

Using eScriptorium to work on documents: eScriptorium applied on a 13th century French cartulary

Usando o eScriptorium para trabalhar em documentos: eScriptorium aplicado em um cartulário francês do século XIII

Michela Galli 

Sapienza University of Rome, Italy.
E-mail: michela.galli@uniroma1.it

Editors-in-chief

Marcus Dores
Célia Lopes

Guest Editors

Ana Pereira Ferreira
Leonor Dias Garcia
Ofélia Sequeira

Dossier

Paleografia e
Diplomática na senda
das Humanidades
Digitais: caminhos
e propostas

How to cite:

GALLI, Michela. Using eScriptorium to work on documents: eScriptorium applied on a 13th century French cartulary. *Revista LaborHistórico*, v.10, n.1, e63294, 2024. doi: <https://doi.org/10.24206/lh.v10i1.63294>

Abstract

As part of the research on Digital Diplomatics or on new tools usable in Diplomatics' studies, this essay discusses the experience of treating a 13th century French cartulary with eScriptorium. By analyzing the functioning of this open-source software created mainly for the automatic transcription of any type of document, in any language and on any medium, we will see step by step its application on the the Saint-Etienne-des-Grès church's cartulary. If the starting hypothesis proposed the software's use for the analysis of both internal and external features on the one hand and on the other the experimentation of several HTR models on a document tipology which by nature can include a great variety of different scripts, such as in this case, the work proved satisfactory for the recognition of the external features of the document and a good result was achieved in terms of automatic transcription. As far as the analysis of the internal features is concerned, it is currently not possible to carry it out directly on the software. By the way, at this stage the work one does on

* This work was prepared for the 1st International Congress of Palaeography and Diplomacy, which took place in 2021.

eScriptorium can be exported and then transformed into TEI, taking into account that the eScriptorium team is working on the integration of the TEI's principles into the tool, promising good results in the near future.

Keywords

Diplomatics, Digital Humanities, Cartulary, eScriptorium.

Resumo

No âmbito da investigação sobre a Diplomática Digital ou sobre novas tecnologias utilizáveis nos estudos de Diplomática, este estudo aborda a experiência de tratamento de um cartulário francês do século XIII com o eScriptorium. Analisando o funcionamento deste open-source software criado principalmente para a transcrição automática de qualquer tipo de documento, em qualquer língua e em qualquer suporte, veremos passo a passo a sua aplicação no cartulário da igreja de Saint-Etienne-des-Grès. Se a hipótese de partida sugeria, por um lado, a utilização do software para a análise das características internas e externas e, por outro, a experimentação de vários modelos HTR sobre uma tipologia documental que, por sua essência, pode incluir uma grande variedade de escritas diferentes, como é o caso, o trabalho revelou-se satisfatório para o reconhecimento das características externas do documento e obteve-se um bom resultado em termos de transcrição automática. Relativamente à análise das características internas, não é atualmente possível realizá-la diretamente no software. Aliás, nesta fase, o trabalho que se faz no eScriptorium pode ser exportado e depois transformado em TEI, tendo em conta que a equipa do eScriptorium está a trabalhar na integração dos fundamentos do TEI na plataforma, prometendo bons resultados num futuro próximo.

Palavras-chave

Diplomática, Humanidades Digitais, Cartulário, eScriptorium.

The context, a still open question: Digital Diplomats or digital tools for Diplomats?

Similar to other subjects within the study of written documents like Paleography and Philology, as well as other disciplines in the Humanities known as *Ancillae Historiae*, Diplomats has also questioned the potential benefits of incorporating computer technology into its practices. This contemplation has evolved over the past few decades with advancements in technology, resulting in an ongoing debate for nearly half a century.

It all started with a Round Table discussion on the topic of «Informatique et Histoire Médiévale» organized by the École française de Rome and the Institute of Medieval History of Pisa in 1975¹, where Robert-Henri Bautier and Alessandro Pratesi² in particular expressed themselves with only lukewarm enthusiasm and many doubts. The exploration of this subject continued in the subsequent years through the research of Michel Parisse, Michael Gervers and Manfred Thaller³.

It was however between 2007 and 2014 that a pivotal change in focus in the field occurred, largely due to three international conferences: *Digital Diplomatics. Historical research with medieval charters in a digital world* (Munich, 2007), *Digital Diplomatics. Tools for the Digital Diplomatist* (Naples, 2011) and *Digital Diplomatics. What is Diplomats in the Digital Environment?* (Paris, 2013). These conferences significantly advanced the exploration of how Diplomats and computers can collaborate in areas such as text encoding standards, digital critical editions, specialized tools, and databases.

