

ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Kutay Sefil

licencié en histoire

L'implémentation de l'OCR dans une bibliothèque patrimoniale

**L'exemple de la Bibliothèque
interuniversitaire de la Sorbonne**

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2024

Résumé

Ce mémoire a été réalisé dans le cadre du master Technologies Numériques Appliquées à l’Histoire à l’École nationale des chartes. Il a été rédigé à la suite d’un stage de quatre mois à la Bibliothèque interuniversitaire de la Sorbonne. Cette dernière souhaitait implémenter de l’OCR aux documents qui sont en ligne sur sa bibliothèque numérique, NuBIS. Ce mémoire expose les réflexions menées à ce sujet en s’intéressant tout d’abord aux particularités de cette bibliothèque et de leur collection avant d’étudier les solutions d’OCR qui soient adaptées à la Sorbonne.

This thesis was produced as part of the “Technologies Numériques Appliquées à l’Histoire” master’s program at the École nationale des chartes. It was written following a four-month internship at the Bibliothèque interuniversitaire de la Sorbonne. The library wished to implement OCR on the documents that are online on its digital library, NuBIS. This thesis outlines the thinking behind the project, focusing first on the particularities of the library and its collection, before looking at OCR solutions that are suitable for the Sorbonne.

Mots-clés : OCR ; Sorbonne ; bibliothèque ; numérisation ; histoire du livre ; logiciel ; Omeka S ; IIIF ; intelligence artificielle.

Informations bibliographiques : Kutay Sefil, *L’implémentation de l’OCR dans une bibliothèque patrimoniale. L’exemple de la Bibliothèque interuniversitaire de la Sorbonne*, mémoire de master « Technologies numériques appliquées à l’histoire », dir. Emmanuelle Bermès, École nationale des chartes, 2024.

Remerciements

MES es remerciements vont tout d'abord à Emmanuelle Bermès pour sa patience et son accompagnement, non seulement pendant la période du mémoire, mais aussi durant la globalité du master TNAH en tant que responsable pédagogique.

J'adresse également mes remerciements à Laurie Aoustet, Cécile Obligi et Juliette Jestaz qui m'ont accompagné pendant les quatres de mois de stage à la Sorbonne et m'ont permis d'acquérir une expérience enrichissante à leurs côtés. Je remercie en particulier Sébastien Clément pour son aide précieuse durant l'ensemble du stage.

Je tiens aussi à remercier mes amis et ma famille pour leur soutien lors de la rédaction de ce mémoire. Enfin, mes remerciements vont à mes camarades de promotion pour ces deux années passées ensemble qui furent pleines de découvertes.

Bibliographie

Histoire de la BIS et du livre

A l'origine : la bibliothèque de la rue Saint-Jacques (1770-1823), URL : <https://www.bis-sorbonne.fr/biu/spip.php?article29> (visité le 11/08/2024).

Aspects techniques, URL : <https://nubis.bis-sorbonne.fr/page/aspects-techniques> (visité le 11/08/2024).

BARBIER (Frédéric), « L'invention de l'imprimerie et l'économie des langues en Europe au XVe siècle », *Histoire et Civilisation Du Livre*, 4 (2008), p. 21-46, URL : https://revues.droz.org/HCL/article/view/HCL_4_21-46.

— *Histoire du livre en Occident*, 2020, DOI : 10.3917/arco.barbi.2020.01.

— *Histoire des bibliothèques. D'Alexandrie aux bibliothèques virtuelles*, 2021, URL : <https://shs.cairn.info/histoire-des-bibliotheques--9782200630126> (visité le 30/09/2024).

BOBIS (Laurence) et NOGUÈS (Boris), *La Bibliothèque de la Sorbonne, 250 ans d'histoire au cœur de l'Université*, Éditions de la Sorbonne, Paris, France, 2022.

Collections, URL : <https://www.bis-sorbonne.fr/biu/spip.php?rubrique10> (visité le 11/08/2024).

Corpus numérisés, URL : <https://nubis.bis-sorbonne.fr/page/le-corpus> (visité le 11/08/2024).

DERROT (Sophie), *La marquise Arconati Visconti*, nov. 2023, URL : http://blog.bibliotheque.inha.fr/fr/posts/marquise_arconati_visconti.html (visité le 02/09/2024).

Équipe, URL : <https://nubis.bis-sorbonne.fr/page/équipe> (visité le 11/08/2024).

Claude Jolly (éd.), *La Bibliothèque de la Sorbonne*, Paris, France, 1989.

La bibliothèque de l'ancienne Sorbonne (1823-1897), URL : <https://www.bis-sorbonne.fr/biu/spip.php?article41> (visité le 11/08/2024).

La bibliothèque de la nouvelle Sorbonne (1897-2013), URL : <https://www.bis-sorbonne.fr/biu/spip.php?article40> (visité le 11/08/2024).

Rapport d'activité 2022 de la BIS, URL : https://www.bis-sorbonne.fr/biu/IMG/pdf/bis_ra_2022_web.pdf.

VEYRIN-FORRER (Jeanne), « Hommage aux premiers imprimeurs de France. 1470-1970 », *Bulletin des bibliothèques de France (BBF)*–2 (1971), p. 65-80, URL : <https://bbf.enssib.fr/consulter/bbf-1971-02-0065-001>.

OCR et HTR

ANDERSON (Niall), MUHLBERGER (Gunter) et ANTONACOPOULOS (Apostolos), *Optical Character Recognition - IMPACT Best Practice Guide*, sept. 2023, URL : https://www.digitisation.eu/wp-content/uploads/2023/09/OpticalCharacterRecognition-IBPG_01.pdf.

CAMPS (Jean-Baptiste), CLÉRICE (Thibault) et PINCHE (Ariane), « Noisy medieval data, from digitized manuscript to stylometric analysis : Evaluating Paul Meyer's hagiographic hypothesis », *Digital Scholarship in the Humanities*, 36–Supplement_2 (oct. 2021), p. ii49-ii71, DOI : [10.1093/l1c/fqab033](https://doi.org/10.1093/l1c/fqab033).

Thierry Claerr et Isabelle Westeel (éd.), *Numériser et mettre en ligne*, Villeurbanne, 2010 (La Boîte à outils), URL : <https://books.openedition.org/pressesensib/414> (visité le 05/09/2024).

CORTADA (James W.), *Before the Computer : IBM, NCR, Burroughs, and Remington Rand and the Industry They Created, 1865-1956*, 2015.

GABAY (Simon), CLÉRICE (Thibault) et REUL (Christian), « OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more) », *Journal of Data Mining and Digital Humanities*, 2023 (juin 2023), DOI : [10.46298/jdmdh.6492](https://doi.org/10.46298/jdmdh.6492).

GABAY (Simon), CLÉRICE (Thibault), JACSONT (Pauline), LEBLANC (Elina), JEANNOT-TIROLE (Marie), SOLFRINI (Sonia), DOLTO (Sophie), GOY (Floriane), LUJÁN (Carmen Carrasco), ZAGLIO (Maddalena), *et al.*, « Reconnaissance des écritures dans les imprimés », dans *Humanistica 2024*, Meknès, Morocco, 2024 (OCR), URL : <https://hal.science/hal-04557457> (visité le 10/08/2024).

HEIL (Jacob) et SAMUELSON (Todd), « Book History in the Early Modern OCR Project, or, Bringing Balance to the Force », *Journal for Early Modern Cultural Studies*, 13–4 (2013), p. 90-103, JSTOR : [jearlmodcultstud.13.4.90](https://www.jstor.org/stable/jearlmodcultstud.13.4.90), URL : <https://www.jstor.org/stable/jearlmodcultstud.13.4.90> (visité le 05/09/2024).

HOLLEY (Rose), « How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs », *D-Lib Magazine*, 15–3/4 (mars 2009), DOI : [10.1045/march2009-holley](https://doi.org/10.1045/march2009-holley).

PATEL (Chirag), PATEL (Atul) et PATEL (Dharmendra), « Optical Character Recognition by Open source OCR Tool Tesseract : A Case Study », *International Journal of Computer Applications*, 55 (oct. 2012), p. 50-56, DOI : [10.5120/8794-2784](https://doi.org/10.5120/8794-2784).

Référentiel OCR version 2, URL : https://www.bnf.fr/sites/default/files/2018-11/ref_num_ocr_v2.pdf.

SOLYM (Clément), *Europe : IMPACT ou améliorer l'accès aux textes numérisés*, URL : <https://actualitte.com/article/86285/bibliotheque/europe-impact-ou-ameliorer-l-acces-aux-textes-numerises> (visité le 30/09/2024).

TRIPATHI (Pankaj), *A Journey Through History : The Evolution of OCR Technology*, 2024, URL : <https://www.docsumo.com/blog/optical-character-recognition-history> (visité le 12/07/2024).

ZHANG (James), HAVERALS (Wouter), NAYDAN (Mary) et KERNIGHAN (Brian W.), « Post-OCR Correction with OpenAI's GPT Models on Challenging English Prosody Texts », dans *Proceedings of the ACM Symposium on Document Engineering 2024*, New York, NY, USA, 2024 (DocEng '24), p. 1-4, DOI : 10.1145/3685650.3685669.

Introduction

483 222 résultats, dont 460 294 livres. Ce chiffre correspond au nombre de documents avec OCR disponibles sur Gallica lorsqu'on effectue une recherche avancée en sélectionnant la présence du « mode texte » comme unique critère. Devant ce chiffre gargantuesque, certaines bibliothèques historiques en France avec une vaste collection sont encore au stade zéro avec aucun document océrisé : c'est le cas de la Bibliothèque interuniversitaire de la Sorbonne (BIS) et de sa bibliothèque numérique NuBIS. C'est dans ce contexte, et dans l'objectif de rechercher les meilleures solutions d'OCR pour leurs documents imprimés en fonction des moyens disponibles, que j'ai effectué un stage de quatre mois d'avril à juillet 2024 au sein de la BIS. Le présent mémoire s'inscrit ainsi dans le cadre de ce stage et plus généralement de mes deux ans de master TNAH (Technologies Numériques Appliquées à l'Histoire) à l'École nationale des chartes.

La reconnaissance optique de caractères (plus connue sous l'abréviation d'OCR, de l'anglais *Optical Character Recognition*) désigne la technologie qui permet d'extraire de manière automatique le texte présent dans l'image numérisée d'un document imprimé et de le transformer en un format de texte lisible par une machine. Lorsqu'il est question d'un document manuscrit, on parle alors d'HTR (*Handwritten Text Recognition*). Les origines de la technologie de l'OCR remontent au moins aux années 1920¹ et elle a connu de nombreuses utilisations, telles que la conversion du texte en paroles pour permettre aux personnes malvoyantes d'accéder aux documents imprimés ou la lecture automatique des codes-barres et des numéros de chèques. Dans le monde des bibliothèques, elle est utilisée dès le début des années 1990 dans le cadre notamment de la numérisation de la presse historique.²

Le fonctionnement général de cette technologie est le suivant : l'OCR effectue tout d'abord une analyse de la mise en page d'une image numérique et décompose cette image en éléments plus petits afin de distinguer les zones qui contiennent du texte. Cette étape

1. Tripathi, Pankaj. « A Journey Through History : The Evolution of OCR Technology », 2024. <https://www.docsumo.com/blog/optical-character-recognition-history>.

2. Holley, Rose. « How Good Can It Get ?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs ». *D-Lib Magazine* 15, n° 3/4 (mars 2009). <https://doi.org/10.1045/march2009-holley>.

se nomme la segmentation. Dans chaque zone, le logiciel d'OCR repère alors les lignes de texte et, dans ces lignes, les mots et les caractères individuels. Une fois que le moteur du logiciel a isolé un caractère unique, il analyse les propriétés visuelles de ce caractère afin de l'identifier grâce à une base de données interne. Il répète ensuite le processus pour tous les caractères d'un mot, certains logiciels utilisant un dictionnaire interne afin de corriger éventuellement le mot en question. L'OCR étend alors ce processus à travers les phrases, les lignes et les blocs de texte jusqu'à ce que l'intégralité du texte de l'image ait été identifié.³

De nombreux logiciels d'OCR sont disponibles : ils permettent des usages différents et fonctionnent par conséquent selon des normes et technologies qui peuvent varier. Pour notre étude, nous nous intéressons aux outils qui sont utilisés dans les bibliothèques numériques afin de savoir lesquels sont les plus efficaces pour notre collection numérisée à la BIS. En particulier, l'objectif initial était de disposer d'un outil qui soit capable d'avoir la recherche en plein texte dans un document et d'exporter la transcription dans un format accessible comme le PDF.

Le consortium IMPACT s'est justement penchée sur les réflexions autour de la question de l'implémentation de l'OCR. Lancé en décembre 2007 et financé par la Commission européenne, IMPACT (IMProving Access to Text) est un projet de recherche européen visant à faciliter et améliorer l'accès aux textes numérisés des institutions partenaires (dont la Bibliothèque nationale de France). Son objectif principal est de surmonter les obstacles à la numérisation du patrimoine culturel européen en partageant des connaissances, données, outils, et savoir-faire à l'échelle européenne. Le projet vise également à accélérer la standardisation des pratiques et à améliorer la qualité de la numérisation de masse. IMPACT cherche en particulier à améliorer le processus de numérisation des imprimés historiques. Dans cet objectif, le consortium accorde une place importante à tout ce qui concerne l'océrisation de ces imprimés.⁴

Dans leur rapport dédié à ce sujet⁵, le consortium IMPACT recense plusieurs axes et considérations à prendre en compte pour une institution qui souhaite mettre en place un tel projet :

3. « Référentiel OCR version 2 ». Bibliothèque nationale de France, 2015. p. 7. https://www.bnf.fr/sites/default/files/2018-11/ref_num_ocr_v2.pdf.

4. Solym, Clément. « Europe: IMPACT ou améliorer l'accès aux textes numérisés ». Actualité.com. Consulté le 30 septembre 2024. <https://actualitte.com/article/86285/bibliotheque/europe-impact-ou-ameliorer-l-acces-aux-textes-numerises>.

5. Anderson, Niall, Gunter Muhlberger, et Apostolos Antonacopoulos. « Optical Character Recognition - IMPACT Best Practice Guide ». Optical Character Recognition, 2023. p. 6-8. https://www.digitisation.eu/wp-content/uploads/2023/09/OpticalCharacterRecognition-IBPG_01.pdf.

- **Objectifs du projet** : La solution OCR choisie doit refléter les objectifs du projet. Si l'objectif est de permettre uniquement la recherche en plein texte dans un document, une OCR simple sans correction manuelle peut suffire. Pour des recherches plus spécifiques au sein du texte, un langage comme XML qui utilise des balises sera alors plus pertinent. De même, si le texte est directement affiché aux utilisateurs, l'OCR doit être de meilleure qualité avec une mise en page du texte adaptée à l'usager.
- **Caractéristiques du matériel source et de la numérisation** : La qualité du papier, la langue, la police et autres aspects matériels et typographiques du document peuvent affecter la qualité de l'OCR. Il faut aussi tenir compte de la présence d'éléments graphiques ou de caractères spéciaux. La qualité de la numérisation autrement dit celle des images en elles-mêmes peut également impacter la précision de l'OCR.
- **Contrôle de qualité** : Un programme de contrôle de qualité de l'OCR est recommandé en fonction des objectifs du projet afin d'évaluer l'efficacité des différents logiciels. Cela peut s'effectuer à travers un échantillon représentatif de la collection à océriser.
- **L'échelle du projet** : Le volumes de documents à océriser influe nécessairement sur la solution d'OCR à choisir : les contraintes techniques, budgétaires et de temps ne seront pas les mêmes en fonction de l'ampleur du projet.
- **Internalisation ou externalisation de l'OCR** : L'OCR peut être effectué en production interne ou en externe par l'intermédiaire d'un prestataire. Cela dépend de la disponibilité du matériel, du personnel, du budget et des compétences spécifiques au prestataire externe.
- **Durée du projet** : La durée de la mise en place de l'OCR et de son application dépend de plusieurs facteurs (le matériel source, le logiciel utilisé, le personnel impliqué dans le projet,...). En fonction du temps exigé pour le projet, certaines solutions seront donc plus adaptées que d'autres.
- **Coûts** : L'aspect budgétaire est bien sûr à prendre en considération aussi. Le coût de l'océrisation serait sensiblement différent en fonction du prix du logiciel mais aussi du volume de documents à océriser ou encore des ressources utilisées pour un tel processus.

En résumé, chaque décision liée à l'OCR doit tenir compte de ces différentes contraintes et objectifs souhaités afin de garantir le succès du projet. Toutes ces interrogations peuvent être résumées en la question suivante, qui sera le fil directeur de notre mémoire :

Quels défis soulèvent la mise en place de l’OCR pour une bibliothèque comme la BIS et quelles sont les solutions les plus adaptées pour notre cas précis ?

Pour y répondre, nous étudierons dans un premier temps la bibliothèque de la Sorbonne, son histoire, sa collection et les caractéristiques matérielles des imprimés qui y sont numérisés. Nous verrons ensuite dans un second temps quelles solutions techniques privilégier face aux particularités de la bibliothèque numérique et de sa collection à travers plusieurs travaux effectués pendant le stage.

Chapitre 1

La BIS et sa collection

I. Histoire de la bibliothèque

Contrairement à une croyance répandue, la bibliothèque de la Sorbonne n'est pas l'héritière directe de la collection du célèbre collège de Sorbonne, qui abritait également la faculté de théologie. Les livres imprimés de ce collège furent dispersés pendant la Révolution, tandis que ses manuscrits furent transférés à la Bibliothèque nationale de France. En réalité, la bibliothèque de la Sorbonne est issue de la bibliothèque de l'Université de Paris de l'Ancien Régime, laquelle regroupait la faculté des arts ainsi que les trois facultés supérieures de théologie, de droit, et de médecine. Celle-ci a ouvert pour la première fois ses portes au public le 3 décembre 1770. L'histoire et l'évolution de la bibliothèque peuvent être découpés en suivant ses trois localisations successives, qui correspondent approximativement aux trois grandes périodes de sa collection : rue Saint-Jacques, dans l'ancien collège Louis-le-Grand, siège de l'université de 1770 à 1823, puis dans l'ancienne Sorbonne de 1823 à 1897, et enfin dans la Sorbonne actuelle jusqu'à aujourd'hui.¹

En 1770, la bibliothèque s'installe tout d'abord dans les galeries Fouquet et Harlay de l'ancien collège Louis-le-Grand, dont les jésuites ont été expulsés en 1763. À cette époque, ses collections comptent 20 000 volumes provenant de quatre sources principales² :

- La bibliothèque personnelle de Jean-Gabriel Petit de Montempuis, janséniste reconnu et ancien recteur de l'Université de Paris, léguée en 1762 (5 000 volumes).
- La bibliothèque du collège des Jésuites, attribuée à l'Université en 1764, qui conserve 10 000 volumes et vend le reste.

