ABBYY FineReader and Transkribus as philological tools: digitizing multilingual and dialphabetic ancient medical dictionaries (16th–18th centuries)

Anaïs CHAMBAT, Cahal TAAFFE

ABBYY FineReader and Transkribus as philological tools: digitizing multilingual and *dialphabetic* ancient medical dictionaries (16th– 18th centuries)

Abstract

We propose in this article a critical comparison of two approaches in handling medical dictionaries digitalization, a specialized and an industrial one. Both methods were applied on similar corpora but required different approaches. We examine the advantages and disadvantages of using ABBYY FineReader and Transkribus as well as the methods by which researchers and engineer can better apply them, to older printed work.

Keywords

ABBYY Finereader ; Transkribus ; OCR ; ancient medical dictionaries

Introduction

Optical character recognition or OCR is a technology that has always had great potential within the humanities[1]. By digitizing thousands of pages researchers were able to do a systematic analysis previously impossible by a single human reading[2]. Most OCR are now based on a neural network which allows the recognition of characters based on values attributed to each pixel determined by a model as well as machine learning[3]. This improvement had led to a huge increase in terms of performance and customization allowing training of model on specific corpus. Yet no technology is perfect and trouble has arisen when trying to digitize older printed work or even hand-written manuscripts. We propose in this article a critical comparison of two digitizing chains used on similar corpora, examine the advantages and disadvantages of the tools as well as the methods by which researchers and engineer can better apply them, to older[4] printed work.

This research is based on the "Multilingual Medical Metadictionary of the Medica Digital Library" (METADICTMEDICA) CollEx-Persée project[5] which aims to establish links between words currently dissociated by spelling, language, or by the evolution of usage in a corpus of some 50 medical dictionaries and encyclopedias covering four centuries (17th-20th). We studied the 1622 edition of the *Definitionum medicarum libri XXIIII*[6] by Jean de Gorris and its abridged French version[7] by François Thévenin

[1] To know more about the beginnings and history of OCR, see [Schantz H. F. 1982].
[2] See for example [Reboul M. 2022].
[3] See [Gefen A., Saint-Raymond L., Venturini T. 2020] 4-5.
[4] Most commercial OCR are designed to work on contemporary documents where printing has moved beyond lead ink and pressure printing to move to offset method that have a higher durability.
[5] See CollEx-Persée Project website page.
[6] Gorris J. *Opera Definitionum medicarum libri XXIIII. A Joanne Gorraeo filio … locupletati & accessione magna adaucti*, Paris, Apud Societatem Minimam, 1622, 742 p. Medica
[7] Thevenin F. *Œuvres contenant un traité des opérations de chirurgie, un traité des tumeurs, & un dictionnaire étymologique de mots grecs servant à la médecine, recueillies par maistre Guillaume Parthon*, Paris, Rocolet, 1658, 202 p. Medica

(1658) as well as the 1746 edition of the *Lexicon medicum graeco-latinum*[8] (1644) by Bartolomeo Castelli. We have also included the three volumes of Robert James' *Medicinal Dictionary*[9] (1743–1745) and its translation in six volumes[10] by Diderot, Eidous and Toussaint (1746–1748), respectively written in English and French, but whose headwords are in Latin. The specificity of dictionaries lies in the diversity of relations between words that can be established and thus marked. The relations between words are indicated by the typography or a verbal element that makes the type of relation explicit or not. They can be spelling variants, synonyms (within the same language or between terms of different languages) and cross-references. Finally, the same relation can be characterized in several different ways: according to the dictionaries and within the same dictionary. The structure and typography of these dictionaries posed problems to the OCR programs that had to be solved.

The main issued we face was the need to work on dictionaries that were the result of Renaissance and early modern printing practices. They were each multilingual, two to three different languages, and *dialphabetical*, containing two alphabets in a continuous fashion meaning it was impossible to isolate each alphabet. In addition to those diachronic and linguistic differences, each corpus varied in size from medium (fewer than a thousand pages) to large (more than a thousand pages, across multiple volumes). Moreover, we had to face many technical issues such as irregular spacing between characters, accent lines, ligatured Greek and typographic variance. The need for both philological and technical analysis of each corpus was paramount as was the need for different approaches in handling their digitization such as a specialized and an industrial tool: ABBYY FineReader and Transkribus.