In 2007, the term «Digital Diplomats» was first coined, and its limits had not yet been theorized. In 2009, thus, it assumes another meaning: Luciana Duranti, in fact, refers to Digital Diplomats in close connection with the InterPARES⁴ project, substantially enhancing the application of Diplomats studies on born-digital documents, while at the same time opening the doors to a reflection on the forensics of the electronic record⁵.

While an exhaustive critical study of what the scope of Digital Diplomats actually should be seems to be lacking, in 2014 Georg Vogeler gave a definition of it, implemented then in 2018, that keeps Diplomats anchored to its more traditional sphere, while giving it all the advantages that new technologies can provide.

According to Vogeler, Digital Diplomats in fact studies the external features, as well as the internal features together with the context of creation of the document from all points of view, «but it uses modern information technologies» and «this can

¹ Michele Ansani, «Diplomatica (e diplomatisti) nell'arena digitale», *Scrineum*, 1 (1999): 1-11 [Online] [Consult. 18 February 2024] Available at <http://www.rmoa.unina.it/19/>

² *Informatique et Histoire Médiévale. Communications et débats de la Table Ronde CNRS, organisée par École française de Rome et l'institut d'Histoire Médiévale de l'Université de Pise* (Rome, 20-22 mai 1975), présentés par L. Fossier, A. Vauchez, C. Violante, Roma, École française de Rome, 19 [Online] [Consult. 18 February 2024] Available at https://www.persee.fr/issue/efr_0000-0000_1977_act_31_1.

³ Antonella Ambrosio, «La Diplomatica e il digitale. Il Fondo della Biblioteca della Società Napoletana di Storia Patria online» *RTH - Education & Philosophy. Riconoscimenti, trasformazioni e misteri. Culture a confronto attraverso secoli di Autor*, 7 (2020) [Online] [Consult. 18 February 2024] Available at <http://www.serena.unina.it/index.php/rth/article/view/8912>.

⁴ <http://www.interpares.org/>.

⁵ Luciana Duranti, «From Digital Diplomats to Digital Records Forensics» *Archivaria*, 68 (2009): 39–66 [Online] [Consult. 18 February 2024] Available at <https://archivaria.ca/index.php/archivaria/article/view/13229>. On the same opinion see for example: Marie-Anne Chabin, «Diplomatique, diplomatique numérique, digital Diplomats» (2013) [Online] [Consult. 18 February 2024] Available at <https://bibliopaf.ebsi.umontreal.ca/bibliographie/RCYRHFZ3>.

change the methods of Diplomatics significantly⁶», for example by allowing work on large corpuses of digitized documents, including through statistical methods, aggregating single works on the same theme, producing online editions according to the use of controlled vocabularies and formal ontologies for metadata⁷.

Vogeler departed from a shared view that Theo Kölzer had well expressed in his essay *Diplomatik, Edition, Computer*, which was that Digital Diplomatics could be a turning point for the further development of traditional Diplomatics, but that it did not create «a new method and therefore not a new Diplomatics⁸».

Hence: Digital Diplomatics or digital tools for Diplomatics? This question remains unresolved, and this essay, along with the conference that inspired it, whose explanatory title is *Paleography And Diplomatics On The Digital Humanities Route: Pathways And Proposals*, seeks to contribute to this ongoing discussion by introducing a potential tool for Diplomatics that in the future could encompass all the tools Vogeler identified as fitting for the discipline: eScriptorium.

The eScriptorium

Between the end of 2017 and the beginning of 2018, an interdisciplinary and strategic research initiative (IRIS) called Scripta PSL was created at the University of Paris Science et Lettres, with the project «Scripta-PSL. Histoire et pratiques de l'écrit⁹». This project aims to integrate the academic disciplines related to Manuscript and Book Studies with those more closely related to Sociology and the Humanities and especially with the Computational and Digital Humanities. The project has involved a number of French academics¹⁰ and, above all, has created a team of experts in the field of Information Technology applied to the Humanities, who have developed, over the years, special tools for the study of writing in the most inclusive sense possible.