1. Jolly, Claude, éd. *La Bibliothèque de la Sorbonne*. Paris, France : Bibliothèque de la Sorbonne, 1989.

2. « À l'origine : la bibliothèque de la rue Saint-Jacques (1770-1823) ». Consulté le 9 août 2024. <https://www.bis-sorbonne.fr/biu/spip.php?article29>.

- Les ouvrages issus de 28 collèges parisiens supprimés en 1764, réunis dans une nouvelle institution reprenant le nom et les bâtiments du collège Louis-le-Grand (3 000 volumes).
- Les livres achetés par le bibliothécaire entre 1766 et 1770 (1 700 volumes).

Jusqu'à la Révolution, la bibliothèque est ouverte trois jours par semaine, accueillant non seulement les étudiants et les professeurs, mais aussi le grand public. Elle continue d'enrichir ses collections régulièrement, atteignant 31 000 volumes en 1791. Avec la suppression de l'Université en septembre 1793 et la transformation de ses locaux en caserne et en prison, la bibliothèque est temporairement transférée au dépôt littéraire de Louis-la-Culture (actuelle église Saint-Paul-Saint-Louis). Entre 1796 et 1798, elle est progressivement ramenée au collège Louis-le-Grand, renommé collège Égalité, puis Institut des boursiers et Prytanée français. Durant cette période, elle s'enrichit de nombreux ouvrages confisqués aux ordres religieux et aux émigrés, notamment des collections des familles Condé, Caylus, et Montmorency. Rebaptisée bibliothèque des lycées de Paris en 1802, elle devient en 1808 la bibliothèque de l'Université de France.

En 1823, elle quitte les locaux du collège Louis-le-Grand pour s'installer dans ceux de l'ancien collège de Sorbonne, supprimé par la Révolution. Dans l'ancienne Sorbonne, reconstruite au début du XVII^e siècle par l'architecte Jacques Lemercier sous l'impulsion du cardinal de Richelieu, la bibliothèque de l'Université est installée de manière peu adéquate dans une série de salles situées au quatrième étage des ailes nord et ouest du bâtiment, et non dans la galerie qui abritait autrefois la bibliothèque du collège. Elle stagne ainsi pendant une vingtaine d'années. L'arrivée de Philippe Le Bas à la direction de l'établissement en 1844 (poste qu'il occupe jusqu'en 1860) marque ce qui peut être considéré comme une seconde fondation de la bibliothèque. De 1846 à 1861, celle-ci est officiellement nommée, pour la première fois, bibliothèque de la Sorbonne. Divisée en cinq départements, elle est désormais ouverte tous les jours, sauf les dimanches et jours fériés. Un nouveau cadre de classement est instauré pour redistribuer les livres, et une véritable politique documentaire est mise en place, poursuivie par Léon Renier, directeur de 1860 à 1885. Les collections connaissent une croissance rapide : 39 451 volumes en 1846, 77 500 volumes en 1867, et 300 000 volumes en 1885.³

En raison de la vétusté des locaux de l'ancienne Sorbonne, rénovée au début du XIX^e siècle pour accueillir l'Université, il est décidé de la reconstruire. Les travaux, menés par l'architecte Henri-Paul Nénot, s'étendent de 1885 à 1901. La nouvelle bibliothèque est inaugurée le 29 décembre 1897. Elle se compose de trois grands espaces :

3. « La bibliothèque de l'ancienne Sorbonne (1823-1897) ». Consulté le 9 août 2024. <https://www.bis-sorbonne.fr/biu/spip.php?article41>.

- Une salle de lecture offrant 264 places.
- Deux magasins de cinq étages, l'un dédié aux lettres et l'autre aux sciences.
- Un ensemble de petites pièces le long d'un long couloir, comprenant la salle des professeurs, la salle des périodiques, la Réserve des manuscrits et divers bureaux.

Cependant, ces installations se révèlent rapidement insuffisantes. En 1930, les fauteuils spacieux de la salle de lecture sont remplacés par des sièges plus étroits, permettant ainsi d'accueillir 400 lecteurs simultanément. En 1932, les magasins sont surélevés de trois étages chacun. Entre 1972 et 1977, un troisième magasin est construit en sous-sol, surmonté d'une salle des catalogues et d'une salle de consultation de la Réserve. La bibliothèque récupère également les locaux de l'ancien musée de géologie (la « salle Saint-Jacques »), ceux de l'ancien laboratoire de biologie génétique et de l'ancien institut d'études indiennes, transformés en magasins. Deux nouvelles sections lui sont rattachées en 1978 : la bibliothèque de l'Institut de géographie et la bibliothèque Victor-Cousin, dont les collections complètent les siennes. Pendant cette période, la bibliothèque change de nom à plusieurs reprises. Redevenue bibliothèque de l'Université de France en 1861, elle est renommée, par décret du 28 juin 1910, section lettres et sciences de la bibliothèque de l'Université de Paris (qui regroupe les bibliothèques des facultés et la bibliothèque Sainte-Geneviève), et prend rapidement le nom de bibliothèque de l'Université de Paris à la Sorbonne, abrégé en bibliothèque de la Sorbonne. Elle connaît également divers statuts, le plus récent faisant d'elle une bibliothèque interuniversitaire. Elle est régie par une convention portant sur son organisation et son fonctionnement qui a signée par deux universités : Paris 1 Panthéon-Sorbonne et Paris 3 Sorbonne Nouvelle.⁴

Aujourd'hui, la Bibliothèque interuniversitaire de la Sorbonne conserve et enrichit des collections spécialisées en lettres et sciences humaines, destinées à un public de chercheurs, qu'ils soient débutants ou confirmés. Elle possède aussi un important fonds patrimonial. Les collections imprimées, incluant la bibliothèque de géographie, comptent près de 2 millions de volumes, dont 19 300 titres de périodiques et plus de 50 000 thèses. Chaque année, ces collections s'enrichissent d'environ 17 000 ouvrages et 3 600 revues en abonnement.⁵

La politique documentaire de la BIS vise à renforcer ses collections et les services qui leur sont associés dans les disciplines suivantes : les sciences de l'Antiquité, l'histoire mé-

4. « La bibliothèque de la nouvelle Sorbonne (1897-2013) ». Consulté le 9 août 2024. <https://www.bis-sorbonne.fr/biu/spip.php?article40>.

5. « Collections ». Consulté le 9 août 2024. <https://www.bis-sorbonne.fr/biu/spip.php?rubrique10>.

diévale et moderne, la géographie, la littérature principalement française, la philosophie. Dans ces domaines, la BIS maintient un haut niveau d'acquisition pour les publications étrangères, qu'elles soient imprimées ou numériques. La bibliothèque faisait notamment partie du réseau CADIST (Centres d'acquisition et de diffusion de l'information scientifique et technique) en histoire moderne (depuis 1983), histoire médiévale (depuis 1993), géographique (depuis 1988) et Antiquité (depuis 2008). Cela signifie que la BIS disposait d'un statut particulier de « tête de pont » dans ces disciplines (engagement à ne pas désherber et à conserver des revues internationales par exemple) en échange d'une subvention. Le réseau CADIST a été remplacé en 2017 par celui de CollEX-Persée dont fait partie la BIS⁶. En complément de ces grands axes, des acquisitions en histoire contemporaine, sciences sociales, sciences religieuses, littératures étrangères et sciences du langage viennent enrichir et renforcer les collections.

Catégorie lecteur	2022
L3	4 277
M1	3 475
CAPES-AGRÉG.	824
ERASMUS & ÉCHANGES	687
DOCTORAT	1 955
M2	3 316
ENSEIGN. UNIV.	445
DOCTORANT-ENSEIGNANT	460
PROF. UNIV./MCF	1 344
DOCTEUR - 4 ANS	82
DOCTEUR + 4 ANS	141
CHERCHEUR	649
PERS. BIS & BIB GÉO.	39
PERS. INST. GÉO. & QUADRIL.	7
PRO. DOC.	94
AUTRE STATUT	73
TOTAL	17 868

TABLE 1.1 – Profil des lecteurs fréquentant la BIS

Le tableau ci-dessus renseigne sur le niveau d'étude des lecteurs qui fréquentent la bibliothèque. À la date du 31 décembre 2022, la BIS comptait au total 17 868 selon le rapport d'activité de 2022⁷. Nous constatons que la majorité des lecteurs de la BIS sont donc des étudiants ayant un niveau Bac + 3 au minimum, le reste étant des chercheurs,

6. <https://www.collexpersee.eu/le-reseau/>.

7. « Rapport d'activité 2022 de la BIS », Bibliothèque interuniversitaire de la Sorbonne, p. 42. https://www.bis-sorbonne.fr/biu/IMG/pdf/bis_ra_2022_web.pdf.

doctorants et enseignants. La question ici est donc de savoir si ces lecteurs sont représentatifs des utilisateurs de la bibliothèque numérique (NuBIS) qui sont eux le public concerné par l'implémentation de l'OCR. Bien que l'accès à NuBIS ne nécessite pas de carte lecteur, nous pouvons tout de même supposer que les deux publics sont proches, les utilisateurs de NuBIS penchant probablement plus vers le groupe des chercheurs dû à sa collection patrimoniale et tournée vers la recherche comme nous le verrons par la suite.

II. Organisation de la bibliothèque

La Bibliothèque interuniversitaire s'organise dans un premier temps autour de cinq départements :

- **Département de l'accueil des publics et de la communication des documents** : Il est chargé d'organiser et de garantir l'accès des lecteurs aux espaces et aux ressources de la bibliothèque. Il élabore et met en œuvre le règlement de la bibliothèque à destination du public.
- **Département du traitement documentaire** : Il a pour mission principale, au sein de la BIS, le catalogage de tous les documents des collections du fonds général de la bibliothèque, ainsi que des e-books acquis de façon pérenne.
- **Le département du développement des collections** : Il a pour fonctions de développer les collections du fonds général en mettant en œuvre la politique documentaire définie par la bibliothèque, en particulier dans les domaines prioritaires des collections (Antiquité, Histoire, Philosophie et Littérature).
- **Le département des Manuscrits et des livres anciens** : Il assure la gestion des collections patrimoniales de la bibliothèque. Il est question ici des manuscrits, estampes et autres documents iconographiques, livres imprimés avant 1801, ouvrages à caractère bibliophilique des XIX^e et XX^e siècles, auxquels s'ajoutent les thèses dactylographiées et certaines thèses imprimées.
- **Bibliothèque de géographie** : Elle constitue elle aussi un département de la BIS. Elle traite toutes les collections géographiques et mène des actions spécifiques visant à la mise en valeur de ses fonds. Elle est liée par son fonctionnement à l'Institut de géographie.

En plus de ces cinq départements, il existe à la Sorbonne quatre services transversaux impliqués dans le fonctionnement de la bibliothèque et qui travaillent avec l'ensemble des départements :

- **Service des moyens généraux** : Il est chargé du traitement de toutes les questions administratives. Il traite avec les services centraux de l’Université Paris 1 Panthéon-Sorbonne, avec les services du ministère chargé de l’enseignement supérieur ainsi qu’avec le rectorat.
- **Service de l’informatique et des systèmes d’information** : Il assure le traitement de toutes les questions liées à l’informatique de la bibliothèque.
- **Service de la conservation et de la gestion matérielle des collections** : Il traite des questions liées à la conservation, à la préservation des collections, à l’implantation et au suivi des exemplaires de monographies et de périodiques.
- **Service de la valorisation numérique des collections et du soutien à la recherche (SERVAL)** : Il s’agit du service au sein duquel notre stage s’est déroulé. Il a pour mission de définir, coordonner et opérer des actions de valorisation des collections sur des outils numériques, ainsi que promouvoir et porter des partenariats scientifiques en lien avec des projets de recherche. Le SERVAL gère donc la bibliothèque numérique NUBIS et coordonne le projet d’OCR dont il est question dans cette étude.

III. La bibliothèque numérique

En 1978, la Sorbonne voit l’installation d’un atelier de microfilmage au sein de sa bibliothèque. Cette date marque les débuts timides d’une politique de reproduction des collections physiques qui aboutiront quatre décennies plus tard à la naissance de sa bibliothèque numérique NuBIS. Ce cheminement s’est en effet inscrit dans un paysage documentaire bouleversé à partir du milieu des années 1990 par l’arrivée de l’Internet le développement des nouvelles technologies puis par la généralisation des programmes de numérisation au sein des bibliothèques de l’enseignement supérieur.⁸ C’est dans ce contexte que la BIS installe en février 2007 son propre atelier de numérisation qui est équipé d’un scanner Digibook 10 000 RGB conçu par la société i2S. Celui-ci est adapté au traitement des documents qui peuvent être anciens et fragiles, et permet une numérisation en haute définition⁹. À partir de cette date, la numérisation en couleur devient la règle avec la production de fichiers au format jpeg compressé à 50% et au format tiff qui lui reste non compressé.

8. Bobis, Laurence, et Boris Noguès. *La Bibliothèque de la Sorbonne, 250 ans d’histoire au cœur de l’Université*. Éditions de la Sorbonne. Paris, France, 2022. p. 257.

9. Bibliothèque numérique de la Sorbonne. « Aspects techniques ». Consulté le 9 août 2024. <https://nubis.bis-sorbonne.fr/page/aspects-techniques>.

Cependant, au début des années 2010, alors qu'elle était pionnière dans le paysage documentaire universitaire en se dotant assez tôt d'un tel équipement de pointe qui est adapté à la numérisation des documents patrimoniaux, la BIS n'a pas encore été en mesure de réaliser le volet informatique du projet de numérisation des sources de l'histoire de l'université. Ce projet avait été élaboré en 2004 et devait déboucher sur la diffusion des données numérisées.¹⁰ Après le retour de la BIS dans ses locaux historiques à la Sorbonne, les réflexions autour de la diffusion des documents numérisés reprennent en 2014.

Deux pistes différentes sont alors envisagées, ces pistes ayant déjà été explorées par d'autres bibliothèques interuniversitaires. La première fait appel à la solution « Gallica Marque blanche »¹¹, dont la bibliothèque nationale et interuniversitaire de Strasbourg a été la première à adopter pour la mise en ligne de sa bibliothèque numérique Numistrat. Il s'agit d'un dispositif de coopération numérique qui s'adresse aux établissements ayant numérisé ou souhaitant numériser une partie de leurs collections, mais qui ne disposent pas de plateforme en ligne qui soit capable de diffuser ces collections numérisées. Cette option serait alors la moins coûteuse pour la BIS. Cependant, elle comporte des limites du point de vue de la valorisation des collections. Ces limites sont causées par le rôle pivot joué par les catalogues de la BnF (Bibliothèque nationale de France) dans le dispositif. En effet, lorsqu'une édition imprimée est déjà signalée dans le catalogue général de la BnF, c'est la notice de ce catalogue qui sert de sources aux métadonnées descriptives qui accompagnent le document, sans la possibilité de les compléter par les données propres à l'exemplaire reproduit (la reliure, la présence d'annotations manuscrites...).

La seconde piste consistait à faire appel à des prestataires externalisées qui permettrait à la BIS de développer sa propre bibliothèque numérique. Le coût d'une telle prestation est estimée alors à 100 000 euros. Compte tenu des difficultés budgétaires auxquelles la Sorbonne se trouve alors confrontée, cette piste est abandonnée elle aussi. Finalement, c'est une solution moins onéreuse reposant sur l'utilisation du logiciel libre Omeka qui est adoptée. Celle-ci est réputée pour sa prise en main aisée et déjà été choisie par plusieurs bibliothèques de l'enseignement supérieur. Le projet de la mise en place de la bibliothèque numérique a été réalisé en 2016 par Mylène Ravereau, étudiante en master « Technologies numériques appliquées à l'histoire », dans le cadre de son stage de fin d'études à l'École nationale des chartes. Ce projet s'est déroulé sous la supervision conjointe du Département des manuscrits et des livres anciens et du Service de l'informatique et des systèmes d'information¹².

10. Bobis, Laurence, et Boris Noguès. *La Bibliothèque de la Sorbonne, 250 ans d'histoire au cœur de l'Université*. Éditions de la Sorbonne. Paris, France, 2022. p. 267.

11. <https://www.bnf.fr/fr/gallica-marque-blanche>.

12. Bibliothèque numérique de la Sorbonne. « Équipe ». Consulté le 9 août 2024. <https://nubis.bis-sorbonne.fr/page/equipe>.

Il fallait ensuite s’assurer de la diffusion et de la pérennisation des données, de paramétrier l’application et d’intégrer les données numérisées sur Omeka. Cette intégration a nécessité le renommage systématique des fichiers suivant un plan de nommage conforme aux préconisations de la norme ISO 9660, leur conversion aux formats de diffusion qui sont requis (jpeg et pdf) ainsi que la création de métadonnées respectant le guide d’interopérabilité rédigé par la BnF sur le protocole OAI-PMH¹³.

La bibliothèque numérique NuBIS a finalement été inaugurée le 25 avril 2017 avec plus de 40 000 images provenant de la numérisation d’environ 1 000 documents issus des fonds patrimoniaux de la bibliothèque.¹⁴ Début 2024, elle rassemble plus de 11 000 documents, totalisant plus de 450 000 images. NuBIS est régulièrement enrichie au fil de l’année par la numérisation de nouveaux documents patrimoniaux conservés à la Bibliothèque interuniversitaire de la Sorbonne.¹⁵ La bibliothèque numérique est aujourd’hui encore gérée par l’intermédiaire du logiciel Omeka S qui a succédé à Omeka¹⁶ tandis que les images sont hébergées sur le visualiseur IIIF Mirador¹⁷. Elle est hébergée sur un serveur de l’université Paris 1 Panthéon-Sorbonne qui relève du département du Système d’information et des usages numériques (DSIUN).