ABBYY FineReader[11] is a proprietary software made by the ABBYY corporation and sold as PDF editing tools and an OCR software mostly to businesses wanting to digitize their archives. Yet it's prebuilt template model and machine learning abilities allowed it to function when dealing with early modern printed documents but the software requires extensive training and correction from the user to achieve a sufficiently high level of recognition. Transkribus[12] is a "comprehensive platform for the digitization, AI-powered text recognition, transcription and searching of historical documents[13]". A solution developed since 2013 thanks to European funding such as Horizon 2020, by several teams gathered by the READ[14] project, Recognition and Enrichment of Archival Documents, only partially open-source[15].

---

[8] CASTELLI B. *Lexicon Medicum graeco-latinum ante a Jacobo Pancratio Brunone iterato editum, nunc denuo ab eodem et aliis plurimis novis accessionibus locupletatum et in multis correctum, Edition nova accuratissima*, Genève, apud Fratres de Tournes, 1746, 760 p. Medica

[9] JAMES R. *A medicinal dictionary*, London, J. Roberts, 1743-1745, 3 vol. Medica

[10] JAMES R. *Dictionnaire universel de médecine traduit de l'anglais par Messieurs Diderot, Eidous et Toussaint*, Paris, Briasson, David l'aîné, Durand, 1746-1748, 6 vol. Medica

[11] https://pdf.abbyy.com/fr/

[12] https://readcoop.eu/transkribus

[13] *Idem*.

[14] https://eadh.org/projects/read

[15] See [MUEHLBERGER G. & al. 2018].

I. Pre-processing

Pre-processing is often put in place to improve the chances of successful recognition. With either approach, it proved impossible to leave it to an automatic process and various changes had to be made to both the software and the source file.

In order to use Transkribus, we set up a chain of automatic treatment of the images. According to the type of document (handwritten or old printed) and the level of wear of the paper such as crooked margins, curves, holes or ink stains, the application of all the parameters will not be necessary. Note, however, that they are all adjustable. In aim to do so, we used ImageMagick[16] a free software for image manipulation. Applied at the pixel level, these treatments may not be visible to the naked eye, but they intrinsically increase the image quality for OCR. They consist in reducing noise, making the image sharper and thus smoothing spots if there are any, automatically adjusting colour levels and increasing contrast without saturating highlights or shadows[17]. If the document was not correctly aligned when scanned, it may be necessary to automatically tilt it a few degrees clockwise or counter-clockwise to make the text lines perfectly horizontal or vertical. This can be done intuitively with ScanTailor[18], a free post-processing software for scanned pages. It performs operations such as page splitting, deskewing, adding/removing borders, and others.

For large formats and non-standardized works, a script for automatic detection of text[19] on a page image can also be set up in order to proceed to a cutting as close as possible to the content. Several actions are performed. First, the image is converted to greyscale and a binary threshold[20] with an optimal value set at 180 is applied. Binarization is a process for converting a colour or greyscale image to black and white. It is a simple way to separate the text, or any other element of the image, from the background. The lines of the image are thickened and the white spaces are reduced. The higher the number of iterations (15 in this case), the greater the dilation. This results in the identification of the contours of the different objects contained in the page image as well as the creation of a bounding rectangle around each *Region of Interest*[21], here text blocks. Sometimes the algorithm encounters intersecting or overlapping objects. It is therefore necessary to discard areas that are too small or too large to be columns. Finally, the other areas of the page are recorded.

---

[16] https://imagemagick.org/script/mogrify.php

[17] Here is the command that we used: **mogrify -despeckle -despeckle -despeckle -despeckle -auto-level -sharpen 10x1 -sigmoidal-contrast 2x50 -path dict -quality 100.jpg*

[18] https://scantailor.org/

[19] Note that a batch processing is possible thanks to the addition of an iterative function of the *glob* python module. The script has been written from the *OpenCv* library and the answers collected on Stackoverflow.

[20] Its efficiency for OCR programs has been studied by [Chamchong R., Fung C. C., Wong K. W. 2010] on manuscripts. The use of a binarization method or another one, will depend on the quality of the input images. To know more about this, see [Gupta M. R., Jacobson N. P., Garcia E. K. 2007] and [Milyaev S., Barinova O., Novikova T., Kohli P., Lempisky V. 2013].

[21] About new segmentation methods and automatic block detection, see [Clérice T. 2022].

There is then a necessary step of checking and renaming the generated files. Indeed, it is not possible for the moment to predict which column will be detected first.

