souvent l'aspect qualitatif des documents numériques exploités pour ne favoriser qu'une approche quantitative imparfaite.

Dans cet article, nous nous intéressons au cas particulier des documents textuels OCéRisés. Ce travail s'inscrit dans la cadre du projet AméliOCR [1] lancé en 2016, dont l'objectif est d'améliorer la qualité du texte dans les documents historiques numérisés au sein de la bibliothèque numérique Gallica [2]. Nous portons une attention particulière à détecter et corriger les erreurs d'OCR qui affectent les termes les plus recherchés dans Gallica. L'enjeu est de taille, sachant l'impact que des mauvais résultats de recherche peuvent avoir sur des analyses automatisées (Traub et al., 2015). À titre d'exemple, on citera l'utilisation de Gallica comme source d'attestation de lexique [3] et ce cas emblématique de biais : une recherche sur « gadget »

1. Fruit d'une collaboration entre la Bibliothèque nationale de France et le laboratoire L3i.
2. Bibliothèque numérique de la BnF en libre accès : http ://gallica.bnf.fr
3. « Alain Rey et Gallica : une grande histoire de mots », http ://gallica.bnf.fr/blog/20102016/alain-rey-et-gallica-une-grande-histoire-de-mots

(à l'étymologie discutée) dans la presse du XIX^{e} renvoie de nombreuses occurrences qui sont en fait des transcriptions OCR erronées de « budget » !

On retrouve dans la littérature un certain nombre de travaux visant à améliorer *a posterori* des résultats d'OCR. Certains s'appuient essentiellement sur des modèles de langages tels que (Bassil et Alwani, 2012) via *Google Suggest*, d'autres utilisent des modèles d'erreurs (Brill et Moore, 2000) voire éventuellement avec un retour à l'image (Lee et Smith, 2012). Ces approches se heurtent généralement à des limites statistiques (Smith, 2011), et c'est d'ailleurs pour cela que bon nombre d'initiatives de correction assistées par l'homme ont été proposées (Taghva et Stofsky, 2001). Néanmoins le problème, notamment pour les documents anciens qui sont particulièrement difficiles à OCéRiser, reste entier.

La première phase du projet AméliOCR nous amène à proposer deux contributions : 1) la constitution d'un corpus pour l'analyse des erreurs d'OCR, qui sera prochainement rendu public ; 2) une méthode d'alignement entre les textes OCéRisés et leur vérité terrain [4].

2 Constitution d'un corpus OCR / VT

1) Compilation des documents – Nous avons rassemblé un corpus de documents anciens en français qui est à notre connaissance le plus conséquent dans son genre. Comme le montre le tableau 1, il regroupe des documents OCéRisés de natures différentes (p. ex. journaux, monographies) dont certains proviennent de projets de recherche européens antérieurs (IMPACT, Europeana Newspapers) et d'autres de projets ou programmes de numérisation de la BnF. La richesse de ce corpus est qu'il dispose pour chaque document OCéRisé (OCR) d'une vérité terrain (VT) sur le texte. La majorité des documents (OCR + VT) sont disponibles en libre accès, mais reposent sur une variété de formats et de versions utilisés dans le domaine (p. ex. ALTO, PAGE, EPUB, texte brut) ainsi qu'un certain nombre de spécificités propres à chacun (p. ex. métadonnées, encodages, ordre de lecture renseigné ou non). Afin de rendre ce corpus accessible d'une part, et permettre l'agrégation de ces données au sein d'un corpus homogène, un important travail d'ingénierie a été réalisé.

Source	**Nature**	**Dates**	**Symboles alignés**
Europeana News. (52 pages)	périodiques	1814 - 1944	1 066 994 (92%)
IMPACT (1004 pages)	monographies	1821 - 1864	1 190 331 (98%)
VT BnF (6656 pages)	mixte	1820 - 1943	8 861 428 (98%)
Marché de masse (151 pages)	mixte	1654 - 2000	270 471 (95%)
Presse autre (32 pages)	périodiques	1897 - 1934	650 720 (90%)
Monog autre (70 livres)	monographies	1610 - 1926	16 518 313 (99%)
	TOTAL	1610 - 2000	27 597 957 (98,7%)

TAB. 1 – *Constitution du corpus FR pour le projet AméliOCR*

2) Alignement OCR / VT – Pour pouvoir identifier les erreurs-types d'OCR et analyser leur fréquence, il est nécessaire d'aligner au symbole près les deux versions. Les outils d'alignement traditionnellement utilisés par la communauté tels que ISRI (Rice et Nartker, 1996) – voire d'autres extensions [5] – n'ont pas été conçus pour gérer des séquences à l'échelle d'un