La Sorbonne a par ailleurs fait le choix de placer les données de sa bibliothèque numérique sous une licence ouverte conçue par Etalab tout en préservant son service de numérisation à la demande qui lui est payant. En outre, dans la perspective d’une politique de numérisation plus large, la gestion et la coordination de ce nouvel outil ont été confiées à une nouvelle mission nommée « Valorisation des collections et soutien à la recherche ». Cette mission a ensuite été renforcée au point de devenir un service à part entière sous le nom de « Service de la valorisation numérique des collections et du soutien à la recherche » (SERVAL) évoqué précédemment.

13. https://www.bnf.fr/sites/default/files/2019-02/Guide_oaipmh.pdf.

14. Bobis, Laurence, et Boris Noguès. *La Bibliothèque de la Sorbonne, 250 ans d’histoire au cœur de l’Université*. Éditions de la Sorbonne. Paris, France, 2022. p. 269.

15. Bibliothèque numérique de la Sorbonne. « Corpus numérisés ». Consulté le 9 août 2024. <https://nubis.bis-sorbonne.fr/page/le-corpus>.

16. « Omeka S ». Consulté le 9 août 2024. <https://omeka.org/s/>.

17. « Mirador — Home ». Consulté le 9 août 2024. <https://projectmirador.org/>.

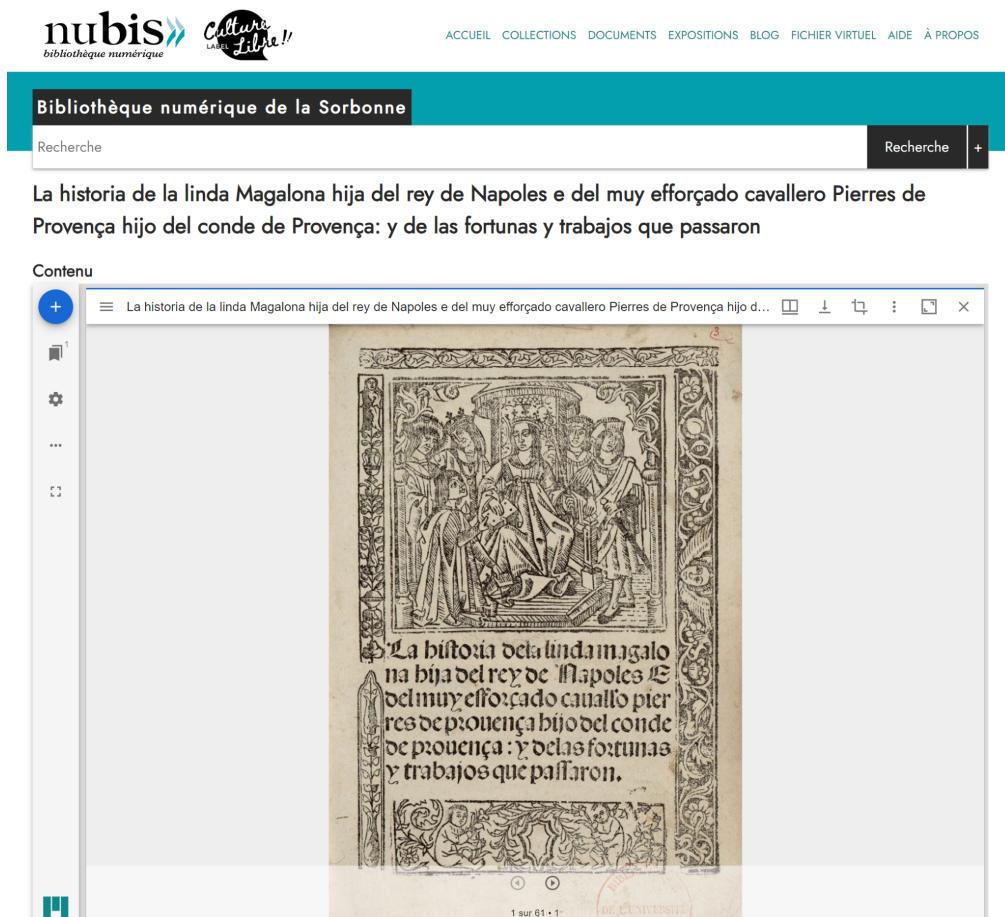


FIGURE 1.1 – Capture d’écran de l’interface de NuBIS avec Mirador

 A screenshot of the Omeka S dashboard. On the left, a dark sidebar contains navigation links for "NuBIS", "CONNEXION COMME Kutay Sefil", "Se déconnecter", "Rechercher des contenus", "Trouver des ressources...", "Sites", "RESSOURCES", "Contenus", "Collections", "Vocabulaires", "Modèles de ressource", "ADMIN", "Utilisateurs", "Modules", "Tâches", "Paramètres", "Fichiers", "Journaux", "MODULES", "Moteurs de recherche", "Ark", "Export en lot", "Import en lot", and "Import CSV". The main content area is titled "Tableau de bord" and includes a "Connexion réussie" message and a green banner stating "Une nouvelle version d'Omeka S est disponible. [Obtenir la nouvelle version](#)". Below this, a message says "Bienvenue dans le tableau de bord NuBIS !". There are two sections: "Gérer les ressources" and "Gérer les sites". "Gérer les ressources" lists "Contenus (11444)", "Collections (146)", "Vocabulaires (10)", and "Modèles de ressource (9)". "Gérer les sites" lists numerous items such as "Bibliothèque numérique de la Sorbonne", "Le Japon et la mer : une cartographie asiatique", "La Roumanie d'Emmanuel de Martonne : carnets de terrain", etc., each with a "Voir" link.

FIGURE 1.2 – Capture d’écran de l’interface d’Omeka S

Des efforts considérables ont été effectués par la suite afin d'améliorer la visibilité de NuBIS, par exemple avec la création systématique de liens pointant vers la bibliothèque numérique depuis les notices se trouvant les catalogues collectifs comme Sudoc pour les imprimés ou Calames pour les manuscrits. Un autre exemple est la mise en place d'expositions virtuelles qui reprennent le plus souvent les contenus des expositions présentées dans les murs de la bibliothèque avec la participation d'enseignants-rechercheurs. Une stratégie de valorisation plus large visant une communication à travers les réseaux sociaux et un enrichissement de Wikipédia a aussi été initiée. Plus important encore, NuBIS est désormais moissonnée par Gallica grâce à la mise en place d'un entrepôt OAI, ainsi que par le moteur de recherche Isidore (spécialisé dans le domaine des sciences humaines et sociales) et enfin par le portail Biblissima (spécialisé dans le patrimoine écrit du Moyen Âge et de la Renaissance en Occident).¹⁸

Les collections disponibles sur NuBIS sont donc issues d'un chantier de numérisation des collections patrimoniales mené depuis une quinzaine d'années. À la rentrée universitaire 2017, le chantier d'alimentation initiale qui consistait à traiter de manière rétrospective l'ensemble des données libres de droits accumulées depuis le début des activités de numérisation internes à la bibliothèque a été quasiment achevé. Désormais, les travaux de reproduction nouvellement effectués par l'atelier de numérisation sont versés et publiés au fil de l'eau sur NuBIS.

La politique de numérisation de la bibliothèque est aujourd'hui définie par le Département des manuscrits et des livres anciens. Elle consiste à numériser les livres rares et les unica de la collection de la BIS c'est-à-dire les livres dont un seul exemplaire est répertorié dans le catalogue Sudoc. Les points forts sont concentrés essentiellement sur l'histoire de l'université à Paris dans toutes les périodes de son histoire et sous tous ses aspects : institution, pédagogie, enseignants, vie étudiante, bâtiments, notamment par la diffusion des archives de l'Université de Paris et des archives de la bibliothèque de la Sorbonne. On y retrouve aussi des numérisations en lien avec les projets de recherche et les expositions, mais dans une quantité bien moindre. Un changement dans cette politique est prévu avec l'introduction d'un marché de numérisation l'an prochain¹⁹ mais à l'heure actuelle, la numérisation s'organise autour de neuf collections en plus des collections invitées qui sont hébergées dans la bibliothèque numérique mais ne sont pas issues des collections conservées à la BIS²⁰ :

18. Bobis, Laurence, et Boris Noguès. *La Bibliothèque de la Sorbonne, 250 ans d'histoire au cœur de l'Université*. Éditions de la Sorbonne. Paris, France, 2022. p. 271.

19. D'après un échange avec Cécile Obligie et Laurie Aoustet, conservatrices à la BIS.

20. Bibliothèque numérique de la Sorbonne. « Corpus numérisés ». Consulté le 9 août 2024. <https://nubis.bis-sorbonne.fr/page/le-corpus>.

- Les sources de l'histoire de l'Université de Paris mentionnées précédemment qui regroupe des documents relatifs à l'évolution des bâtiments de l'Université et des collèges qui lui étaient rattachés ainsi qu'aux enseignements qui y étaient dispensés.
- Des documents manuscrits, imprimés et iconographiques dédiés à l'histoire de la bibliothèque et mis en ligne à l'occasion du 250e anniversaire de la BIS.
- Des manuscrits médiévaux hérités des bibliothèques médiévales de plusieurs collèges parisiens.
- Des papiers d'universitaires et d'érudits qui valorisent les documents des fonds comme celles de Léon Chestov ou encore de la marquise Arconati-Visconti.
- Des cartes postales sous forme de caricatures de la Première Guerre mondiale.
- Des cartes géographiques rares allant du XVI^e au XIX^e siècles.
- Des imprimés et manuscrits de l'école française de géographie.
- Des manuscrits philosophiques clandestins du XVIII^e siècle.
- Depuis mai 2024, « Matrimoine » qui est une collection d'écrits de femmes ou d'images témoignant de la présence des femmes à l'université.

NuBIS se positionne donc dans le paysage foisonnant des bibliothèques numériques d'aujourd'hui avec des atouts considérables. Elle fait face aux nouveaux enjeux que représente la valorisation des collections par la recherche. Grâce à sa politique de numérisation structurée, elle s'engage à soutenir des projets de recherche en prenant appui sur ses collections. Outre une organisation sous forme de collection, les documents en ligne sur NuBIS peuvent être distingués selon les catégories suivantes qui renseignent sur leur type : manuscrits, monographies imprimées, cartes postales, photographies, documents cartographiques, estampes, dessins, et enfin tapuscrits. Le diagramme ci-dessous montre la répartition des documents sur NuBIS en fonction de leur type :

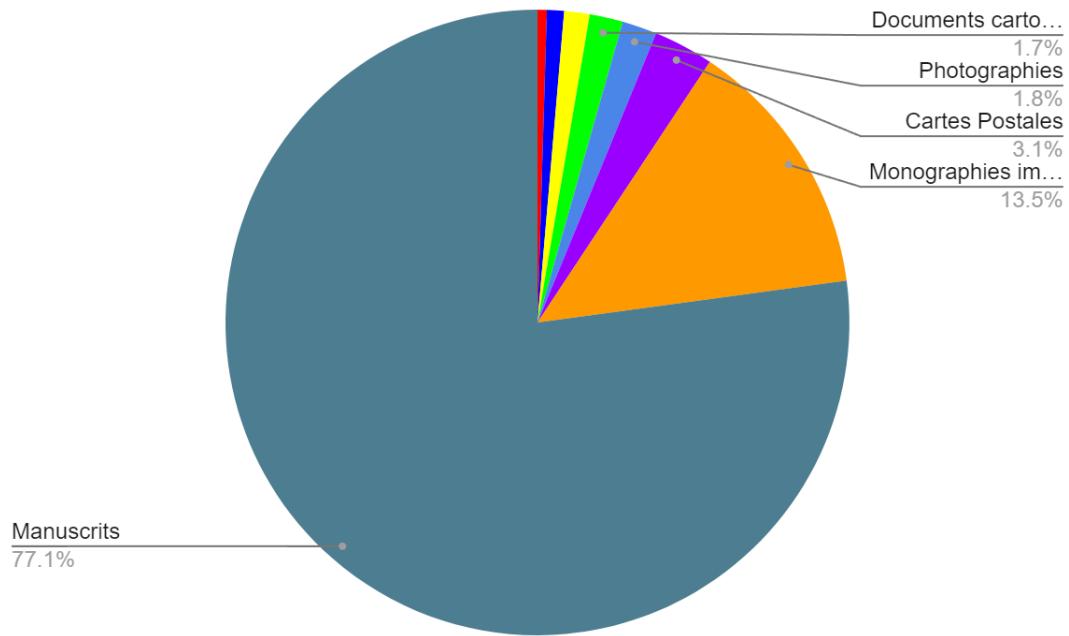


FIGURE 1.3 – Répartition des documents par type sur NuBIS

Par définition, la reconnaissance optique de caractères (OCR) concerne uniquement les documents imprimés. Ainsi, nous nous intéresserons ici uniquement aux monographies imprimées (en orange) ainsi qu'aux tapuscrits (en rouge), ce qui représente environ 1500 documents soit 14% de l'ensemble des documents sur NuBIS. Il convient donc à présent de nous intéresser plus en détail à cette collection afin d'étudier la qualité attendue de l'OCR sur cet ensemble de documents. Pour cela, nous avons jugé pertinent d'établir une typologie de ces imprimés et tapuscrits selon leurs caractéristiques matérielles et typographiques, lesquelles peuvent influer sur la qualité de l'OCR et donc sur le choix du logiciel d'OCR.

IV. Typologie des documents à océriser

En ce qui concerne les tapuscrits tout d'abord, il en existe actuellement 52 sur NuBIS dont 33 sont des lettres dactylographiées du début du XX^e siècle adressées à la marquise Arconati-Visconti, similaires entre elles et présentant une impression de bonne qualité. De son nom de naissance Marie Peyrat (1840-1923), elle est la fille de l'homme de lettres Alphonse Peyrat, journaliste républicain devenu après 1870 député puis sénateur. En 1873, elle devient par mariage la marquise Arconati-Visconti. Son mari, le jeune Giammartino, issu d'une des plus grandes familles italiennes et propriétaire de nombreuses propriétés à travers l'Europe, décède trois ans plus tard sans laisser d'héritier, léguant à la nouvelle marquise une immense fortune. Elle utilise cet héritage pour entretenir ses résidences, devenir une mécène, mais surtout pour soutenir la jeune République qui voit le jour en

France en 1870 dans les domaines de l'art et des savoirs. Passionnée par l'histoire et la politique, dreyfusarde, Marie Arconati-Visconti réunit autour de sa table les principaux acteurs de la Troisième République : députés, ministres, hauts fonctionnaires, ainsi que des savants et professeurs du Collège de France, de l'École des chartes et de l'École des hautes études²¹. Ses correspondances entretenues avec ces figures notables de la Troisième République ont été numérisées après un chantier d'un an et mises en ligne sur NuBIS en décembre 2019.²² Elles ont par ailleurs donné lieu à dans le cadre d'une exposition virtuelle.²³

Les autres tapuscrits ont une typographie très lisible eux aussi car ils sont exempts de toute lettre ou caractère archaïque puisque les machines à écrire sont une invention relativement récente, datant de la fin du XIX^e siècle²⁴. Cependant, les tapuscrits présentent parfois un problème particulier ce type de support : les caractères imprimés au verso sont lisibles au recto et vice-versa, ce qui peut causer des difficultés de lecture à un logiciel d'OCR. Hormis ce point, la qualité visuelle des tapuscrits en ligne sur NuBIS ne pose pas de difficulté pour l'OCR avec une typographie qui est visuellement très proche d'un tapuscrit à l'autre.

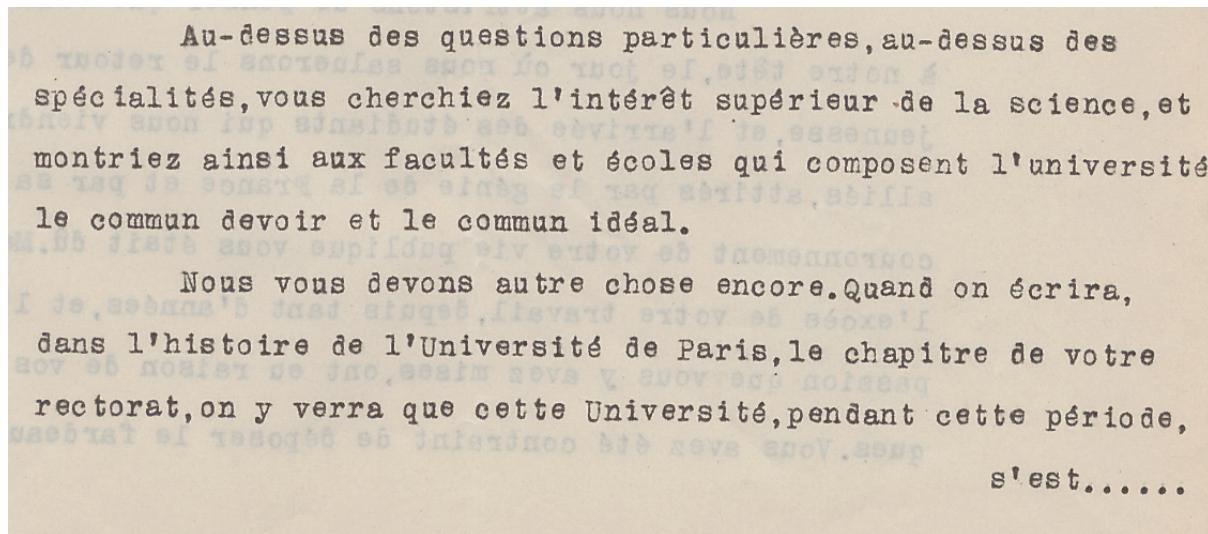


FIGURE 1.4 – Exemple de tapuscrit avec le texte au verso lisible (<https://nubis.bis-sorbonne.fr/ark:/15733/9qp6?mirador-1=1>)

21. DERROT, Sophie. « La marquise Arconati Visconti », 22 novembre 2023. http://blog.bibliotheque.inha.fr/fr/posts/marquise_arconati_visconti.html.