ABBYY FineReader offers built-in pre-processing which mas simpler to apply and didn't necessitate a custom approach as with Transkribus, as it is not mandatory to exit the software to apply them. The programme can automatically determine and apply the necessary corrections depending on the type of image. Applicable corrections may include removing noise and blur, inverting colours to lighten the background relative to the text, correcting skew, straightening text lines, correcting trapezoidal distortions and cropping image borders. The most important and impactful treatment is standardizing the resolution around 300 to 400 dpi to allow for the best quality of each image across multiple files[22].

With either approach, we then proceed to the segmentation of the page images. This involves the recognition of text areas, form typing to ordering (layout analysis). A page layout can be segmented into one or several text regions, each with their own coordinates. Text lines can be nested in a text region (green). A line of text combines a baseline (blue) defined by at least 2 points and a text node. On Transkribus, the "Printed Block Detection" option is the most suitable for managing the lettering, images (red), columns and printed matter. In the case of a single page, it is also possible to use the "CITlab Advanced" option which allows the joint detection of text zones and lines. Although this option is relevant in most cases, words or expressions, Greek ligatures or even diacritics may escape detection. It is then necessary to manually adjust the polygons. The need for quality control across both software is evident has neither option managed to perfectly recognize the ordering of the texts zone.

II. First text recognition

The first pass of the OCR can indicate necessary changes with the pre-processing, to improve certain aspects, but will mostly serve as quality benchmarks. With ABBYY FineReader the first pass is made with its stock template allowing for a first test to see where improvement can be made and to start the training and the constitution of a template library. Whereas ABBYY FineReader is simple to use and necessitate afterward a lengthy training process left to an operator knowledgeable in both Greek palaeography and philology, Tranksribus require a different approach.Transkribus allows you to load images and perform automatic segmentation, transcribe them manually, transcribe them automatically thanks to a pre-existing template (for a fee) or train a transcription template. Training a template is not useful on a heterogeneous or small corpus. It is indeed necessary to manually transcribe at least 75 pages from the same hand[23] (or more for difficult texts).

[22] The limitation in both memory and computing power forces us to limit ourselves to about 400 pages by OCR files depending on image resolution. Each template and user's dictionary can be reused across the different files.

[23] To read feedbacks about training a model on Transkribus, see for example [VENTRESQUE V., MASSOT M.-L., WALTER R. 2022] or [SCHLAGDENHAUFFEN R. 2020].

If the pre/post processing ratio does not make you spend more time to obtain the same result as by transcribing by hand, then it can be useful to train a model on its data[24].

For our ancient multilingual medical dictionary corpus, we have chosen an integrated Transkribus model, the one trained by Stefan Zathammer in the framework of the NOSCEMUS[25], Nova Scientia: Early Modern Scientific Literature and Latin project funded by the European Research Council (ERC) from October 2017 to September 2022. It is able to read neo-Latin printed texts, in particular those of the 16th, 17th and 18th centuries.

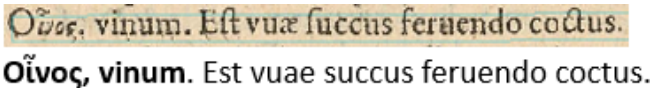

Figure 1 Excerpt of the transcription under Transkribus of the article "Οἶνος".
Jean De Gorris, *Definitionum medicarum libri XXIIII*, 1622, 443.

The latest versions include Greek and Latin prints from the 15th and 19th centuries. The technology used is of a mixed type since it uses the contributions of both printed matter, *Computational Intelligence Technology*, and manuscripts, *Handwritten Text Recognition*. The model was enriched throughout the project. It is therefore essential to consider its several evolutions.