One of the outcomes of this initiative is the development of eScriptorium. The main creator's team is composed of Daniel Stökl Ben Ezra, Peter A. Stokes, Marc Bui, Benjamin Kiessling, Robin Tissot and Elhassan Gargem.

⁶ Georg Vogeler, “Digital Diplomatics: The Evolution of a European Tradition or a Generic Concept?”, in Simon Cubelic, Axel Zotter, Aastrid Michaels, *Studies in Historical Documents from Nepal and India*, 2018, 85–109.

⁷ Georg Vogeler, *Op. Cit.* see also Georg Vogeler, “Digital Diplomatics: What Could the Computer Change in the Study of Medieval Documents?”, *Initial. A Review of Medieval Studies*, 2 (2014): 163-185 [Online] [Consult. 18 February 2024] Available at https://www.academia.edu/11893608/Digital_Diplomatics_What_could_the_Computer_Change_in_the_Study_of_Medieval_Documents.

⁸ Theo Kölzer, “Diplomatik, Edition, Computer” in Georg Vogeler, *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*, (Archiv für Diplomatik - Beiheft 12; Köln-Weimar-Wien 2009), 13-27: 16.

⁹ <https://scripta.psl.eu/>.

¹⁰ Such as: École normale supérieure (ENS), École pratique des hautes études (EPHE), École nationale de chartes (ENC), École des hautes études en sciences sociales (EHESS), Institut de recherche d'histoire de textes (IRHT CNRS) e Collège de France.

Since then, the eScriptorium software has been continued in particular in the Resilience project, which is a European project¹¹ seeking to develop a 35-year long-term research infrastructure for religious studies¹².

What is eScriptorium?

If one opens the main eScriptorium web page¹³, will read that «eScriptorium is a project that provides digital recognition of handwritten documents using machine learning techniques» but it is actually more than that.

In fact, eScriptorium is:

- an open-source software that allows multiple instances;
- it provides open models thanks to an engine called Kraken¹⁴;
- it is cooperative and a strength of the software is data exchange, as well as manual editing and automatic transcription.

The main goal of eScriptorium is to serve as a valuable tool for creating digital editions. Originally designed as transcription software¹⁵, eScriptorium has evolved steadily, with the team working to integrate image annotation based on the principles of Archetype¹⁶ and text annotation in TEI¹⁷.

The software has been created for the purpose of automatic or partially automatic transcription of any written source, whether printed, manuscript, engraved or other, in any language or writing system placed on any type of support: once a writing model has been trained, for eScriptorium it will be indifferent to deal, for example,

¹¹ <https://www.resilience-ri.eu/> note that EPHE is among the 12 academic structures participating in the project.

¹² Although the Scripta project already includes a wide range of languages and scripts studied, the Resilience project aims to include even more by wanting to study all languages relevant to any aspect of religious studies in Europe, see Benjamin Kiessling et al., “The eScriptorium VRE for Manuscript Cultures” *Classic@ Journal* 18, 1 (2021) [Online] [Consult. 18 February 2024] Available at: <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>

¹³ <https://gitlab.com/scripta/escriptorium>.

¹⁴ <https://github.com/mittagessen/kraken>. Kraken is language-agnostic and fully open-source, it is developed in the same eScripta research group, see also note nineteen.

¹⁵ Peter Stokes, “eScriptorium : un outil pour la transcription automatique des documents.” *ÉpheNum: Veille, agenda et actualités des humanités numériques à l’EPHE*. 2020 [Online] [Consult. 18 February 2024] Available at <https://ephenum.hypotheses.org/1412>, English version available at <https://www.resilience-ri.eu/blog/resilience-tool-escriptorium/>.

¹⁶ Archetype is an open-source software for the analysis and presentation of scripts from a paleographic perspective. In 2010 the European Research Council awarded a Starting Grant to Peter Stokes for the project Digital Resource and Database for Paleography, Manuscript Studies and Diplomatic (DigiPal), which ended in 2014 and now lives on through several projects and in Archetype. Archetype allows scholars to search, visualize and organize detailed features of handwriting or other material in both verbal and visual form. More information can be found at: <https://github.com/kcl-ddh/digipal/wiki>.

¹⁷ The Text Encoding Initiative is a consortium of international linguistic and literary institutions that has developed a standard for the representation of texts in digital format. More information can be found at: <https://tei-c.org/>.

with a corpus of wax seals whose inscriptions are in Latin and French uncial script, or a stele written in hieroglyphics with an arrangement of signs in columns.