22. L'ensemble des correspondances manuscrites et tapuscrites se trouvent ici : <https://nubis.bis-sorbonne.fr/ark:/15733/ffzf>.

23. Site de l'exposition virtuelle : <https://nubis.bis-sorbonne.fr/web/marquise-arconati-visconti/accueil.html>.

24. Cortada, James W. *Before the Computer : IBM, NCR, Burroughs, and Remington Rand and the Industry They Created, 1865-1956*. Princeton University Press, 2015. p. 38.

Quant aux monographies imprimées, elles sont bien plus nombreuses sur NuBIS puisque nous en dénombrons plus de 1400 en ligne actuellement. Cette collection d'imprimés est ancienne et a une provenance universitaire : son noyau d'origine provient d'établissements d'enseignement et d'anciens professeurs. On peut dès le début présupposer que la précision de l'OCR sera très variable elle aussi puisqu'elle va dépendre des caractéristiques du matériel source qui sont mentionnés dans le rapport sur l'OCR du consortium IMPACT et que nous avons rappelé dans notre introduction²⁵. Ceux-ci sont en effet liés à l'aspect matériel de l'ouvrage en lui-même puisque la qualité de la numérisation est excellente sur NuBIS quelle que soit la date de publication de l'ouvrage avec des images de très haute définition en termes de pixels. Pour cette raison, il est nécessaire de distinguer tout d'abord les monographies imprimées en plusieurs grandes catégories avant d'ensuite étudier les particularités des ouvrages qui peuvent causer des difficultés pour l'OCR.

Il a été envisagé dans un premier temps d'effectuer une typologie par type d'imprimé : lettres, actes, thèses, pièces de théâtre, ouvrages scientifiques... Or, pour un même type d'imprimé, la qualité de l'OCR n'est pas la même en fonction de la date du document, la qualité de l'impression et les autres caractéristiques typographiques. Par conséquent, il nous a paru plus pertinent d'établir une typologie de nature chronologique, fondée sur la date de publication des documents. En effet, malgré les particularités propres à chaque ouvrage, nous avons observé qu'un découpage par période traduit assez bien l'évolution de la qualité de l'OCR pour la collection des imprimés : plus ces documents sont récents, meilleure est la qualité de l'OCR. Nous avons donc effectué un découpage chronologique en quatre périodes distinctes, chacune étant caractérisée par des éléments typographiques qui lui sont caractéristiques ou du moins plus récurrents que dans les ouvrages des autres périodes. Ce découpage a été validé par la responsable de la Réserve de la bibliothèque.

Les graphiques suivants montrent respectivement la répartition des imprimés par décennie et par siècle sur NuBIS :

25. Anderson, Niall, Gunter Muhlberger, et Apostolos Antonacopoulos. « Optical Character Recognition - IMPACT Best Practice Guide ». *Optical Character Recognition*, 2023. p. 8-11. https://www.digitisation.eu/wp-content/uploads/2023/09/OpticalCharacterRecognition-IBPG_01.pdf.

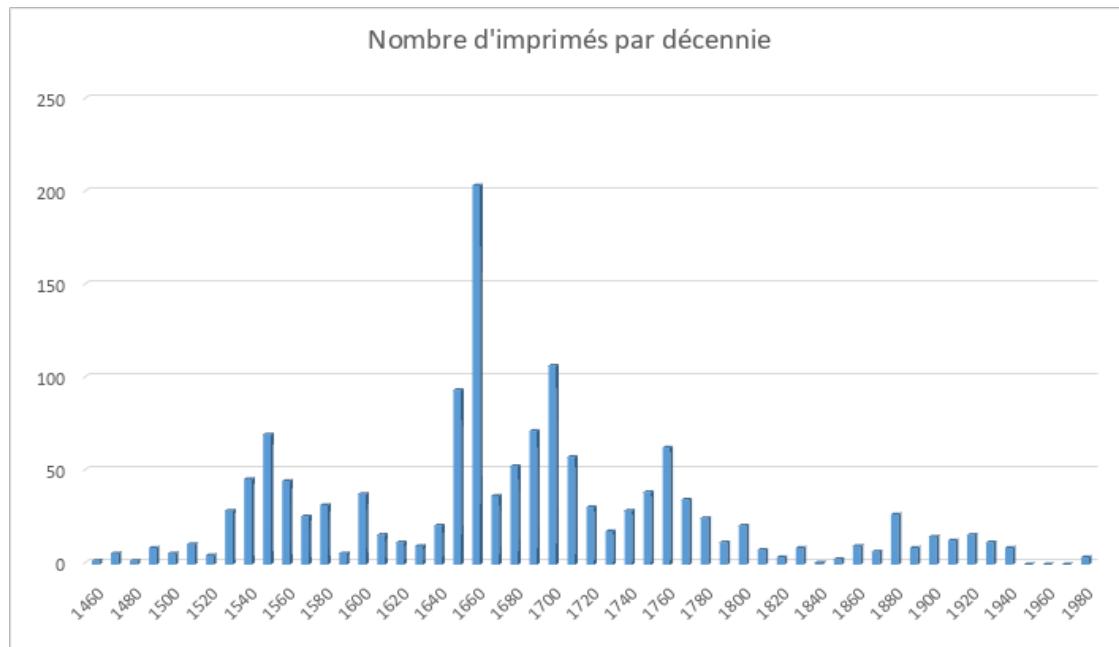


FIGURE 1.5 – Répartition du nombre d'imprimés par décennie

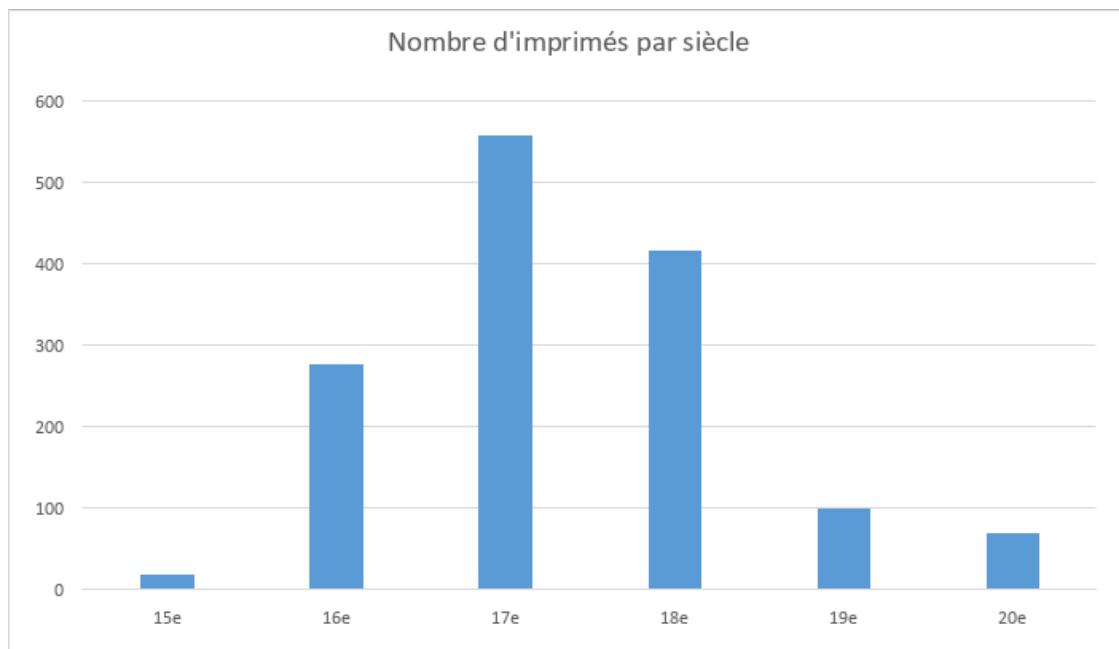


FIGURE 1.6 – Répartition du nombre d'imprimés par décennie

1. 1467 - 1539

Cette période comprend l'intégralité des incunables et une partie des documents du XVI^e siècle. La date de 1467 correspond au plus ancien document imprimé qu'on trouve sur NuBIS²⁶, tandis que celle de 1539 correspond au dernier ouvrage en écriture gothique

26. Gerson, Jean, et Ulrich Zell. *Conclusiones de diversis materiis moralibus, sive De regulis mandatorum.* Köln, France : Ulrich Zell, 1467. <https://nubis.bis-sorbonne.fr/ark:/15733/nkd1>.

actuellement disponible sur NuBIS²⁷. La présence d’incunables rares au sein de la BIS et de sa bibliothèque numérique n’est pas surprenante puisque l’apparition de l’imprimerie en France naît d’une volonté de l’Université de Paris de posséder des exemplaires de livres en nombre²⁸. Plus précisément, ce sont Guillaume Fichet et Jean Heynlin, condisciples à l’Université de Paris, qui sont à l’origine du premier livre imprimé en France en 1470 (les *Epistolae* de Gasparin de Bergame).²⁹

Ainsi, cette tranche chronologique est composée majoritairement d’ouvrages en écriture gothique, ce qui est dans la continuité des manuscrits médiévaux gothique du Moyen Âge tardif. Cette graphie représente un défi considérable pour l’ensemble des moteurs d’OCR. Par exemple, Gallica (la bibliothèque numérique de la Bibliothèque nationale de France) ne propose pas d’OCR pour les documents de cette période. De manière générale, la grande majorité des ouvrages de cette période est en latin, le reste étant en moyen français, espagnol, allemand et grec. Toujours dans la continuité de l’époque médiévale, le latin continue donc à dominer largement les écrits mais cette domination tend à s’affaiblir avec l’invention de l’imprimerie au profit des langues vernaculaires comme le français : nous passons de 97% de titres en latin en 1467 à 74% en 1477 puis 65% en 1492.³⁰. Comme l’ont montré les graphiques précédents, les ouvrages de cette période constituent seulement une très petite partie des imprimés en ligne sur NuBIS.

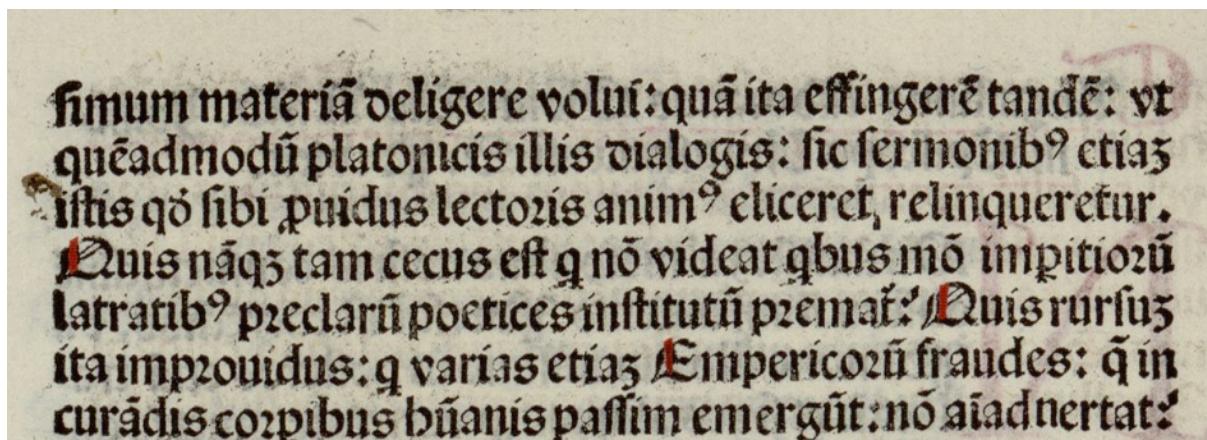


FIGURE 1.7 – Extrait d’un ouvrage de la fin du XV^e siècle. (<https://nubis.bis-sorbonne.fr/ark:/15733/ng7n?mirador-1=4>)

27. « Das Vater Unser kurz ausgelegt/ und inn Gesang weyse gebracht durch D. Mar. Luth. » Consulté le 12 août 2024. <https://nubis.bis-sorbonne.fr/ark:/15733/4frd>.

28. Barbier, Frédéric. « Chapitre 2. Gutenberg et l’invention de l’imprimerie ». *Histoire du livre en Occident*. Armand Colin, 2020. p. 114.

29. Veyrin-Forrer, Jeanne. « Hommage aux premiers imprimeurs de France. 1470-1970 », *Bulletin des bibliothèques de France (BBF)*, 1971, n° 2, p. 65-80. <https://bbf.enssib.fr/consulter/bbf-1971-02-0065-001>

30. Barbier, Frédéric. « L’invention de l’imprimerie et l’économie des langues en Europe au XVe siècle ». *Histoire et Civilisation Du Livre* 4 (2008) : p. 21-46.

Nous pouvons observer ici une graphie gothique typique des imprimés de cette période. L'ouvrage est en latin comme c'est le cas encore une fois pour la plupart des imprimés de l'époque. Nous remarquons aussi la présence de nombreuses abréviations sous la forme de caractères spéciaux, ce qui peut compliquer le travail de l'OCR. Le papier rend ce travail d'autant plus difficile puisque le l'encre et le texte de la page au dos sont lisibles ici. Enfin, les majuscules et lettrines sont en rouge et donc dans une couleur différente du reste du texte, ce qui peut impacter la qualité de l'OCR là aussi.

2. 1539 - fin du XVII^e siècle

Les ouvrages du reste du XVI^e siècle et du XVII^e siècle ont été regroupés au sein d'une même catégorie. Ceux-ci constituent environ la moitié de la collection des monographies imprimées de la NuBIS. Ces monographies sont caractérisées par une qualité d'impression souvent médiocre et des particularités récurrentes : présence d'annotations manuscrites, caractères peu lisibles, courbure du texte, abondance de l'écriture italique. De plus, l'expansion rapide de l'imprimerie en Europe à cette période a donné lieu au développement de nombreuses polices de caractères, chaque imprimerie utilisant ses propres caractères mobiles³¹. En France, nous pouvons citer Claude Garamond, l'un des premiers graveurs et fondeurs typographiques indépendants du royaume, qui produit des fontes romaines et italiennes employées par Robert Estienne à partir des années 1540³².

Cette multitude de polices de caractères influent ainsi de manière négative sur les performances des outils d'OCR. Environ les trois quarts des ouvrages de cette tranche chronologique sont en latin et un quart en français. Pourtant, l'historien du livre Henri-Jean Martin a montré que l'équilibre entre le latin et le français s'inversait déjà en faveur de cette dernière dans la décennie 1560³³. La domination du latin au seins des imprimés de la BIS encore à cette période peut s'expliquer par l'abondance des imprimés universitaires et religieuses où la langue latine restait la référence. Les exemples ci-dessous, pris dans la période concernée, illustrent des cas récurrents susceptibles de diminuer la qualité de l'OCR.

31. Heil, Jacob, et Todd Samuelson. « Book History in the Early Modern OCR Project, or, Bringing Balance to the Force ». *Journal for Early Modern Cultural Studies* 13, n° 4 (2013) : p. 90-103.<https://www.jstor.org/stable/jearlmodcultstud.13.4.90>.

32. Barbier, Frédéric. « Chapitre 4. L'Ancien Régime : formes de l'imprimé ». *Histoire du livre en Occident*. Armand Colin, 2020. p. 239.

33. « Chapitre 2. Le paradigme de l'absolutisme : l'Europe classique et l'imprimé ». *Ibid.*, p. 197.

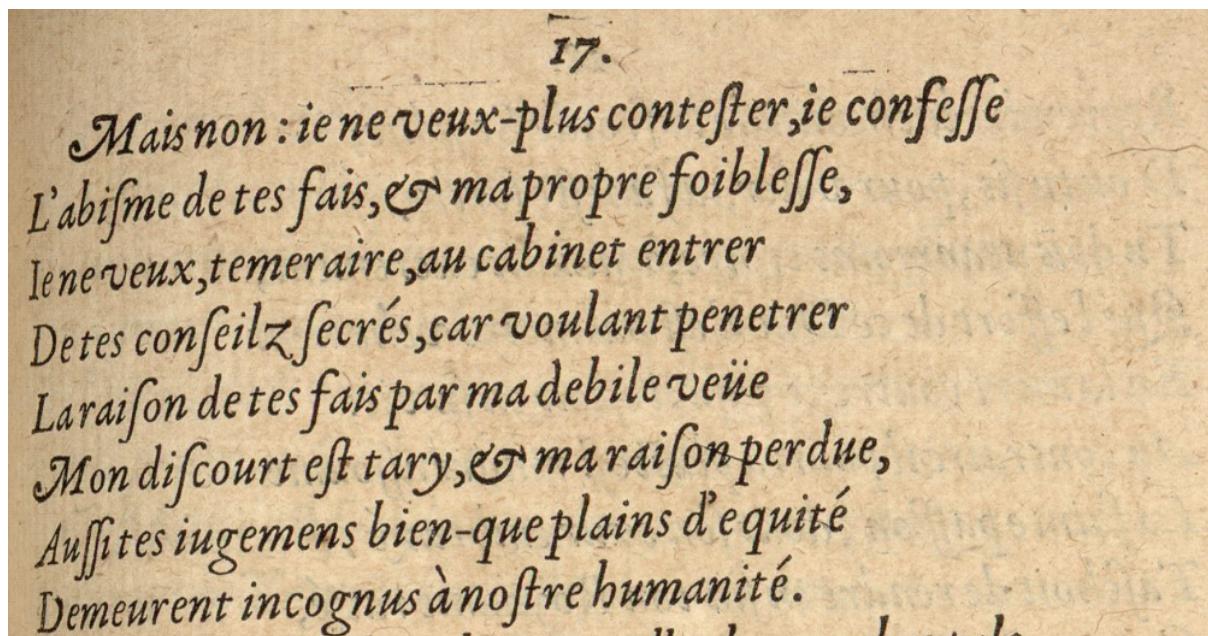


FIGURE 1.8 – Extrait d'un ouvrage du début XVII^e siècle. (<https://nubis.bis-sorbonne.fr/ark:/15733/49bk?mirador-1=17>)

Dans l'extrait ci-dessus, le texte en italique et la courbure de la page sur le côté gauche diminuent grandement la lisibilité du texte. L'espacement des mots est lui aussi difficile à identifier à certains moments et le texte au dos de la page ressort encore une fois.