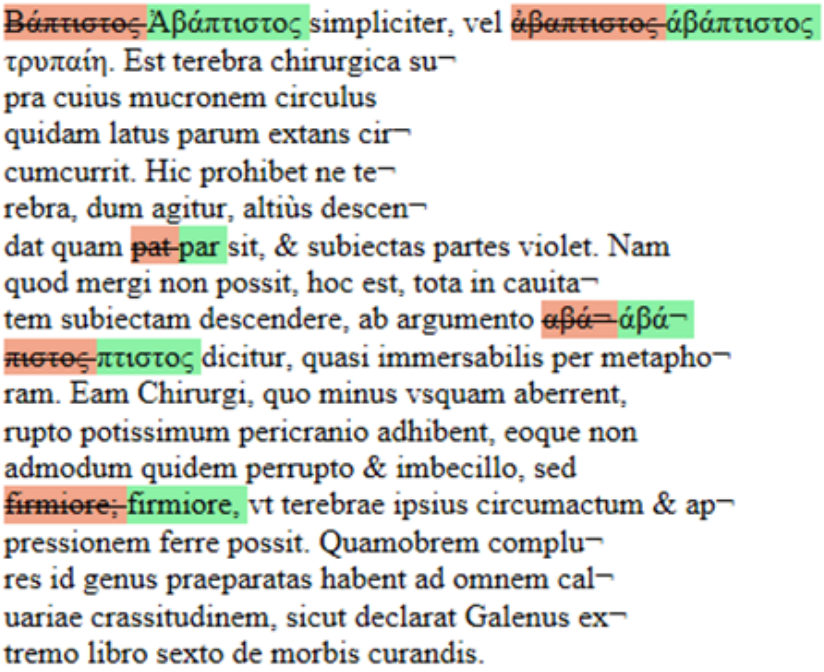

Figure 2 Comparison of Noscemus GM 4.0 and Noscemus GM 5.0 models.
Jean De Gorris, 1622, 1st column, 1st definition "Ἀβάπτιστος".

[24] To know more about this, see [CHAGUÉ A., CLÉRICE T., ROMARY L. 2021].
[25] https://www.uibk.ac.at/projects/noscemus/

With the last evolutions of the model, we notice that the lettering has been taken into account. Diacritics are also much better supported. Typographical errors have been corrected, as well as an error in Latin recognition. The terms left blank are those that have been uniformly recognized and therefore for which there is no significant difference between the application of one or the other.

Noscemus model	**GM 3.0**	**GM 4.0**	**GM 5.0**
Number of trained words	448 464	541 611	607 837
Number of trained lines	66 575	81 555	92 476

Table 1 Training evolution of Noscemus GM 3.0, GM 4.0 and GM 5.0

Compared to our corpus, we note that the latest version of the model performs better. This is mainly due to the increase in the amount of training data: between GM 3.0 and GM 4.0, we note an increase of 20.77% in the number of trained words and 22.5% in the number of trained lines. Between GM 4.0 and GM 5.0, we note increases of 12.23% and 13.39% respectively[26]. So, variety of characters forms has been introduced in the model, making it more complete but also rending the user dependent of the model's next improvements.

III. Post-processing and improvements

Training ABBYY FineReader to read Grec du Roi[27] font might seem like using a supercomputer to lemmatize Sappho's poetry, an exercise in using an overly powerful tool on a small ancient corpus. Yet the software provides two advantages one its huge stock templates allowing for recognition of bold and italic scripts[28]. This was essential to keep the structure of the dictionaries to automatically styled them during the passage to XML-TEI. Second, it's machine learning can be used, for an important time cost, to learn ligatured Greek usually providing few confusions with other characters. Where ABBYY FineReader struggle when it comes to ancient text is with a confusion between Latin and Greek scripts, something of a legacy issue[29]. This confusion tends to add noise to the resulting text and had the potential to misrecognize the lemma of a dictionary falsifying the linking of this entry with the rest. ABBYY FineReader can recognize suspect characters and attributes to each page a percentage of suspected mistake this is incredibly useful for improving the user model but cannot be taken as a quality assessment as custom and often rare character in comparison with Latin script such a ligatured Greek can often count as a suspected mistake even though it

[26] We can measure the character error rate (CER) and the word error rate (WER) of a training model with KaMI a python library developed by Alix CHAGUÉ and Lucas TERRIEL at INRIA (ALMAnaCH).
[27] See [BARKER, N. 1996] 99.
[28] It has already been used to digitize Greek critical edition [BOSCHETTI F., ROMANELLO M., BABEU A., BAMMAN D., 2009] 3.
[29] The issue was noted almost twenty years ago. [LE D., STRAUGHAN S., THOMA G. 2003] 1, 6.

is perfectly recognized. Any digitized text will, of course, come with imperfection and can be improved using several methods.