4. Texte transcrit à la main à partir des images.
5. `https://github.com/kba/awesome-ocr`

livre. Des approches récentes telles que (Yalniz et Manmatha, 2011; Al Azawi et al., 2013), plus performantes, offrent une solution à ce problème. Celles-ci fonctionnent de manière récursive via l'identification de sous-chaînes similaires de plus en plus petites entre l'OCR et la VT. Le positionnement d'ancres à différentes échelles permet ainsi un appariement allant jusqu'au niveau du caractère. Ce mécanisme d'ancrage strict peut poser problème lorsque l'OCR est trop dégradé, faute de pouvoir identifier suffisamment de sous-chaînes similaires. C'est pourquoi nous avons développé une approche d'alignement à base d'ancrages flous où des sous-chaînes légèrement différentes peuvent servir de repère pour l'alignement. Les séquences non-similaires (par opposition aux séquences similaires) entre les repères sont ensuite alignées à l'aide d'une méthode originalement utilisée pour l'alignement de paires d'ADN (Smith et Waterman, 1981). En dépit d'un temps de calcul plus important, cette nouvelle méthode conduit à des résultats d'alignement satisfaisant sur les parties bruitées de l'OCR. Au total, le corpus a été aligné à plus de 98%, ce qui représente près de 27,5 millions de symboles appariés. Cela permet entre autres de retrouver les erreurs réalisées par les logiciels de reconnaissance de caractères du marché d'une part, et les termes originaux impliquées d'autre part (cf. fig 1). Par exemple, nous avons pu constater que les mesures de la qualité renseignées par certains moteurs OCR sont décorrélées des erreurs réellement avérées. Ces métriques (p. ex. « Word Confidence » renvoyé par l'OCR FineReader) sont donc peu fiables dans un contexte d'exploitation à grande échelle via une approche automatique.

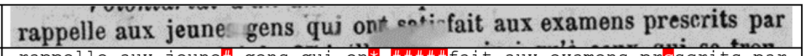
IMG	
OCR	rappelle aux jeune# gens qui on* #####fait aux examens prescrits par
VT	rappelle aux jeunes gens qui ont satisfait aux examens prescrits par
WC	0.70 0.99 0.99 0.99 0.99 0.99 0.98 0.99 0.99 1.00 0.99

FIG. 1 – *Alignement entre l'OCR et la VT, où le "#" symbolise les symboles manquants (ou illisibles). Word Confidence (WC) est une métrique exprimant le taux qualité estimé au mot.*

3 Croisement entre les erreurs et les recherches dans Gallica

L'analyse porte sur les deux points suivants : 1) la nature et la fréquence des erreurs d'OCR ; 2) le croisement de ces erreurs avec les termes les plus fréquemment recherchés sur Gallica :

1) Fréquence des erreurs – La figure 2 donne un aperçu des erreurs d'OCR les plus fréquentes. Sur les 5 millions de mots constituant le corpus, on compte plus de 100k mots concernés par des erreurs d'OCR (hors erreurs de ponctuation), ce qui représente 2 % des mots corpus. Parmi ces erreurs, on estime que 1) 15 % des mots mals OCéRisés sont des noms propres (repérés ici par une majuscule non précédée d'un point), et que 2) près de la moitié des erreurs concerne des termes non présents dans un dictionnaire classique (i.e. dictionnaire OpenOfficeFr).

2) Croisement avec les recherches Gallica – Nous exploitons pour cela la liste des 26 000 termes les plus fréquemment requêtés sur Gallica sur une période de 4 mois (de décembre 2015 à mars 2016). On note qu'un nombre important de noms propres – soit environ 79% sur les 500 premières requêtes – sont la cible de recherches. Globalement, nous observons que 21% des mots (ou occurrences) de la vérité terrain se retrouvent dans les logs de recherche. Ce chiffre important s'explique par les termes communément utilisés (p. ex. « sont », « pour »). Parmi les

1 symbole		
OCR	VT	Freq.
f	s	5069
l	t	1915
c	e	1633
o	e	1344
i	l	1252
	s	1074
t	l	1066
u	n	1007

2 symboles		
OCR	VT	Freq.
fi	st	286
li	h	264
ff	ss	246
11	ll	226
fl	st	223
in	m	186
ni	m	178
ii	n	169