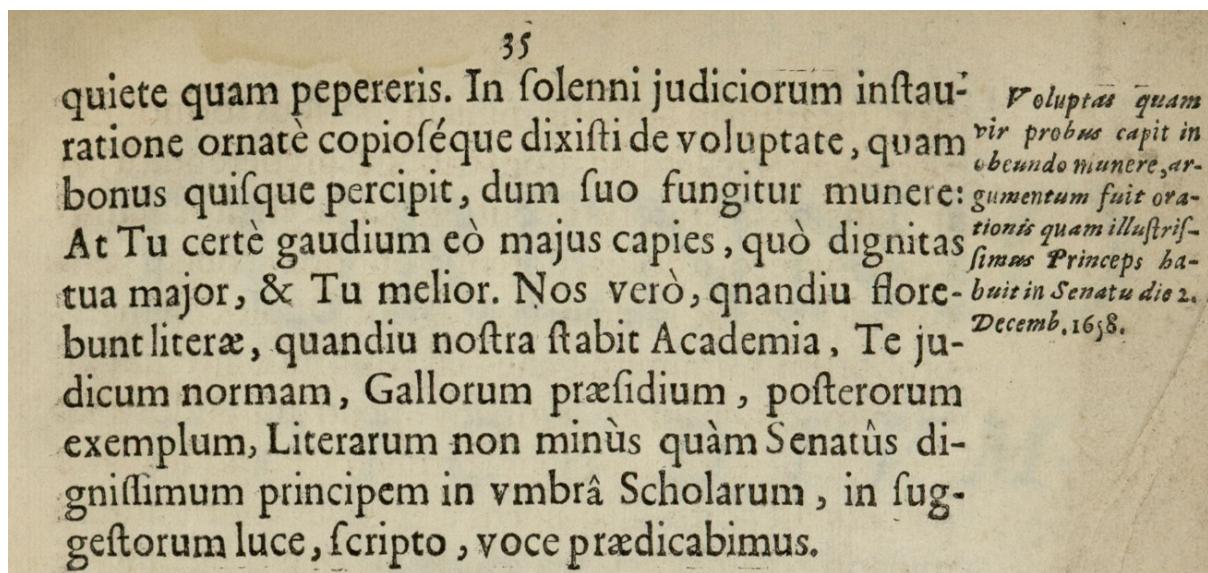


FIGURE 1.9 – Extrait d'un ouvrage du milieu du XVII^e siècle. (<https://nubis.bis-sorbonne.fr/ark:/15733/1khw?mirador-1=43>)

Dans cet exemple, nous remarquons la présence d'un texte en italique situé en marge sur le côté droit, compliquant ainsi la segmentation de la page. En outre, ces notes en italique sont peu lisibles dû à leur petite taille.

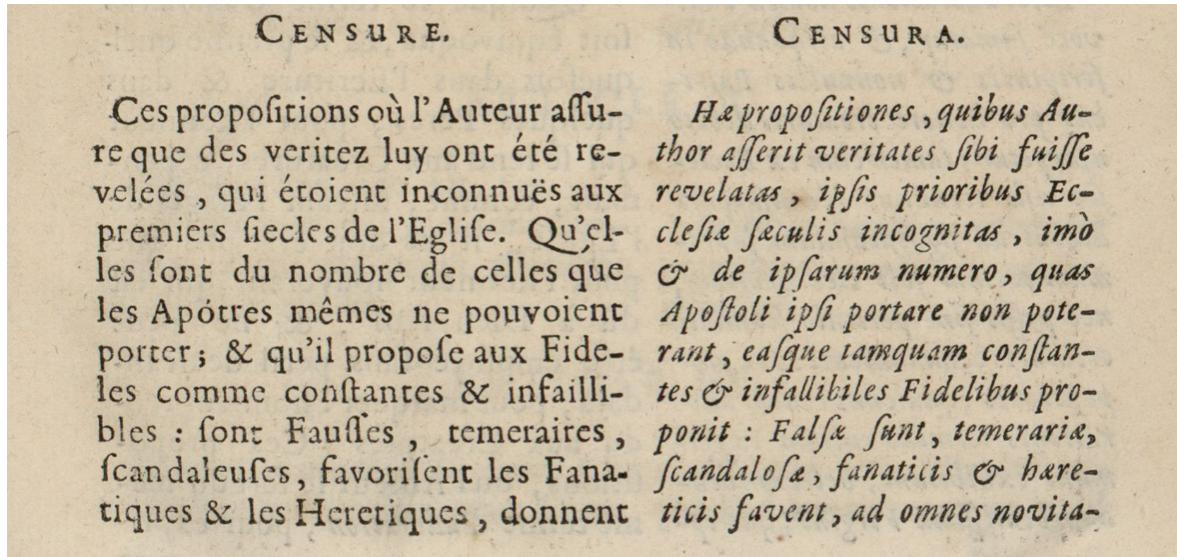


FIGURE 1.10 – Extrait d'un ouvrage de la fin du XVII^e siècle. (<https://nubis.bis-sorbonne.fr/ark:/15733/17zw?mirador-1=7>)

Enfin, nous pouvons voir ici un exemple de texte réparti sur deux colonnes au sein d'une même page. Non seulement les styles sont différents (romain et italien), mais les langues le sont aussi (français et latin). Cela complique à la fois la segmentation de la page et la lecture des caractères.

3. XVIII^e siècle

Le siècle suivant est marqué par une qualité typographique bien meilleure et un faible usage de l'italique, ce qui facilite grandement le travail d'OCR. Le XVIII^e siècle voit en effet une nouvelle grande vague de travail sur le dessin des caractères sous le règne de Louis XIV avec l'influence du directeur de l'Imprimerie royale Jean Anisson et les travaux du graveur Philippe Grandjean : c'est la naissance du « Romain du Roi ». Il devient la référence pour le siècle des Lumières avec ses caractères plus raides qui sont qualifiés de « mathématiques » dû à la conception géométrique des lettres³⁴. Toutefois, l'usage des « s » longs constitue toujours une difficulté pour la plupart des moteurs d'OCR puisque ces caractères sont souvent confondus avec des « f ». La moitié environ des imprimés de cette période est en latin tandis que le reste est en français.

34. « Chapitre 4. L'Ancien Régime : formes de l'imprimé ». *Ibid.*, p. 241.

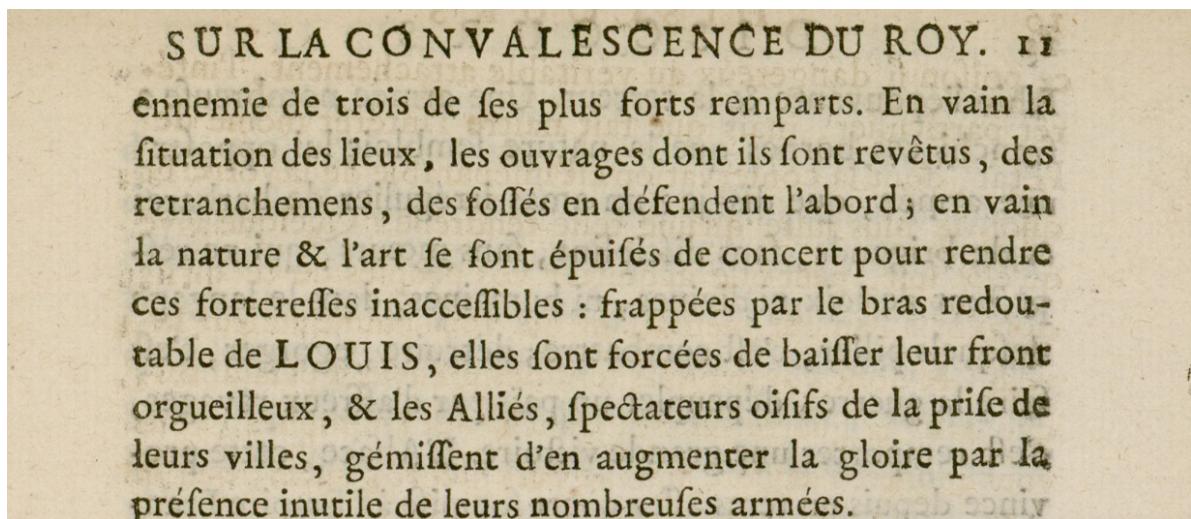


FIGURE 1.11 – Extrait d'un ouvrage typique du XVIII^e siècle. (<https://nubis.bis-sorbonne.fr/ark:/15733/1181?mirador-1=11>)

Dans l'exemple ci-dessus, nous pouvons voir que l'impression est de bonne qualité mais il existe toujours des imperfections au niveau de l'encre avec notamment la transparence du texte au dos de la page. En outre, des caractères archaïques tels que le « s » longs et les ligatures « ct » subsistent encore.

4. XIX^e et XX^e siècles

Un peu moins de 200 ouvrages datent des XIX^e et XX^e siècles, la quasi-totalité étant en français. Les imprimés de cette période sont majoritairement des ouvrages scientifiques, parfois numérisés dans le cadre des expositions à la BIS. Ils comportent souvent des images, Frédéric Barbier qualifiant le XIX^e siècle du « siècle de l'image »³⁵. En effet, le début du siècle voit un retour spectaculaire de la xylographie (la gravure sur bois) avec l'invention de la gravure sur bois de bout par l'anglais Thomas Bewick qui permet d'imprimer conjointement texte et image de façon efficace. D'autres inventions telles que la lithographie et l'héliographie viennent contribuer à l'émergence de l'image au sein des livres datant de cette période. De manière générale, la qualité excellente de la typographie et l'abandon des « s » longs et autres caractères désuets rendent le travail d'OCR sensiblement plus simple. L'exemple ci-dessous illustre bien cela :

35. « Chapitre 3. Le XIXe siècle industriel ». *Ibid.*, p. 325.

Un des meilleurs moyens pour les boursiers de montrer leur reconnaissance au collège qui les avait recueillis était d'enrichir la bibliothèque de manuscrits copiés par eux; les volumes de la bibliothèque de l'Université en offrent de nombreux exemples.

On peut conjecturer, d'après la souscription mentionnée ci-dessus, que sous l'influence de

FIGURE 1.12 – Extrait d'un ouvrage de la fin du XIX^e siècle représentatif des imprimés de cette période avec aucune difficulté éventuelle pour l'OCR. (<https://nubis-bis-sorbonne.fr/ark:/15733/17b9?mirador-1=12>)

Ainsi, la répartition chronologique des imprimés dans la bibliothèque numérique de la BIS et les quelques extraits que nous avons vus nous montrent qu'une bibliothèque interuniversitaire ayant un fonds patrimonial important comme la BIS doit disposer d'un outil d'OCR performant sur les documents anciens et qui soit capable de surmonter différentes qualités d'impression et des particularités typographiques récurrentes dans ces ouvrages anciens. En particulier, nous nous attendons à ce que les ouvrages antérieurs au XVIII^e soient ceux qui causent le plus de difficultés à l'OCR. Pour vérifier ces hypothèses, il est nécessaire d'étudier les solutions d'OCR qui existent actuellement sur le marché et déterminer lesquelles sont les plus adaptées pour une bibliothèque numérique comme NuBIS et sa collection d'imprimés.

Chapitre 2

Une solution d'OCR adaptée à la BIS

I. État des lieux

Avant de commencer à étudier les différentes solutions d'OCR qui existent sur le marché, nous avons tout d'abord établi un état des lieux en examinant ce que les bibliothèques disposant d'une collection numérique similaire à NuBIS ont adopté comme solution pour leur traitement OCR. Pour cela, nous avons interrogé principalement des établissements gérant des bibliothèques numériques patrimoniales qui proposent un fonds de documents anciens. Le tableau suivant montre les données recensées auprès des bibliothèques à la suite de notre enquête (y sont inclus uniquement les bibliothèques ayant répondu à nos questions) :

	OCR	Solution d'OCR
Gallica (Bibliothèque nationale de France)	Oui sous forme de transcription	En interne : ABBYY FineReader
Bibliothèque patrimoniale de la Côte d'Azur	Oui pour certains pdf	- Avec un prestataire (Arkhénum) entre 2016 et 2019 - En interne depuis 2019 : Adobe Acrobat Pro
Bibliothèque patrimoniale de Toulouse	Oui pour certains pdf	En interne : LIMB Processing avec ABBYY
Bibliothèque patrimoniale de l'Alsace	Oui sous forme de text brut	Non connu
Bibliothèque patrimoniale de Lille	Oui pour les pdf	En interne : LIMB Processing avec ABBYY 10 Engine
Bibliothèque patrimoniale de Grenoble	Oui pour les pdf	En interne : Abbyy FineReader 15 sur les documents post-1800 avec correction manuelle
Bibliothèque patrimoniale de Bordeaux	Oui pour les pdf	Pas de réponse
Bibliothèque CUJAS	Oui pour certains pdf	En interne jusqu'en 2015 : Abbyy FineReader 9 et correction manuelle par les magasiniers

TABLE 2.1 – Liste des bibliothèques numériques avec OCR examinées dans le cadre de cette étude

Nous constatons que, parmi les bibliothèques ayant répondu à notre enquête, la grande majorité utilise le logiciel ABBYY pour leur traitement OCR. Ainsi, il semble y avoir un quasi-consensus dans le paysage des bibliothèques numériques étudiées en ce qui concerne le choix du logiciel pour ce traitement. Nous estimons que cela est dû à l'ancienneté du logiciel et à son utilisation par la Bibliothèque nationale de France puisque qu'ABBYY est un des premiers logiciels dans le domaine de l'OCR comme nous le verrons juste après, ce qui peut expliquer la fiabilité du logiciel chez les autres bibliothèques.

Concernant Gallica, un entretien en visioconférence a été effectué avec Sébastien Cretin qui est le chef de projet OCR à la BnF. Cet échange nous a fourni des infor-

mation des informations utiles dans le cadre du stage et de notre problématique. Nous avons notamment appris que le processus d’océrisation se faisait par l’intermédiaire d’un prestataire jusqu’en avril 2019, mais qu’il se fait dorénavant en interne avec ABBYY FineReader dans l’objectif de réduire les coûts, qui est pour rappel l’un des critères que nous avons énoncés dans notre introduction. En outre, l’internalisation de l’OCR permettrait un fonctionnement plus souple et plus pratique selon Sébastien Cretin. Leur licence ABBYY leur permet d’océriser 20 millions de pages A4 par an (ce qui est largement au-dessus de notre volumétrie qui est de 124 459 pages pour les imprimés et tapuscrits). Ensuite, lorsque nous lui avons demandé quels étaient les critères sur lesquels la BnF se basait afin de sélectionner les documents à océriser, nous avons appris que ce processus était effectué de manière automatisée en se basant sur trois critères :

- La langue du document.
- La date de publication du document
- Le « type » du document (les monographies par exemples).

II. Réalisation des tests

1. Méthodologie

Dans un second temps, des tests ont été réalisés sur un ensemble de logiciels d’OCR afin d’évaluer leur performance au regard de la collection présente dans cette bibliothèque. Ces tests ont constitué la majorité des travaux effectués pendant la période du stage. Ainsi, cinq outils différents ont été évalués dans le cadre de ces tests :

- Tesseract¹ : un logiciel *open source* directement intégré en tant que module sur Omeka S, outil utilisé par la BIS pour gérer sa bibliothèque numérique. Le logiciel a été publié en 2005 et son développement a été pris en charge par Google de 2006 à novembre 2018. Il s’agissait du logiciel envisagé par la BIS pour son OCR au commencement du stage car la bibliothèque avait eu des bons retours sur ce logiciel.
- ABBYY FineReader² : un logiciel payant pour lequel la BIS dispose d’une licence sous sa version 16 et qui est utilisé par les entreprises et bibliothèques. C’est l’un des logiciels pionniers dans le domaine de l’OCR puisqu’il est disponible depuis 1993. Il s’agissait du second logiciel choisi pour les tests après que nous ayons observé le consensus parmi les bibliothèques interrogées.

1. <https://github.com/tesseract-ocr/tesseract>.
2. <https://pdf.abbyy.com/>.

- Nanonets³ : un logiciel payant à destination des entreprises qui souhaitent traiter des documents de type administratif (factures, reçus...). L’entreprise éponyme à l’origine du logiciel est relativement jeune puisqu’elle a été fondée en 2017. Cet outil a été découvert à la suite de recherches de nouveaux logiciels d’OCR sur le Web.
- eScriptorium⁴ : une application web *open source* développée par l’université PSL (Paris Sciences & Lettres) et spécialisée dans l’HTR avec le moteur Kraken. Sa première version a été publiée en 2018. Ce logiciel a été choisi car nous avons voulu tester l’efficacité des outils d’HTR sur des imprimés.
- Transkribus⁵ : une application Web payante développée par l’université d’Innsbruck. Elle a été choisie pour la même raison qu’eScriptorium puisque Transkribus est spécialisée dans l’HTR elle aussi. La plateforme a été mise en point dans le cadre des deux projets européens transScrip-torium (2013-2015) et READ (Recognition and Enrichment of Archival Documents – 2016-2019).

La méthodologie utilisée pour les tests a été la suivante : a été sélectionné un imprimé en ligne sur NuBIS tous les vingt ans à partir de 1600 et jusqu’au XX^e siècle, soit en tout 20 imprimés, avec à chaque fois un échantillon de trois pages à transcrire manuellement. A été ajouté à ces imprimés un tapuscrit du début du XX^e siècle. Ensuite, les logiciels d’OCR ont été appliqués sur les pages concernées et le taux de précision au mot pour chaque document a été calculé grâce à la commande *wordacc* de l’outil ocreval⁶ sur Ubuntu en lui fournissant en entrée la transcription manuelle d’une part et celle réalisée par l’OCR d’autre part. Dans le cas d’eScriptorium et Transkribus, les modèles utilisés pour la transcription sont respectivement CATMuS Print et Transkribus Print M1 car ce sont celles qui ont obtenu les meilleurs résultats pour chaque logiciel.