eScriptorium[30] is an open-source software, developed by research team Scripta (PSL). It provides an interface for document segmentation, layout annotation, transcribing (manually or automatically) and training OCR/HTR models[31]. The image files should be imported first and then the transcriptions in XML Page format[32]. It is imperative to proceed to a new segmentation of the page images according to the "Only line masks" option before trying to train a model from the data such as Kraken[33], OCR/HTR neural models trained on ancient manuscripts, for a refined segmentation method. In order to over-represent the Greek in our model, one possibility would be to use some pages of the *Theriacs of Nicander* which are edited afterwards the dictionary of De Gorris (1622). In order to ensure the correspondence of a Greek word with the approximate equivalents found, it is possible to use Python libraries for automatic processing of ancient languages such as cltk[34] or pie-extended[35]. They are complete and practical, but have been developed according to neural network technology and are therefore contextualized. Another way would be to compare the terms found with the entries of the Liddell Scott or in Pyrrha[36] api developed by ENC which contains lemmatized words. These are lists of general language lemmas whereas medical dictionaries are specialized ones. Levenshtein distance measure the difference between two strings can also be useful, but with dealing with diacritics is tricky. A combination of user verification and automatized word checking seems to be the only solution to increase the text quality.

Conclusion

The result of the METADICTMEDICA project was that each dictionary, each entity of the same corpus was made interoperable thanks to OCR and its applications. Our purpose in writing this paper was to offer a case study and a critical analysis of a digitization process and to see what could be learned from our case study. These techniques contribute to the creation of true knowledge systems. There is not one OCR, but many OCR/HTR ecosystems. OCR and HTR algorithms are complementary and can be adapted to different kinds of textual digital images, independently on their complexity and heterogeneity. They can be combined to produce the most efficient processing chain for your data. These technologies are innovative and promising, but the documentation often needs to be written to make them accessible[37].

[30] https://traces6.paris.inria.fr/
[31] To know more about the platform and the projet, see https://escripta.hypotheses.org and [KIESSLING B., TISSOT R., STOKES P. & STÖKL BEN EZRA D. 2019].
[32] To know more about the eScriptorium chain, see [SCHEITHAUER H., CHAGUÉ A., ROMARY L. 2021].
[33] https://kraken.re/master/index.html ; see [KIESSLING B. 2019].
[34] http://cltk.org/
[35] https://github.com/hipster-philology/nlp-pie-taggers
[36] https://dh.chartes.psl.eu/pyrrha
[37] For a French Tutorial of Transkribus, see [PERRIN E. 2019].

ABBYY FineReader is designed to process pages in series. While it can train Greek ligatures and thus handle the complexity of ancient Greek diacritics, the user is made dependent on built-in templates and user templates, limited to a certain number of characters. It is therefore necessary to find a compromise between a better recognition of characters and typography. Finally, being a proprietary software, its price can be prohibitive without institutional support, especially for an individual. It is possible to take advantage of ABBYY FineReader's features thanks to the free license. For longer use, it is possible to turn to open-source by building a multimodal chain. The 500 transcription credits offered on Transkribus allow for an estimate of nearly 2,500 pages transcribed with an OCR template or about 400 pages with an HTR[38]. It should be noted, however, that typography is not supported in the templates. There are, however, options to add markup elements. This is a free option and useful with small corpora. While the pre-existing templates perform well, they depend largely on the corpus they were trained on and whether they are updated. The resulting transcript can be improved to train a model with eScriptorium at no extra cost as long as the ground truth is respected. It is also possible to use the platform directly if no pre-existing model is adapted to your data and if you do not wish to use a command-line program like Tesseract[39], very efficient for OCR of contemporary prints as model increase in quality with each passing year.

Anaïs Chambat
Contractual PhD Student in Language Sciences
LT2D Laboratory - Jean Pruvost Center (EA 7518)
CY Cergy Paris University

Cahal Taaffe
Researcher in Byzantine history
Engineer in digital humanities

Acknowledgments

This research was carried out within the framework of the CollEx-Persée project "Multilingual medical metadictionary" of the French digital library Medica, supported by the Bibliothèque interuniversitaire de santé - pôle médecine, Paris Cité University, Sorbonne University, the Institut universitaire de France and Lorraine University. We would like to thank for their assistance as well as their trust and advice Frédéric Glorieux, Christophe Rey, Nathalie Rousseau and Jean-François Vincent.

References

BARKER N. "The relationship of Greek manuscripts and printing types in 15th century Italy." In *Greek Letters: From Tablets to Pixels* edited by MACRAKIS M. S. Delaware: Oak Knoll Press. 1996. 93–107.

[38] https://readcoop.eu/transkribus/credits/
[39] https://github.com/tesseract-ocr/tesseract/releases/tag/5.2.0
For a comparative point of view of several OCR tools, see [TAFTI & al., 2016].