Regarding models, we can describe a writing model as the perfect depiction of a visual system that conveys a mode of communication. In order to identify the symbols and generate the appropriate writing formats, eScriptorium utilizes the Kraken engine¹⁸. This engine was developed by Benjamin Kiessling from the OCRopus system and is an OCR/HTR¹⁹ engine for recognizing written text, optimized for historical and non-Latin script material. Written in Python, the Kraken engine is particularly well suited to serve a wide range of graphic communication models since it breaks the usual process of traditional OCR by segmenting page images into lines, then each line into words, and then each word into characters, which is a problematic attitude when faced with scripts which aren't set, be it cursive such as 8th century Merovingian or Arabic script. Kraken eliminates the issue of word/character segmentation by instead employing a form of machine learning called *Neural Network* that enable Kraken to «learn» from transcriptions (training data) to recognize letters in the images of entire lines of text²⁰. One of its key features is its flexibility: designed to minimize assumptions about writing systems, eScriptorium is highly modular, with numerous parameters in each module. Being open-source, users can customize settings to suit their specific requirements or even develop new modules, as long as they are proficient in Python programming.

Not only is Kraken well suited to working with a large number of scripts, but it also performs well when faced with *mise en page* that is more complex than full-page or columnar text, as well as in scenarios where the direction of writing is not left-to-right²¹.

¹⁸ Some templates are accessible in Zenodo's «OCR/HTR model repository» at this link: https://zenodo.org/communities/ocr_models/search?page=1&size=20, and in HTR-United reachable at <https://htr-united.github.io/catalog.html>. Note that not all the models implemented so far are available and that the repository is open-source and collaborative, this means that it is possible to request that a model of one's own creation be included among those proposed. The model before being made available to the whole community is checked by Benjamin Kiessling, who manages the repository as well as being the developer of the Kraken engine.

¹⁹ OCR stands for Optical Character Recognition and HTR for Handwritten Text Recognition. For more specifications see note three of Benjamin Kiessling et al., *Op. Cit.* e Benjamin Kiessling, «Kraken – A Universal Text Recognizer for the Humanities» DH 2019 (2019) [Online] [Consult. 18 February 2024] Available at <https://dh-abstracts.library.virginia.edu/works/9912>.

²⁰ See note number three in Matthew Thomas Miller, Maxim G. Romanov and Sarah Bowen Savant, «Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans», *International Journal of Middle East Studies* 50, 1 (2018): 103–109, <https://doi.org/10.1017/S0020743817000964> [Consult. 18 February 2024] Note that Kraken is particularly able to handle scripts with a wide variety of ligatures such as Arabic and Persian on which it has already been tested and projects are underway.

²¹ Benjamin Kiessling, *Op. Cit.*: «Full trainability of layout analysis is of utmost importance to a truly universal OCR system, as text layout and its semantics varies widely across time and space, e.g. hand-crafted methods for printed Latin text are unlikely to work reliably on Arabic text or manuscripts with extensive interlinear annotation».

How does eScriptorium work?

Let's now take a brief look at how eScriptorium works²². An example and a more detailed analysis will be given on the following pages where we will study the cartulary under exam. For a more detailed description of the workflow please read the step-by-step instructions written by Alix Chague from the LECTAUREP project²³.

First of all, it is necessary to access eScriptorium

To use *eScriptorium online*, you must request permission to access the platform using the «Contact» form which can be found in the header of the homepage of institutions using eScriptorium, such as <https://escriptorium.paris.inria.fr/> or <https://msia.escriptorium.fr/>, on the left side. eScriptorium can also be installed locally using two solutions: one involves the use of Docker, and the other way by installing and configuring everything yourself, both of which require knowledge of programming languages and a high-performance computer²⁴.

Recent developments in the eScriptorium project have led to a modification in the initial step required to begin a task, showcasing the project's ongoing advancements. Whereas until recently, the first step involved creating a work through the «My Documents» function, which allowed the upload and definition of parameters for a single document, it is now possible to start by setting up an entire project that can include one or more documents, to which customized parameters can be assigned one by one.

Once logged in, one will find the «My Projects» and «My Models» functions in the header on the right side of the site, the latter concerning writing and segmentation models. Entering «My Projects», one can access the repository of previously developed projects or can create new ones using the «Create new Project» function, which will ask to give the project a name.