Pour ces tests, il a été décidé de ne pas comptabiliser les mauvaises lectures des « s » longs comme des erreurs s’ils étaient remplacés par des lettres « f » car cela risquait de trop pénaliser les taux de précisions pour les outils d’OCR incapables de les détecter. En effet, nous ne voulions pas que les taux calculés soient « faussés » par une même erreur qui est récurrente dans les ouvrages les plus anciens pour les logiciels concernées. De même, il a été décidé de ne pas prendre en considération dans le calcul du taux de précision l’espacement entre les mots car l’utilisateur effectue le plus souvent la recherche en plein texte sur des mots uniques au sein des documents plutôt que sur des ensembles de mots

3. <https://nanonets.com/>.

4. <https://escriptorium.inria.fr/>.

5. <https://www.transkribus.org/>.

6. <https://github.com/eddieantonio/ocreval>.

ou des phrases. En outre, tout ce qui relève de la mise en page et qui n’entre donc pas dans le corps du texte n’est pas pris en compte. Cependant, ces éléments n’étant pas non plus négligeables dans l’estimation de la qualité d’un logiciel d’OCR, nous les détaillerons tout de même dans la suite du document.

2. Corpus

La construction du corpus comprenant les 21 échantillons s'est faite en concertation avec différentes conservatrices à la BIS. L'objectif était d'établir un corpus qui soit représentatif des difficultés que peuvent poser la collection d'imprimés disponible sur NuBIS aux différents logiciels d'OCR qui allaient être testés. Il était donc nécessaire de choisir des documents et des pages qui soient suffisamment variés sur le plan typographique du texte et sur l'aspect matériel des ouvrages. Nous pouvons organiser ce corpus selon le même découpage chronologique effectué dans la partie précédente consacrée à la typologie des documents à océriser.

1467 - 1539

Aucun document du corpus ne figure dans cette tranche chronologique. La raison pour cette absence est qu'au moment de la construction du corpus, nous ne pensions pas qu'il était envisageable d'appliquer l'OCR sur des incunables et imprimés aussi anciens, en particulier ceux en écriture gothique. L'OCR étant encore au stade zéro à la Sorbonne, et les ouvrages de cette période peu nombreux sur NuBIS, il avait donc été de laisser de côté les imprimés les plus anciens pour le moment à l'instar de ce que fait Gallica. Toutefois, des tests sur les imprimés gothiques ont aussi finis par être réalisés ultérieurement durant le stage sur les logiciels d'HTR pour évaluer leur efficacité sur cette graphie particulière. Bien que ne figurant pas dans le corpus, les résultats de ces tests seront tout de même commentés par la suite.

1539 - fin du XVII^e siècle

1. *Petri Rami Veromandui rhetoricae distinctiones, ad Carolum Lotharingum cardinalem Guisianum. Oratio ejusdem de studiis philosophiae & eloquentiae conjungendis.* Cet ouvrage en latin de Pierre de La Ramée date de 1549. Il a comme caractéristique un texte en italique du 16e siècle avec le texte de la page imprimée au verso qui est visible au recto.
2. *Les larmes publiques sur le trespass de feu tres-haut, tres-valeureux & redouté prince, Philippe Emanuel de Lorraine, duc de Mercoeur, & de Penthievre, prince du S. Empire, de Martigues, pair de France, marquis*

de Nominy, Bauge, &c. Lieutenant general de la Majesté imperiale és armées d’Hongrie contre les infideles. Cet ouvrage en français de 1602 par Alphonse de Rambervillers contient du texte en italique avec une courbure de la page sur les bordures intérieures.

3. *Le chois des epistres de Lipse.* Cet imprimé en français de Juste Lipse date de 1619 et a pour particularités le mélange de texte normal et italique sur une même page avec la présence de réclames.
4. *Sententia dominorum deputatorum, quibus sacra Facultas theologiae Parisiensis curam commisit observandi ea omnia quae spectant approbationes librorum, & cautiones quae in iis concedendis debent adhiberi.* Il s’agit d’un document en latin de l’Université de Paris publié en 1643. Il mélange lui aussi le texte normal et italique en latin avec la présence d’un large tampon sur les pages choisies.
5. *Panegyricus illustrissimo Domino D. Guillelmo de Lamoignon de nuperrima ejusdem in principem senatûs Galliarum promotione. Dictus die undecima januarii an. 1659. apud Maturinenses in majoribus comitiis Parisiensis académiae, ad aedem Deo sacram sub invocatione S. Pauli mox processurae.* C’est un texte en latin de 1659 rédigé par Guillaume Cauvet. Nous observons la présence de notes de marge en italique, de réclames et de signatures.
6. *Theses mathematicæ de hydrostatica, architectura militari, et astronominia. Has theses, Deo duce, & auspice Deipara, propugnabuntur in collegio Claromontano Societatis Jesu, diebus XIX. XX. XXI. mensis junii anni M.DC.LXXVI. a tertia ad vesperam.* Ce document en latin de 1676 a été édité par le Collège de Louis le Grand. Le texte de la page imprimée au verso est visible au recto et nous avons relevé la présence de lettrines au début des paragraphes.
7. *Censure faite par la Faculté de theologie de Paris, d’un livre qui a pour titre : La mystique cité de Dieu , miracle de sa toute-puissance, abîme de la grace, histoire divine, & la vie de la tres-sainte Vierge Marie, mere de Dieu, notre reine & maîtresse ; manifestée dans ces derniers siecles par la sainte Vierge, à la sœur Marie de Jesus, abbesse du couvent de l’Immaculée Conception de la ville d’Agreda, de l’ordre de Saint François ; & écrite par cette même sœur, par ordre de ses supérieurs & de ses confesseurs. Traduite de l’espagnol par le Pere Thomas Croset, recolet.* Ce texte de l’Université de Paris mélange du texte normal et italique en français et en latin sur deux colonnes dans une même page.

XVIII^e siècle

8. *Dessein du theatre dressé au college de Louis le Grand, en l'honneur de Louis XV. fondateur des prix.* Il s'agit d'un ouvrage en français écrit par Joseph de Blainville en 1720. Il comporte de courts passages en italique
9. *Discours sur la convalescence du roy, prononcé à l'ouverture des classes, par M. Crevier, professeur de rhétorique au collège de Beauvais. Le lundi 19. octobre 1744.* Ce document en français de Jean-Baptiste-Louis Crevier daté de 1744. Il ne présente aucune particularité sur l'aspect matériel et typographique.
10. *Extrait des registres du Parlement, du 7 septembre 1762.* Comme son nom l'indique, il s'agit de registres en français du Parlement de Paris datant de 1762. Les pages présentent de légères taches.
11. *Histoire de Paris, et description de ses plus beaux monuments.* Cet ouvrage en français de 1782 par Jean-Charles Poncelin de La Roche-Tilhac ne présente aucune particularité.
12. *Lettre de Lakanal à Louis de Fontanes.* Ce document en français de Joseph Lakanal écrit en 1800 a des bordures intérieures qui sont très peu lisibles avec des guillemets en début et en bout de ligne.

XIX^e et XX^e siècles

13. *Oeuvres complètes de Descartes, publiées par Victor Cousin professeur-suppléant de l'histoire de la philosophie moderne à la faculté des lettres de l'Académie de Paris, maître de conférences à l'ancienne école normale. Prospectus.* Cet ouvrage en français de 1842 écrit par René Descartes présente des taches au niveau des pages.
14. *Messianisme : union finale de la philosophie et de la religion, constituant la philosophie absolue. Tome II. Métapolitique messianique : désordre révolutionnaire du monde civilisé.* Il s'agit d'un livre en français de Józef Maria Hoënne-Wroński et qui date de 1840. La numérisation de ce document a été effectué en niveaux de gris et nous observons ici aussi des taches.
15. *Sainte-Barbe et les barbistes.* C'est un ouvrage en français écrit par un auteur nommé Célestin en 1863. Il présente des taches très visibles au niveau des pages.
16. *Notice sur un ouvrage de médecine orné de miniatures, copié en 1379 : imprimé pour le mariage Avalle-Bassereau, 5 juin 1886.* Cet imprimé en

français date de 1886 et a été écrit par Émile Chatelain. Nous avons relevé la présence de notes de bas de page.

17. *La Valachie : essai de monographie géographique*. Cet essai en français de 1902 par Emmanuel de Martonne présente une courbure du texte sur les bordures intérieures des pages.
18. *Conférence faite le 20 février 1921 sur la Roumanie nouvelle*. Il s’agit d’un ouvrage en français d’Emmanuel de Martonne datant de 1921. Nous observons des pages contenant à la fois du texte et images qui sont légendées.
19. *Les Alpes (géographie générale)*. Cette monographie française de 1941 écrite encore une fois par Emmanuel de Martonne a une qualité d’impression médiocre.
20. *La Bibliothèque de la Sorbonne*. Il s’agit d’un livre en français publié par la Bibliothèque interuniversitaire de la Sorbonne (auteur), Claude Jolly en 1989. Il ne présente aucune particularité.
21. *Lettre dactylographiée de Paul Appell à la marquise Arconati-Visconti, Paris, 13 mars 1922*. Cette lettre en Français par Paul Appell de 1922 est le seul document tapuscrit présent parmi nos échantillons et fait partie du fonds de la marquise Arconati-Vasconti.

Un premier tableau recensant ces 21 échantillons se trouve en annexes. Celui-ci comprend respectivement :

- L’année de publication de l’ouvrage dont est extrait l’échantillon.
- La ou les langues utilisées.
- Les différentes autorités relatives au document.
- Les caractéristiques matérielles et typographiques susceptibles d’influer sur la qualité de l’OCR.
- L’URL qui renvoie au document sur NuBIS. Les liens cliquables sont disponibles dans fichier le nommé « Annexes » du GitHub.
- Les vues (pages) correspondant à l’échantillon.

3. Résultats

Les résultats des tests avec les taux de précision pour chaque logiciel en fonction de l’échantillon (identifiée par l’année de publication de l’ouvrage) sont référencés dans un second tableau disponible également en annexes. Dans la suite du mémoire, il sera référé

aux documents de ce tableau en citant uniquement leur année de publication dans un souci de clarté. Les transcriptions générées par eScriptorium ainsi que les vérités de terrain c'est-à-dire les corrections manuelles ont été directement publiés sous format ALTO et texte dans le répertoire GitHub de HTR-United, un projet visant à regrouper des données d’entraînement pour la reconnaissance de texte au sein d’un même catalogue⁷. Les transcriptions sous format texte pour l’ensemble des logiciels ainsi que les deux tableaux sont disponibles dans le dépôt GitHub du mémoire⁸.

Ces tests se sont ainsi étalées plusieurs semaines et ont constitué le cœur du travail effectué pendant le stage. En effet, il était tout d’abord nécessaire de se familiariser avec chaque logiciel et son fonctionnement. Tesseract était censé être intégré sur la plateforme NumaHOP par exemple mais nous avons préféré utiliser la version directement intégré à Omeka S. ABBYY avait sa propre application sur Windows grâce à la licence dont nous disposons à la BIS avec ABBYY FineReader 16. Quant à Nanonets, eScriptorium et Transkribus, le traitement de l’OCR s’effectuait directement sur leurs sites Web. Nanonets et Transkribus étaient tous les deux en version gratuite avec un nombre de crédits limités mais la qualité de l’OCR n’a pas été impactée par cela. Ensuite, ne pas comptabiliser les « s » longs et les espacements des mots comme des erreurs signifie qu’il était nécessaire de reconstruire à chaque fois une nouvelle vérité de terrain pour un même document qui soit propre à chaque logiciel (plutôt qu’utiliser le même) afin que ces erreurs ne soient pas comptabilisées par ocreval.

Voici donc une représentation graphique qui permet de visualiser les résultats des cinq différents outils en fonction de la date de publication de l’ouvrage :

7. <https://github.com/HTR-United/htr-united/blob/master/catalog/nubis-ocr/nubis-ocr.yml>.

8. <https://github.com/ksefil/Memoire-TNAH>.

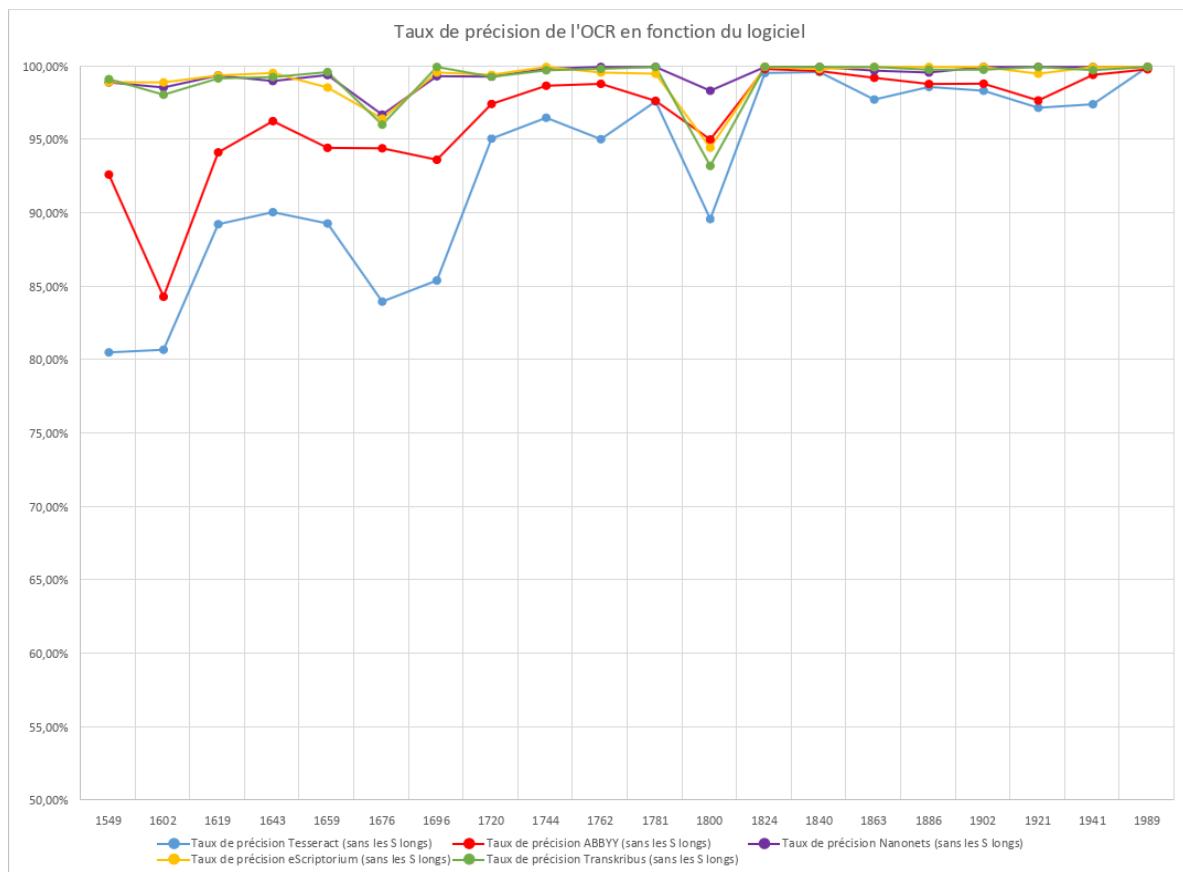


FIGURE 2.1 – Taux de précision de l'OCR en fonction du logiciel et de la date de l'ouvrage

4. Observations

Tesseract et ABBYY

La représentation graphique montre que ces deux outils ont des résultats satisfaisants (supérieurs à 97%) sur les ouvrages postérieurs au XIX^e siècle. Cependant, ils font face aux mêmes difficultés pour les documents qui sont antérieurs à cette date. Nous retrouvons ce même constat dans un article scientifique qui étudie l'OCR pour les imprimés anciens⁹. Il existe en effet plusieurs caractéristiques liées à la qualité de l'ouvrage et à la typographie qui ont un impact négatif sur la performance de Tesseract, et dans une moindre mesure sur celle d'ABBYY. Ces caractéristiques se traduisent par des difficultés suivantes :

- Confusion des lettres « c » et « e », ainsi que « c » et « t ».
- Confusion des accents avec des lettres différentes
- Confusion des ligatures « ct » avec des « é ».

9. Gabay, Simon, Thibault Clérice, et Christian Reul. « OCR17 : Ground Truth and Models for 17th c. French Prints (and hopefully more) ». *Journal of Data Mining and Digital Humanities* 2023 (juin 2023). <https://doi.org/10.46298/jdmdh.6492>.

- Mauvaise lecture des lettrines.
- Notes en marge qui sont très mal lues et directement intégrées dans le corps du texte en fin ou début de ligne.
- Mauvaise lecture due à la courbure du texte sur les bordures intérieures.
- Confusion des textes qui sont écrits sur 2 colonnes différentes.
- Mauvaise lecture des caractères qui se situent sur des tâches visibles sur la page.
- Mauvaise lecture des chiffres et donc des dates.

Ces deux logiciels, en particulier Tesseract, rencontrent parfois des problèmes à reconnaître les espaces entre les mots, ce qui est confirmé dans un autre article scientifique, cette fois consacré à l’HTR¹⁰. Tesseract a aussi beaucoup de difficultés à mettre en page sa transcription et peut parfois même inverser des lignes, tandis qu’ABBYY a l’avantage de pouvoir lire les césures comme un seul mot. Cependant, ABBYY a le défaut de changer quelquefois les terminaisons « -oit » en « -ait ». La différence majeure entre ces deux outils est le fait que Tesseract est incapable de reconnaître les « s » longs contrairement à ABBYY qui les reconnaît dans la majorité des cas, ce qui a un impact considérable sur la qualité de l’OCR pour les imprimés qui sont antérieurs au XIX^e siècle, lorsque l’usage de celui-ci était fréquent.