BOSCHETTI F., ROMANELLO M., BABEU A. & BAMMAN D. “Improving OCR Accuracy for Classical Critical Editions”. 2009. 156–67. https://doi.org/10.1007/978-3-642-04346-8_17

CHAGUE A., CLERICE T. & ROMARY L. “HTR-United: Mutualisons La Vérité de Terrain !” In *DHNord2021 - Publier, Partager, Réutiliser Les Données de La Recherche : Les Data Papers et Leurs Enjeux*. (Lille), MESHS, 2021. https://hal.archives-ouvertes.fr/hal-03398740

CHAMCHONG R., FUNG C. C. & WONG K. W. “Comparing Binarisation Techniques for the Processing of Ancient Manuscripts”. In *Cultural Computing*, edited by NAKATSU R., TOSA N., NAGHDY F., WONG K. W. & CODOGNET P. Berlin: Springer. 2010. 333:55-64. https://doi.org/10.1007/978-3-642-15214-6_6

CLÉRICE T. *You Actually Look Twice At It (YALTAi): Using an Object Detection Approach Instead of Region Segmentation within the Kraken Engine*. 2022. https://hal-enc.archives-ouvertes.fr/hal-03723208

GEFEN A., SAINT-RAYMOND L. & VENTURINI T. “AI for Digital Humanities and Computational Social Sciences.” In *Reflections on AI for Humanity*, edited by BRAUNSCHWEIG B. & GHALLAB M. Berlin: Springer. 2020. https://hal.archives-ouvertes.fr/hal-03043393

GUPTA M. R., JACOBSON N. P. & GARCIA E. K. “OCR Binarization and Image Pre-Processing for Searching Historical Documents.” In *Pattern Recognition*. 40. 2007. 389–97. https://doi.org/10.1016/j.patcog.2006.04.043

KIESSLING B., “Kraken – A Universal Text Recognizer for the Humanities”. *DH2019*. Utrecht. 2019.

KIESSLING B., TISSOT R., STOKES P. & STÖKL BEN EZRA D., “eScriptorium: An Open Source Platform for Historical Document Analysis”. *International Conference on Document Analysis and Recognition Workshops* (ICDARW). Sydney. 2019. p. 19. https://ieeexplore.ieee.org/document/8893029

LE D. X., STRAUGHAN S. R. & THOMA G. R. “Greek Alphabet Recognition Technique for Biomedical Documents” In *LHC*. 2005. 6.

MILYAEV S., BARINOVA O., NOVIKOVA T., KOHLI P., and LEMPITSKY V. “Image Binarization for End-to-End Text Understanding in Natural Images.” In *12th International Conference on Document Analysis and Recognition*. Washington DC: IEEE. 2013. 128–32.https://doi.org/10.1109/ICDAR.2013.33

MUEHLBERGER G., SEAWARD L., TERRAS M., OLIVEIRA S., BOSCH V., BRYAN M., COLUTTO S. & al. “Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study.” In *Journal of Documentation* 75. no. 5. 2019. 954–76. https://doi.org/10.1108/JD-07-2018-0114

PERRIN E. “Tutoriel Transkribus”. November 2019.https://hal.archives-ouvertes.fr/hal-02472234

REBOUL M. *Comparaison Semi-Automatique Des Traductions Françaises de l’Odyssée d’Homère (1547-1955)*, Paris: Classiques Garnier, 2022. https://doi.org/10.48611/

SCHANTZ H. F. *The History of OCR Optical Character Recognition*. Manchester: Recognition Technologies Users Association, 1982.

SCHEITHAUER H., CHAGUÉ A. & ROMARY L. “From EScriptorium to TEI Publisher.” In *Brace Your Digital Scholarly Edition!* Berlin. France. 2021. https://hal.inria.fr/hal-03538115

SCHLAGDENHAUFFEN R. “Optical Recognition Assisted Transcription with Transkribus: The Experiment Concerning Eugène Wilhelm’s Personal Diary (1885-1951).” In *Journal of Data Mining & Digital Humanities.* 2020. https://doi.org/10.46298/jdmdh.6249

TAFTI A. P., BAGHAIE A., ASSEFI M., ARABNIA H. R., YU Z., PEISSIG P. “OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym.” In *ISCV*. Springer. 2016. 735–746.

VENTRESQUE V., MASSOT M.-L. & WALTER R. “Mettre en ligne, annoter et explorer les fiches de lecture de Michel Foucault.” In *Savoir(s)*. 2022. https://hal.archives-ouvertes.fr/hal-03694668