The subsequent stage involves creating the primary description and ontology, which is a kind of ID card of the document, necessitating an initial examination of the document being worked on. Information such as the document's title, primary script, reading direction, region, and line types must be specified. Each choice is reversible at any time, and indeed the use of eScriptorium will help to understand which are the best study choices to apply to the document.

²² For all information about eScriptorium and its functionalities, please refer to <https://escripta.hypotheses.org/>.

²³ The French version is accessible at <https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium> and the English version at <https://lectaurep.hypotheses.org/documentation/escriptorium-tutorial-en>. LECTAUREP is a notarial archives study platform developed with eScriptorium.

²⁴ The resource code from which the software can be installed is <https://gitlab.com/scripta/escriptorium/>.

The third step is to prepare the images to work on. There are various format options including PDF files or XML format, but if the document has been scanned in IIIF you can easily import the images automatically thanks to the IIIF manifest which will also add the metadata it already holds to the description²⁵.

In the fourth step, the initial task is to identify lines of text and other significant elements (Line types) and to find significant areas (Region types) that can be pinpointed in the image. This allows the machine to learn from the prepared images, enabling it to independently locate lines and regions thereafter. One can alternatively load an existing segmentation model.

In the fifth stage, the process involves transcribing a set of images to enable the machine to learn script recognition (Ground Truth). Another option is to load and utilize an existing model. Subsequently, the machine is trained using the previously prepared images, allowing for automatic transcription. It is important to review the transcript generated by eScriptorium, correct any errors, and retrain the machine for improved accuracy.

At this point, whether one is satisfied, the result can be exported in Alto, Text or XML, for markup, publication or any other requirement.

This concise overview of eScriptorium provides insight into the benefits of HTR software in automatically transcribing books and manuscripts in any language and on any medium from digitized images. It not only offers a significant advantage for transcribing collections of documents belonging to the same environment, but it can also be helpful for rough transcription of unfamiliar documents and for handling materials that may be in a fragile state of preservation.

Before carefully examining the application of eScriptorium on the cartulary under exam, it will be useful to say a few words about cartularies in general and to provide an overview of the cartulary of Saint Étienne-des-Grès.

Cartularies in short

A cartulary is a collection of copies of one's own documents, established by a natural or legal person, who, in a volume or, more rarely, in a scroll, transcribes or has transcribed, in full or sometimes in extracts, titles relating to his or her property and rights, and documents concerning his or her history or administration, in order to ensure their preservation and to facilitate their consultation²⁶.

²⁵ <https://iiif.io/> International Image Interoperability Framework (IIIF), develops an interoperable environment that allows the different software applications that manage digital images via the Web to talk to each other effectively, so as to provide users with a level of uniform and rich access to image-based resources hosted around the world.

²⁶ Translated from: *Vocabulaire international de la diplomatie* [Online] [Consult. 18 February 2024] Available at <https://www.cei.lmu.de/VID/>.

The study of cartulary has experienced a significant increase since the early 1990s with the organization of a dedicated Round Table in Paris in 1991²⁷. This event marked a period of renewed recognition of cartularies as a major historical source, particularly in terms of their significance as witnesses to the social practices of the written word and scripturality and from social history perspective too²⁸. The growth of investigations in the field has been steady since the 2000s thanks to further colloquia, thesis work and the expansion of IRHT's CartuIr project²⁹ (now migrating to Regecart³⁰) and culminates with the launch of the MECA, or Medieval European Cartularies³¹, program sponsored by the École française de Rome in 2018. Thanks to this, the last few years have seen remarkable contributions on the topic flourish across Europe and research is operational and current.

An overview of the Saint Etienne-des-Grès' cartulary

A necessary premise: a cartulary is not just a manuscript which serves as a repository of the acts in which an legal or natural person is involved, but is also a document of use, and the manuscript itself bears the physical traces of this actual usage. Moreover, the structure of the manuscript itself is not linear, but probably determined by reasons of utility.

A perfect codicological reconstruction is not possible. The manuscript isn't in an ideal state of preservation, and handling it requires great care, as the front plate is detached, held together with the manuscript only by a cardboard cover used by the Archives Nationales de France.

Collocation: A.N. LL\465³²

The manuscript is made of parchment and consists of 63 folios. There are two parchment guard-leaves with fragments of accounts at the top and at the bottom, the above-mentioned at the top being marked LL 558, today corresponding to the «Deliberations» dossier of the Chapitre de Saint-Marcel, 18 avril 1557 - 12 décembre 1561. A false guard-leaf in paper at the head and tail.