Nanonets, eScriptorium et Transkribus

Comme le prouvent les taux de précision qui avoisinent les 100%, ces trois outils donnent de très bons résultats quels que soient le type d’ouvrage et les caractéristiques typographiques. Ils présentent toutefois quelques différences :

- Nanonets est incapable de lire les « s » longs, ce qui le rend en pratique beaucoup moins performant que les deux autres outils pour les documents qui datent d’avant le XIX^e siècle. De plus, Nanonets peut parfois transcrire par erreur le texte au verso de la page imprimée quand celle-ci est transparente. L’outil est en revanche excellent sur les tapuscrits. Cela fait sens car comme nous l’avons dit précédemment, Nanonets se spécialise dans les documents comptables, les tapuscrits étant visuellement proches des documents de ce type.
- eScriptorium a des difficultés à lire les documents tapuscrits mais il est capable de très bien transcrire les ouvrages en caractères gothiques.

10. Camps, Jean-Baptiste, Thibault Clérice, et Ariane Pinche. « Noisy medieval data, from digitized manuscript to stylometric analysis : Evaluating Paul Meyer’s hagiographic hypothesis ». *Digital Scholarship in the Humanities* 36, Supplement_2 (1 octobre 2021) : p. 49-71. <https://doi.org/10.1093/llc/fqab033>.

- Transkribus est plutôt moyen pour la lecture des caractères gothiques mais c'est le seul outil parmi ceux testés qui est capable de développer les abréviations à certains moments.

Globalement, les trois logiciels ne parviennent pas à identifier correctement les lettrines, à l'instar des deux outils précédents. Ils ont eux aussi plus ou moins des difficultés en présence de courbures du texte sur l'intérieur des pages comme cela est le cas avec l'ouvrage testé datant de 1800, expliquant ainsi la baisse de la courbe à cette date dans le graphique précédent. En outre, ils ont tendance à transcrire les notes manuscrites dans les ouvrages, ce qui peut impacter de manière négative la lisibilité de la transcription. Malgré ces quelques défauts, ces trois logiciels obtiennent des taux de précision excellents comme dit précédemment. Ce résultat est d'autant plus étonnant lorsqu'on peut lire dans le manuel de numérisation de l'Enssib datant de 2010 : « Même en combinant plusieurs logiciels d’OCR sur un même texte, atteindre une haute qualité (entre 98 et 100 %) implique une reprise manuelle et devient donc plus coûteux que l’OCR brut qui peut fournir de bons résultats si le support et l'impression sont de qualité ».¹¹ Cela prouve que ces dix dernières années ont vu le développement de logiciels d’OCR sensiblement plus performants que leurs prédecesseurs.

Observations générales

De manière générale, les ouvrages ayant des pages grises, la présence d’images dans le corps du texte ainsi que les notes de bas de page ne posent pas de difficulté particulière aux différents outils d’OCR. Il faut néanmoins faire attention à bien fournir à l’OCR une image avec une définition suffisamment élevée pour obtenir la meilleure transcription possible en sortie. D’après les résultats obtenus, Nanonets, eScriptorium et Transkribus sont donc les meilleurs outils en termes de performance, suivis par ABBYY et enfin par Tesseract. À l’exception du document datant de 1800, les trois premiers outils ont des taux de précision très proches pour l’ensemble des documents. Avant le XIX^e siècle, ils sont nettement meilleurs que les deux derniers outils mais après cette date, la différence de qualité entre les différents OCR devient marginale.

Il est toutefois nécessaire de prendre en considération le fait que Nanonets et Tesseract sont tous les deux incapables de lire les « s » longs, ce qui les handicape grandement pour les ouvrages antérieurs à cette période. En outre, le taux de précision ne peut pas être le seul critère sur lequel se baser pour déterminer le choix de l’outil d’OCR. Il est

11. Claerr, Thierry, et Isabelle Westeel, éd. *Numériser et mettre en ligne. Numériser et mettre en ligne*. La Boîte à outils. Villeurbanne : Presses de l'Enssib, 2010. p. 33. <https://books.openedition.org/pressesenssib/414>.

aussi nécessaire de s’interroger sur les formats d’exportation proposés par les logiciels ainsi que sur leur disponibilité. Le tableau ci-dessous renseigne ces informations et récapitule les performances pour chaque outil.

	Tesseract	ABBYY	Nanonets	eScriptorium	Transkribus
Qualité de l’OCR sans les « s » longs	Moyenne	Bonne	Excellente	Excellente	Excellente
Lecture des « s » longs	Non	Partielle	Non	Oui	Oui
Formats d’exportation	Texte, PDF	Texte, PDF	Texte, PDF	Texte, ALTO	Texte, PDF, ALTO
Disponibilité du logiciel	Gratuite	Payante	Payante	Gratuite	Payante

TABLE 2.2 – Bilan des caractéristiques et performances des logiciels testés

III. Choix du logiciel

Les résultats des tests nous ont paru pour le moins surprenants. En effet, malgré un quasi-monopole du logiciel ABBYY au sein des bibliothèques que nous avons interrogées, les résultats prouvent que ce logiciel est dans l’ensemble peu adapté pour des bibliothèques ayant une collection ancienne. Dans le cas d’une bibliothèque numérique comme NuBIS dont la majorité des imprimés est antérieure au XIX^e siècle, il nous a ainsi paru peu pertinent de faire appel à ABBYY car celui-ci est incapable de lire les « s » longs. Il en est de même pour Tesseract et Nanonets. Étonnamment, ce sont donc les logiciels spécialisés dans l’HTR, à l’instar d’eScriptorium et Transkribus, qui donnent les meilleurs résultats pour les bibliothèques numériques de ce type en termes de performance brute. Il s’agissait donc pour la BIS d’effectuer un choix parmi ces deux logiciels. C’est ici qu’il faut prendre en considérations les autres paramètres définies par le consortium IMPACT que nous avons mentionnés dans l’introduction.

De fait, le tableau précédent montre deux différences notables pour eScriptorium et Transkribus : le premier est *open source* et gratuit, tandis que le second a l’avantage de proposer le format PDF avec du texte recherchable en sortie. L’argument budgétaire penche donc en faveur d’eScriptorium : il est donc question de déterminer si le format PDF était utile dans le projet d’OCR souhaité par la BIS. Après réflexion, nous avons déterminé qu’il faudrait trois fonctionnalités principaux pour un utilisateur au regard de

l'OCR :

- Avoir la recherche en plein texte sur le site de la NuBIS.
- Avoir la recherche en plein texte au sein même du document.
- Afficher la transcription sous format texte à l'utilisateur.

En ce qui concerne la dernière fonctionnalité tout d'abord, si la transcription est en format ALTO, il est possible d'intégrer directement la transcription obtenue par l'OCR au sein d'un visualiseur IIIF tel que Mirador par l'intermédiaire du module Extract OCR¹² disponible sur Omeka S. En effet, ce format XML effectue une segmentation de la page en différents sous-éléments, permettant ainsi de localiser chaque mot au sein de la page grâce à des coordonnées. Il est très utilisé pour la conversion en mode texte de documents patrimoniaux¹³. Les tests effectués pendant le stage nous ont paru convaincants avec le texte transcrit qui apparaissait sur le côté droit du visualiseur IIIF. Le texte est par ailleurs interactif puisque cliquer sur une ligne de la transcription permet de zoomer sur la page avec la ligne sélectionnée qui apparaît de manière surlignée (et vice versa), ce qui nous a paru être un atout du point de vue de l'utilisateur. Il faut souligner qu'il nous a paru impossible dans un premier temps d'utiliser les fichiers ALTO générées par eScriptorium mais en analysant les codes sources de ces fichiers, nous nous sommes rendus compte que cela était dû à un problème de pagination des fichiers (celle-ci commençant à 0 au lieu de 1). Ce problème a donc pu être résolu sans grande difficulté.

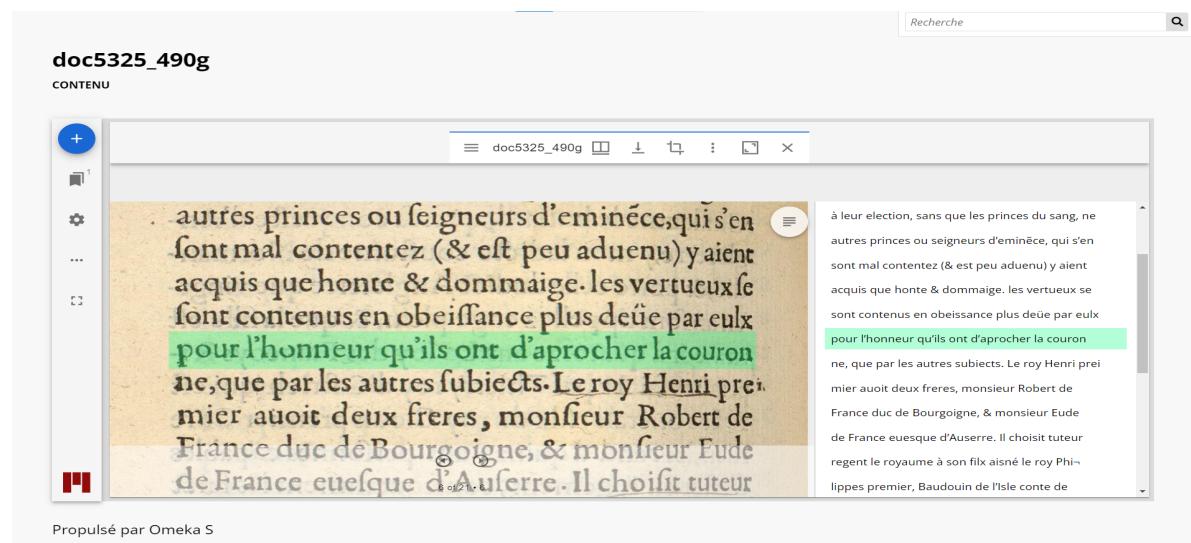


FIGURE 2.2 – Capture d'écran de l'interface de Mirador avec la transcription

12. <https://omeka.org/s/modules/ExtractOcr/>.

13. « Référentiel OCR version 2 ». Bibliothèque nationale de France, 2015. p. 8. https://www.bnf.fr/sites/default/files/2018-11/ref_num_ocr_v2.pdf.

Quant à la recherche au sein du document, nous avons aussi pu la mettre en place grâce au module IIIF Search¹⁴ sur Omeka S qui permet d’avoir la recherche en plein texte directement dans le visualiseur IIIF. Enfin, concernant la recherche sur NuBIS en général, il est tout à fait possible de l’avoir en versant le fichier .txt de la transcription sur la page Omeka S du document correspondant. Ainsi, la présence d’un fichier PDF avec du texte recherchable nous a paru peu pertinent dans notre cas puisque l’interface IIIF nous a semblé bien plus intuitive à utiliser en sachant que le texte contenu dans un fichier PDF ne peut être recherché sur NuBIS. Par conséquent, en raison de son aspect *open source* et gratuit, la BIS a fait le choix de s’orienter vers eScriptorium comme solution d’OCR pour sa bibliothèque numérique.

Après avoir contacté l’INRIA, responsables de l’application, celui-ci nous a proposé de réaliser l’océrisation des imprimés gratuitement pour la Sorbonne et de nous fournir le script pour effectuer ce processus d’océrisation en interne par la suite. En échange, la BIS leur accorde le libre accès aux données produites et s’engage à verser les données d’entraînement sur HTR-United. Cela a ainsi épargné la Sorbonne d’une contrainte supplémentaire qui aurait été de dédier du personnel supplémentaire et donc un budget spécial pour effectuer le processus de l’océrisation. Un disque dur et un fichier .csv comprenant les données d’identification (date de publication, auteur, titre...) de chaque document avec l’emplacement au sein du disque de leurs images ont ainsi été fournies à l’INRIA au vu de cet accord. La fin de cette opération d’océrisation est prévue pour le mois d’octobre 2024.

IV. Correction de l’OCR avec l’intelligence artificielle

Quel que soit le taux de précision obtenu pour l’OCR, celui-ci ne peut être toujours égal à 100% comme nous avons pu le voir. Cela nous a conduit à réfléchir à des moyens d’améliorer la transcription obtenue par le logiciel d’OCR. Une correction manuelle et humaine peut être envisageable. Celle-ci a l’avantage d’être fiable et assez précise. Toutefois, un tel processus serait peu réaliste pour la BIS compte tenu du volume de documents océrisés. Pour cette raison, nous nous sommes intéressés à la piste de l’intelligence artificielle. En effet, les outils d’intelligence artificielle générative ont connu une explosion ces dernières années. Naturellement, nous nous sommes demandés si ces outils étaient capables de corriger et d’améliorer les transcriptions déjà produites par un logiciel d’OCR. Pour cela, nous avons interrogé ChatGPT dans sa version GPT-4o¹⁵ afin de voir si celui-ci pouvait nous être utile.

14. <https://github.com/smachefer/Omeka-S-module-IIIFSearch>.

15. <https://chatgpt.com/>.

Les grands modèles de langage (aussi connus avec l’abréviation anglais LLM) dont fait partie ChatGPT sont en théorie bien adaptés à la tâche de correction de l’OCR car ils sont entraînés à prédire le mot le plus probable. Cependant, ces modèles sont entraînés en grande partie sur du contenu web natif plutôt que sur des documents historiques numérisés. Les progrès récents des LLM offrent un horizon prometteur pour l’amélioration de la qualité des résultats de l’OCR. En particulier, les modèles ChatGPT d’OpenAI ont montré des performances impressionnantes sur des problèmes qui se rapprochent de la tâche de correction post-OCR comme la traduction automatique ou encore la correction des fautes d’orthographe et de grammaire dans un texte.

Les recherches déjà menées à ce sujet dans le milieu académique montrent que les modèles de LLM ont des difficultés avec les textes OCR qui sont fortement erronés ou à l’inverse avec qui sont déjà excellents. De plus, fournir à l’outil d’IA des informations contextuelles (par exemple le nom de l’auteur ou l’année de publication) n’apporterait pas d’améliorations significatives au résultat obtenu. Néanmoins, les résultats obtenus seraient tout de même satisfaisants de manière générale avec un taux d’erreur par caractère qui serait 18.92% moindre avec ChatGPT.¹⁶ Il s’agit donc de vérifier ces observations en utilisant ChatGPT sur quelques échantillons de notre propre corpus et de juger si l’utilisation de cet outil serait pertinente dans le cadre de la BIS.

Nous avons commencé par lui fournir l’OCR faite par eScriptorium sur trois pages de l’imprimé datant de 1800. Il est donc écrit dans un français proche du français actuel mais les caractéristiques de l’ouvrage font que l’OCR présente quelques erreurs, le taux de précision au mot étant de 94,50%. Voici l’instruction (ou « prompt » en anglais) fourni à ChatGPT :

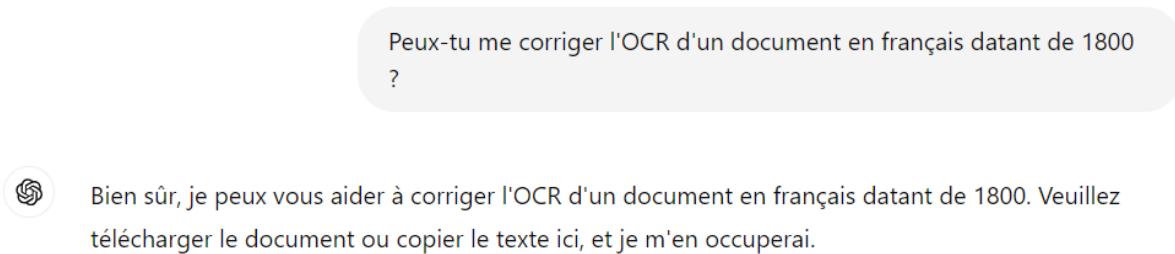


FIGURE 2.3 – Capture d’écran de la première instruction fournie à ChatGPT

16. Zhang, James, Wouter Haverals, Mary Naydan, et Brian W. Kernighan. « Post-OCR Correction with OpenAI’s GPT Models on Challenging English Prosody Texts ». In *Proceedings of the ACM Symposium on Document Engineering 2024*, 1-4. DocEng ’24. New York, NY, USA : Association for Computing Machinery, 2024. <https://doi.org/10.1145/3685650.3685669>.

Le texte corrigé par ChatGPT obtient un taux de précision au mot de 99,42%, ce qui est donc considérablement meilleur. Cette correction a duré environ une minute pour l'ensemble des trois pages.

Ensuite, nous lui avons donné la même instruction pour un document plus ancien avec l'échantillon en français datant de 1602 et ayant un taux de précision de 98,95% sur eScriptorium. ChatGPT a corrigé le texte dans un français actuel, ce que nous ne voulons surtout pas. Il a donc fallu lui spécifier le prompt ci-dessous pour qu'il effectue la bonne opération :

Peux-tu corriger les fautes d'OCR sur ce document datant de 1602 tout en préservant le français de cette époque ?



Oui, je peux vous aider avec cela. Veuillez télécharger le document contenant le texte OCR afin que je puisse le corriger tout en respectant le français du début du XVIIe siècle.

FIGURE 2.4 – Capture d'écran de la deuxième instruction fournie à ChatGPT

Un taux de précision légèrement meilleure de 99,24 % est calculé avec cette instruction. Il faut néanmoins souligner que malgré le respect de l'orthographe de l'époque, ChatGPT a tout de même changé les lettres « i » en « j » et les lettres « u » en « v » lorsqu'elles se prononcent respectivement avec le son [j] et [v], ces lettres se confondant entre elles à cette époque. Ainsi, le mot « ie » se transforme en « je » par exemple. La pertinence d'un tel choix est sujette à discussion car nous obtenons une meilleure lisibilité des mots, mais cela se fait au détriment d'une authenticité moindre de la graphie du texte transcrit.

Que se passe-t-il cependant si on lui demande de corriger une transcription qui ne présente aucune erreur ? Nous avons effectué le test sur la transcription produite par eScriptorium sur l'ouvrage datant de 1941, celui-ci ayant déjà un taux de précision au mot de 100 % :

Peux-tu me corriger les fautes d'OCR d'un document en français datant de 1941 ?



Bien sûr, je peux vous aider avec cela. Veuillez télécharger le document contenant le texte OCR en français de 1941, afin que je puisse le corriger.

FIGURE 2.5 – Capture d'écran de la troisième instruction fournie à ChatGPT

Le texte en sortie était identique avec un taux de 100% aussi. Cela prouve que ChatGPT ne cherche pas à modifier la transcription à tout prix si celle-ci est déjà exacte initialement.