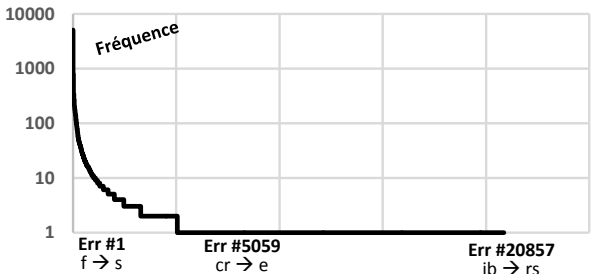


FIG. 2 – *Fréquences des 20857 erreurs d'OCR constatées sur le corpus, avec un tableau détaillé montrant les 8 premières touchant 1 et 2 symboles.*

termes présents à la fois dans les logs de Gallica et dans la VT (au nombre de 3492), 1% sont victimes d'erreurs d'OCR. Voici quelques exemples de termes recherchés sur Gallica les plus fréquemment mal OCéRisés, ainsi que le nombre d'occurrences de ces erreurs :
- *TERMES FRÉQUENTS = [sont*420, sous*199, pour*194, dans*176, cette*150, ...]*
- *ENTITÉS NOMMÉES = [France*35, Egypte*27, Rome*17, Russie*12, Edouard*11, ...]*

4 Conclusion

Cet article présente les prémisses du travail effectué dans le cadre du projet AméliOCR qui vise à proposer une approche de correction automatique des erreurs d'OCR sur les documents numérisés. Les enjeux sont la réduction des biais d'indexation et de recherche, lesquels peuvent altérer les résultats d'analyses menées en aval par les utilisateurs de portails documentaires ou de corpus numériques.

Ainsi, la première phase du projet a donné lieu à deux contributions : 1) la constitution d'un corpus pour l'analyse des erreurs d'OCR ; 2) une approche d'alignement entre les textes OCéRisés et leur vérité terrain. Bien que les chiffres – issus de calculs automatisés – soient à considérer comme des estimations, ce travail vise à alerter les professionnels de l'information des possibles biais rencontrés lors d'analyses massives et automatisées de corpus numérisés, en dépit des bonnes pratiques en vigueur dans leur métier.

Références

Al Azawi, M., M. Liwicki, et T. M. Breuel (2013). Wfst-based ground truth alignment for difficult historical documents with text modification and layout variations. In *IS&T/SPIE Electronic Imaging*, pp. 865818–865818. International Society for Optics and Photonics.

Bassil, Y. et M. Alwani (2012). Ocr post-processing error correction algorithm using google's online spelling suggestion. *Journal of Emerging Trends in Comp. and Info. Sciences 3*.

Brill, E. et R. C. Moore (2000). An improved error model for noisy channel spelling correction. In *Proceedings of Annual Meeting on Association for Comp. Linguistics*, pp. 286–293.

Lee, D.-S. et R. Smith (2012). Improving book ocr by adaptive language and image models. In *Document Analysis Systems, 10th IAPR International Workshop on*, pp. 115–119. IEEE.

Rice, S. V. et T. A. Nartker (1996). The isri analytic tools for ocr evaluation. *UNLV/Information Science Research Institute, TR-96-02*.

Smith, R. (2011). Limits on the application of frequency-based language models to ocr. In *2011 International Conference on Document Analysis and Recognition*, pp. 538–542. IEEE.

Smith, T. F. et M. S. Waterman (1981). Identification of common molecular subsequences. *Journal of molecular biology 147*(1), 195–197.

Taghva, K. et E. Stofsky (2001). Ocrspell : an interactive spelling correction system for ocr errors in text. *International Journal on Document Analysis and Recognition 3*(3), 125–137.

Traub, M. C. et al. (2015). Impact analysis of ocr quality on research tasks in digital archives. In *International Conf. on Theory and Practice of Digital Libraries*, pp. 252–263. Springer.

Yalniz, I. Z. et R. Manmatha (2011). A fast alignment scheme for automatic ocr evaluation of books. In *2011 International Conference on Document Analysis and Recognition*, pp. 754–758. IEEE.

Summary

The processing methods conventionally applied in Big Data often cause a "black box" effect into which the quality of digitized documents is often neglected. Despite the good practices of data-journalists, arises the problem of statistical biases induced by this lack of transparency on the reliability of the sources. As part of the AméliOCR project, this study aims to estimate these potential biases on indexing and searching. This work is based on a corpora of OCR-ized documents associated with their ground truth, as well as historical search logs gathered from Gallica.