²⁷ The results of which are available here: Olivier Guyotjeannin, Laurent Morelle and Michel Parisse (eds), *Les cartulaires: Actes de la table ronde organisée par l'école nationale des chartes et le G.D.R. 121 du C.N.R.S. (Paris, 5-7 décembre, 1991), Mémoires et documents de l'École Nationale des Chartes* 39 (Paris, 1993).

²⁸ Cfr. Within others the studies of Paul Bertrand and Pierre Chastang.

²⁹ <https://telma-repertoires.irht.cnrs.fr/cartulr/page/presentation>.

³⁰ <http://regecart.irht.cnrs.fr/>.

³¹ <https://www.efrome.it/en/meca>.

³² In detail: LL - « Monuments ecclésiastiques » - Titre IV : Collégiales et paroisses du diocèse de Paris - 465. Cartulaire. (XIIIe s.).

Measurements in mm are 120x162 (f. 20).

The binding is modern: it includes cardboard plates covered in red leather with titles and decorations stamped in gold, the front plate is detached. On the spine there are 5 ribs; it reads: ST / ETIENNE / DES GRES / CARTULEIRE / XIIIe SIECLE. The binding is currently covered with beige cardboard so as not to lose the front plate. The place of copying is Paris, at the Church of Saint-Etienne-des-Grès and the manuscript could be dated to the 13-14th centuries; the chronological limits of the acts copied in it correspond to 1187-1314.

The fasciculation is particularly heterogeneous, with the following details (modern numbering is used to indicate the folios):

- I. fascicule = Sexternion, ff. 1-12, catch-word *annuatim* in the lower right corner;
 - II. fascicule = Sexternion, ff. 13-24, catch-word *in singulis* in the lower right corner;
 - III. fascicule = Quinion, ff. 25-36, catch-word *salutem* in lower right corner;
 - IV. fascicule = Binion, ff. 37-40;
 - V. fascicule = Ternion, ff. 41-46;
 - VI. fascicule = Binion, ff. 47-50 (where f. 50 contemporary numbering is missing);
 - VII. fascicule = Ternion, ff. 51-56 (where f. 56 contemporary numbering is missing);
 - VIII. fascicule = Ternion, ff. 57-62 (where f. 57 is contemporary numbering a missing);
- f. 63 was added by a stub to which the bottom guard-leaves, numbered 64-66 in pencil, are also attached.

The sheets are not ploughed, so the guiding holes for scoring appear on each folio, the scoring appears averaging³³ to be dry-pointed, and there is a full-page writing mirror averaging 105x138 mm (f. 21) of 21/22 lines.

Writing and hands

It will be extremely interesting to analyse handwriting, which will be particularly useful later when looking at the transcription results in eScriptorium.

We present here only the principal text of the document, leaving out the hands that have made comments in the margins of the codex. At least eleven hands can be identified in the document, all of which can be assimilated into the domain of gothica:

³³ See the next page «Copyist 5».

- Copyist 1: his hand is found from f. 1r to f. 50v with the exception of f. 24v lines 13-21 and f. 25r lines 1-2. Similar to a praegothic³⁴, the handwriting is set and legible. The strokes of the individual letters are clearly identifiable and the contrast is not very marked. The letters «l», «h» and «b» present the forking of ascenders; «d» is always uncial and sometimes ends in a curve to the right, even after straight letters; «i» often presents a stroke on the top; «r» has two shapes: both straight and in the form of 2 after a curved letter; «s» mostly straight, is sometimes found round at the end of the word. Ligatures «ct» and «st» are present. The abbreviation of «p» with a cut on the descender is used for «per», «par» and «por».
- Copyist 2: his hand is found in ff. 51v-53r. Very similar to the previous one, it is characterized by the descending final minim below the line of «m», «n», «h» and the ascending stroke of «d» which ends with a clockwise stroke.
- Copyist 3: the hand is found at ff. 24v ll. 13-21, 25r ll. 1-2; 53v-54v; 58r. Similar to Copyist 1, it is distinguished by more marked strokes and contrast. Note uncial «a» with the usually straight stroke curved and the absence of rounded «s» at the end of the word.
- Copyist 4: the hand is found at ff. 55r-57v. It's a documentary gothica: set, clearly legible and characterized by a huge presence of gracious elements and «i» with a stroke on the top.
- Copyist 5: the hand is found at ff. 58v-60r ll. 1-10, is a documentary gothica. Very similar to the previous one, it is distinguished by a more elongated and narrow appearance. Note the scoring, which here appears to be made from a plumb line.
- Copyist 6 is found at f. 60r ll. 1-11 and f. 60v ll. 1-14, the writing is a gothic cursive with single-compartmented «a»; «b», «h», «l» with ascenders with loops at the right side, looped «d»; there are not excessive ligatures.
- Copyist 7: his hand is found on f. 60v ll. 15-21 and 61r ll. 1-12. The hand copies a document underwritten in the interline, both in gothic cursive rich of ligatures.
- Copyist 8: his hand is on f. 62r, ll. 13-22. It's a gothic cursive characterized by uncial «a»; «h» has a right stroke that ends curved below the line.
- Copyist 9: this hand is on f. 61v. It's a gothic cursive: uncial «a»; «b», «h», «l» with ascenders with loops at the right side, looped «d»; the final minim of «m» and «n» is curved.
- Copyist 10: his hand is found on f. 62. It can be compared to a documentary gothic, set, airy. The ascending and descending strokes terminate with an