Enfin, nous avons demandé à ChatGPT s'il peut traiter les transcriptions à partir des fichiers ALTO et nous fournir un fichier ALTO corrigé en sortie, ce qui serait très utile pour notre cas :

Peux-tu me corriger l'OCR de ce fichier ALTO et me fournir un fichier ALTO corrigé ?



Malheureusement, je ne peux pas directement traiter ou corriger des fichiers spécifiques comme des fichiers ALTO. Mon domaine d'expertise se situe plutôt dans la génération de texte et la réponse à des questions. Cependant, je peux vous aider à comprendre comment corriger un fichier ALTO ou discuter des étapes générales que vous pourriez suivre pour le faire vous-même.

FIGURE 2.6 – Capture d'écran de l'instruction sur le fichier ALTO rejetée par ChatGPT

L'échange ci-dessus montre que ChatGPT, en tant qu'agent conversationnel (*chatbot*), est incapable de nous exporter directement un fichier ALTO. Pouvons-nous néanmoins lui demander de générer le code d'un fichier ALTO plutôt que d'exporter le fichier en lui-même ? Nous avons essayé cela avec le prompt suivant :

Bien sûr, je vais examiner et corriger le texte du fichier ALTO que vous avez fourni. Voici le texte extrait et corrigé:

```
<?xml version="1.0" encoding="UTF-8"?>
<alto xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://www.loc.gov/standards/alto/ns-v4#"
      xsi:schemaLocation="http://www.loc.gov/standards/alto/ns-v4# http://www.loc.gov/standards/alto/ns-v4.xsd">
    <Description>
      <MeasurementUnit>pixel</MeasurementUnit>
      <sourceImageInformation>
        <fileName>BIS_00_00214_P027_0018.jpg</fileName>
      </sourceImageInformation>
    </Description>
  <Tags>
```

FIGURE 2.7 – Capture d’écran de l’instruction permettant à ChatGPT de corriger un fichier ALTO

Cette fois-ci, ChatGPT accepte notre requête et il corrige effectivement le texte de l’OCR contenu dans les balises. L’opération est cependant très lente : la génération du code prend beaucoup de temps, et il est nécessaire de lui indiquer toutes les quelques secondes par un clic de continuer à générer le code. Or, pour un ouvrage océrisé, il existe autant de fichiers ALTO que d’images autrement dit de pages. Ainsi, il nous paraît peu pratique d’utiliser ChatGPT pour corriger individuellement les fichiers ALTO de la plupart des documents car cela prendrait un temps excessif.

En conclusion, les tests réalisés sur ces quelques échantillons montrent que ChatGPT peut améliorer la précision de l’OCR et ce, quelle que soit l’ancienneté de l’ouvrage. Cependant, la lenteur de cette opération couplée à l’incapacité d’exporter directement un fichier ALTO rendent cet outil peu adapté pour une institution comme la BIS qui souhaite traiter un large volume de documents pour son projet d’OCR.

Conclusion

Notre mémoire consistait à étudier l'implémentation de l'OCR dans une bibliothèque numérique comme celle de la Sorbonne à travers un stage effectué dans le même établissement. Il fallait donc repérer les défis qu'un tel processus soulevait et ensuite proposer des solutions à ces difficultés. Comme nous l'avons vu tout au long de notre étude, il n'existe pas de solution unique qui satisfasse chaque institution souhaitant mettre en place l'OCR pour ses documents. En effet, chaque institution a ses propres contraintes et objectifs au regard de l'utilisation qu'elle souhaite faire de l'OCR.

Dans le cas de la BIS, il était ainsi nécessaire de se pencher sur l'histoire de la bibliothèque afin de comprendre les enjeux liés à sa collection aujourd'hui. Nous avons donc retracé la fondation de sa bibliothèque NuBIS puis nous nous sommes intéressés à leur politique de numérisation. L'enjeu étant la numérisation des manuscrits et imprimés, une typologie chronologique de ces derniers a été mise en place afin de préparer en amont les tests d'OCR à venir et avoir une estimation du taux de précision selon leurs caractéristiques matérielles et typographiques.

La seconde étape consistait alors à essayer différents logiciels d'OCR qui existent actuellement sur le marché. Un état de l'art des autres bibliothèques numériques patrimoniales ayant une collection similaire à NuBIS accompagné par des recherches personnelles ont permis d'isoler cinq logiciels différents à tester. Après avoir choisi un corpus représentatif des imprimés numérisés, les tests ont eu lieu et ils ont montré que Transkribus et eScriptorium conviennent le mieux à une collection ancienne comme celle de la BIS. De la même manière, des réflexions sur ce que serait le plus adapté aux attentes précises de la Sorbonne ont permis de trancher la question et de s'orienter vers eScriptorium. Enfin, une piste de la correction des transcriptions par ChatGPT a aussi été explorée mais malgré des résultats prometteurs, elle ne pouvait s'appliquer à volume large comme cela est le cas pour la BIS.

Désormais, nous pouvons reprendre les différents critères du consortium IMPACT que nous avons énumérés dans l'introduction et qui ont servi de fil directeur pour notre mémoire. Ce sont toutes les interrogations autours de ces paramètres qui nous ont permis

d’aboutir à ce choix précis de logiciel OCR avec eScriptorium. Les recherches et travaux menés jusqu’ici nous ont en effet permis de répondre à chacun de ces critères selon les objectifs, moyens et contraintes de la BIS :

- **Objectifs du projet** : Nous avons soulevés trois objectifs principaux pour le projet d’OCR à la BIS. Il s’agit de la recherche en plein texte sur l’ensemble de la bibliothèque numérique, la recherche en plein texte à l’intérieur d’un document, et enfin afficher un rendu textuel de la transcription à l’utilisateur.
- **Caractéristiques du matériel source et de la numérisation** : La grande majorité des documents à numériser sont des imprimés antérieurs au XIX^e siècle. Ils sont donc caractérisés par la présence de « s » longs ainsi que des défauts liés à la typographie ancienne, ce qui nécessite un logiciel qui soit performant sur les textes et documents anciens.
- **Contrôle de qualité** : Un programme de contrôle de qualité de l’OCR a bien été mis en place en évaluant l’efficacité de cinq logiciels sur un échantillon de 21 documents qui sont représentatifs de la collection disponible sur NuBIS. Ce programme s’est étalé sur plusieurs semaines et a constitué la majorité du travail réalisé pendant le stage.
- **L’échelle du projet** : Le volumes de documents à océriser est largement inférieur à Gallica, mais il s’agit tout de même d’un volume conséquent avec 124 459 pages à traiter. Cela rend toute correction manuelle ou par l’IA de l’OCR peu réaliste dans notre cas.
- **Internalisation ou externalisation de l’OCR** : Après avoir obtenu des résultats très satisfaisants avec les tests effectués sur notre corpus, il a été décidé de ne pas faire appel à un prestataire et de réaliser l’OCR en interne. La proposition de l’INRIA nous a toutefois convaincu : l’océrisation est donc faite partiellement en externe dans un premier temps mais par la suite la BIS disposera du script utilisé par l’INRIA et pourra donc océriser ses nouveaux imprimés complètement en interne.
- **Durée du projet** : La mise en place de l’OCR aurait finalement duré six mois avec une période allant d’avril à octobre 2024. Celle-ci peut se découper en deux parties avec les quatre mois de stage dans un premier temps et les deux mois de traitement de l’OCR effectué par l’INRIA.
- **Coûts** : Le coût du projet s’est avéré quasi nul pour la BIS au final avec comme seule dépense le paiement du stagiaire. Aucun logiciel payant n’a été acquis dans le cadre du stage et l’accord d’océrisation avec l’INRIA n’a impliqué aucun aspect financier.

C'est donc sur ce récapitulatif que s'achève ce mémoire ainsi que plusieurs mois de réflexions sur la mise en place d'une OCR conforme à la BIS. Ce travail constitue uniquement le premier jalon d'un long processus : à terme, il s'agirait d'avoir de la reconnaissance de texte sur l'ensemble des collections sur NuBIS, y compris les manuscrits. Cela sortirait alors du cadre de l'OCR pour entrer dans celui de l'HTR, ouvrant ainsi de nouvelles perspectives de réflexion avec leurs propres défis.

Annexes

Année	Langue	Autorités	Caractéristiques matérielles et typographiques	URL	Vues (pages)
1549	LAT	Pierre de La Ramée (auteur), Charles de Lorraine (dédicataire), Louis Grandin (imprimeur)	Texte en italique du 16e siècle avec visibilité du texte de la page imprimée au dos	https://nubis.univ-paris1.fr/ark:/15733/3sxg	10, 11, 12
1602	FR	Alphonse de Rambervillers (auteur), Melchior Bernard (imprimeur)	Texte en italique et courbure sur les bordures intérieures	https://nubis.univ-paris1.fr/ark:/15733/49bk	6, 7, 8
1619	FR	Juste Lipse (auteur), Antoine Brun (traducteur), Barthélemy Ancelin (imprimeur)	Mélange texte normal et italique et présence de réclames	https://nubis.univ-paris1.fr/ark:/15733/1cz0	44, 45, 46
1643	LAT	Université de Paris (auteur)	Mélange texte normal et italique en latin avec présence d'un gros tampon	https://nubis.univ-paris1.fr/ark:/15733/1f71	12, 13, 14
1659	LAT	Guillaume Cauvet (auteur), Jean Julien (imprimeur)	Présence de notes de marge en italique, de réclames et de signatures	https://nubis.univ-paris1.fr/ark:/15733/1khw	15, 16, 17
1676	LAT	Collège de Louis le Grand (éditeur)	Texte avec visibilité de la page imprimée au dos et présence de lettrines	https://nubis.univ-paris1.fr/ark:/15733/33m5	6, 7, 8
1696	FR + LAT	Université de Paris (auteur), Louis Josse (imprimeur)	Mélange de texte normal et italique en français et en latin sur deux colonnes	https://nubis.univ-paris1.fr/ark:/15733/17zw	9, 10, 11
1720	FR	Joseph de Blainville (auteur)	Courts passages en italique	https://nubis.univ-paris1.fr/ark:/15733/wz1	5, 6, 7
1744	FR	Jean-Baptiste-Louis Crevier (auteur), Jean Desaint (imprimeur), Charles Saillant (imprimeur)	Aucune particularité	https://nubis.univ-paris1.fr/ark:/15733/1181	5, 6, 7
1762	FR	Parlement de Paris, Pierre-Guillaume Simon (imprimeur)	Légères taches	https://nubis.univ-paris1.fr/ark:/15733/1wtw	3, 4, 5

Année	Langue	Autorités	Caractéristiques matérielles et typographiques	URL	Vues (pages)
1781	FR	Jean-Charles Poncelin de La Roche-Tilhac (auteur), quatre imprimeurs différents	Aucune particularité	https://nubis.univ-paris1.fr/ark:/15733/47w0	17, 18, 19
1800	FR	Joseph Lakanal (auteur)	Bordures intérieures peu lisibles et guillemets en début et bout de ligne	https://nubis.univ-paris1.fr/ark:/15733/212d	19, 20, 21
1824	FR	René Descartes (auteur)	Présence de taches	https://nubis.univ-paris1.fr/ark:/15733/343s	3, 4, 5
1840	FR	Józef Maria Hoënè-Wroński (auteur)	Numérisation en niveaux de gris et présence de taches	https://nubis.univ-paris1.fr/ark:/15733/1msc	7, 8, 9
1863	FR	M. Célestin (auteur)	Présence de taches très visibles	https://nubis.univ-paris1.fr/ark:/15733/1dkv	7, 8, 9
1886	FR	Émile Chatelain (auteur)	Présence de notes de bas de page	https://nubis.univ-paris1.fr/ark:/15733/17b9	10, 11, 12
1902	FR	Emmanuel de Martonne (auteur)	Courbure du texte sur les bordures intérieures	https://nubis.univ-paris1.fr/ark:/15733/m38p	18, 19, 20
1921	FR	Emmanuel de Martonne (auteur)	Pages mixtes avec texte et images	https://nubis.univ-paris1.fr/ark:/15733/m35r	10, 11, 12
1941	FR	Emmanuel de Martonne (auteur)	Qualité d'impression médiocre	https://nubis.univ-paris1.fr/ark:/15733/m3j5	16, 17, 18
1989	FR	Bibliothèque interuniversitaire de la Sorbonne (auteur), Claude Jolly (éditeur)	Aucune particularité	https://nubis.univ-paris1.fr/ark:/15733/3sgf	12, 13, 14
1922	FR	Paul Appell (auteur)	Exemple de tapuscrit	https://nubis.univ-paris1.fr/ark:/15733/k2rh	1

TABLE 3 – Caractéristiques du corpus établi

Année	Taux de précision Tesseract	Taux de précision ABBYY	Taux de précision Nanonets	Taux de précision eScriptorium	Taux de précision Transkribus
1549	80.56 %	92.67 %	98.95 %	98.96 %	99.17 %
1602	80.74 %	84.35 %	98.60 %	98.95 %	98.11 %
1619	89.29 %	94.18 %	99.42 %	99.42 %	99.22 %
1643	90.11 %	96.31 %	99.04 %	99.59 %	99.31 %
1659	89.34 %	94.49 %	99.45 %	98.60 %	99.65 %
1676	84.02 %	94.46 %	96.75 %	96.48 %	96.07 %
1696	85.46 %	93.68 %	99.37 %	99.62 %	100.00 %
1720	95.12 %	97.48 %	99.34 %	99.47 %	99.33 %
1744	96.54 %	98.72 %	99.88 %	100.00 %	99.77 %
1762	95.08 %	98.85 %	100.00 %	99.62 %	99.87 %

Année	Taux de précision Tesseract	Taux de précision ABBYY	Taux de précision Nanonets	Taux de précision eScriptorium	Taux de précision Transkribus
1781	97.65 %	97.69 %	100.00 %	99.54 %	100.00 %
1800	89.64 %	95.06 %	98.38 %	94.50 %	93.26 %
1824	99.59 %	99.86 %	100.00 %	100.00 %	100.00 %
1840	99.65 %	99.72 %	100.00 %	99.86 %	100.00 %
1863	97.78 %	99.26 %	99.75 %	100.00 %	100.00 %
1886	98.64 %	98.84 %	99.62 %	100.00 %	99.81 %
1902	98.38 %	98.86 %	100.00 %	100.00 %	99.81 %
1921	97.22 %	97.72 %	100.00 %	99.54 %	100.00 %
1941	97.46 %	99.46 %	100.00 %	100.00 %	99.78 %
1989	100.00 %	99.84 %	100.00 %	100.00 %	100.00 %
1922	94 %	93.92 %	100.00 %	78.43 %	96.86 %

TABLE 4 – Taux de précision au mot des logiciels testés

Liste des tableaux

1.1	Profil des lecteurs fréquentant la BIS	4
2.1	Liste des bibliothèques numériques avec OCR examinées dans le cadre de cette étude	24
2.2	Bilan des caractéristiques et performances des logiciels testés	35
3	Caractéristiques du corpus établi	49
4	Taux de précision au mot des logiciels testés	51

Table des figures

1.1	Capture d'écran de l'interface de NuBIS avec Mirador	9
1.2	Capture d'écran de l'interface d'Omeka S	9
1.3	Répartition des documents par type sur NuBIS	12
1.4	Exemple de tapuscrit avec le texte au verso lisible (https://nubis.bis-sorbonne.fr/ark:/15733/9qp6?mirador-1=1)	13
1.5	Répartition du nombre d'imprimés par décennie	15
1.6	Répartition du nombre d'imprimés par décennie	15
1.7	Extrait d'un ouvrage de la fin du XV ^e siècle. (https://nubis.bis-sorbonne.fr/ark:/15733/ng7n?mirador-1=4)	16
1.8	Extrait d'un ouvrage du début XVII ^e siècle. (https://nubis.bis-sorbonne.fr/ark:/15733/49bk?mirador-1=17)	18
1.9	Extrait d'un ouvrage du milieu du XVII ^e siècle. (https://nubis.bis-sorbonne.fr/ark:/15733/1khn?mirador-1=43)	18
1.10	Extrait d'un ouvrage de la fin du XVII ^e siècle. (https://nubis.bis-sorbonne.fr/ark:/15733/17zw?mirador-1=7)	19
1.11	Extrait d'un ouvrage typique du XVIII ^e siècle. (https://nubis.bis-sorbonne.fr/ark:/15733/1181?mirador-1=11)	20
1.12	Extrait d'un ouvrage de la fin du XIX ^e siècle représentatif des imprimés de cette période avec aucune difficulté éventuelle pour l'OCR. (https://nubis.bis-sorbonne.fr/ark:/15733/17b9?mirador-1=12)	21
2.1	Taux de précision de l'OCR en fonction du logiciel et de la date de l'ouvrage	32
2.2	Capture d'écran de l'interface de Mirador avec la transcription	36
2.3	Capture d'écran de la première instruction fournie à ChatGPT	38
2.4	Capture d'écran de la deuxième instruction fournie à ChatGPT	39
2.5	Capture d'écran de la troisième instruction fournie à ChatGPT	40
2.6	Capture d'écran de l'instruction sur le fichier ALTO rejetée par ChatGPT	40
2.7	Capture d'écran de l'instruction permettant à ChatGPT de corriger un fichier ALTO	41

Table des matières

Résumé	i
Remerciements	iii
Bibliographie	v
Introduction	ix
1 La BIS et sa collection	1
I. Histoire de la bibliothèque	1
II. Organisation de la bibliothèque	5
III. La bibliothèque numérique	6
IV. Typologie des documents à océriser	12
1. 1467 - 1539	15
2. 1539 - fin du XVII ^e siècle	17
3. XVIII ^e siècle	19
4. XIX ^e et XX ^e siècles	20
2 Une solution d'OCR adaptée à la BIS	23
I. État des lieux	23
II. Réalisation des tests	25
1. Méthodologie	25
2. Corpus	27
3. Résultats	30
4. Observations	32
III. Choix du logiciel	35
IV. Correction de l'OCR avec l'intelligence artificielle	37
Conclusion	43
Annexes	48