³⁴ Albert Derolez, *The Palaeography of Gothic Manuscript Books: From the Twelfth to the Early Sixteenth Century*, (Cambridge: Cambridge University Press, 2003) 56-71. Although Derolez only studies book specimens, we will adopt his classification here even if it is a manuscript belonging to the field of documentation.

extension of the horizontal stroke. «d» is always uncial; «r» has two forms: both straight and in the shape of a 2 after a curved letter; «s» is round at the end of the word ending with a clockwise circular stroke.

- Copyist 11: the hand is found on f. 63. It can be defined as a semi-cursive documentary. The writing is oriented to the right, and although the hatching of the letters is typically cursive (i.e. «a» round, «r» always in the form of a 2, looped «d») the writing appears rather set. Note the leftward slant of looped «d» and the thickness of the descending stroke of «s».

The titles, which are rubricated, are written by the same hand as the person writing the document.

Black ink is used for the main text, red ink for the title given to the document and to decorate the small capital letters, as well as for the initial letter of each document and the corresponding paragraph mark. Small decorative interventions in red ink are also used to fill in spaces. The red ink is excluded from f. 58v onwards.

A brownish-yellow ink is used for the first complete numbering, and then only for the Roman numbering in the center of the upper margin, for the numbering of the copies, for some notes in the margin; the writing of old numbering seems to be contemporary with that of the document.

Other inks of a rather brownish color are used for notes in the margin, for underlining the text and sometimes for tracing *maniculae* indicating underlining or important parts of the text. The language of the notes is sometimes Latin, sometimes French or both. All documents are written in Latin.

There are five types of numbering at the beginning, becoming three as the folios pass.

- Upper margin in the center of the recto: Roman numerals are found, in brown ink, they indicate the folios for the whole length of the codex except ff. 62-63.
- The previous numbering is repeated on the verso, in Roman numerals, in the upper left corner in brown ink, for the entire length of the codex.
- On folio 1v it is indicated *prima charta* and from the following folios we understand that this is the numbering given to the copied documents, which are placed side by side in the margin and next to the title or next to the first lines.
- Upper right corner of the recto of the folios: on the first folio marked «I fo» in brown ink up to the V f. In the same place, we have a successive numbering in fine pen which numbers in Arabic from «6. fol» to «63. fol».

- In the same place, but only for a, b, 1-5 and 64-66, we have a contemporary numbering in pencil in Arabic numerals.
- Other numbering is found in the margin, not yet identified.

The cartulary contains 108 copies of documents, most of which are *lettres d'officialités*, where words and diplomatic formulas are standardized contributing to making these acts almost interchangeable³⁵. The text also abounds in technical, precise expressions, charged with the relevant description of family possessions, relationships and legal situations, of the land and its status. Only one of the 108 documents dates from the 14th century.

We turn to a relevant codicological aspect to see the content of the acts in more detail.

In the margins of some folios there are sometimes 1 or 3 cuts and in two occasions some parchment has been removed.

The folios concerned are: 2r, 3r, 13v, 18v, 21v, 28v, 35v, 41r, 44v, 53r, 56r and 60r (based on the modern numbering, placed in the upper right corner). As only the cuts remain, it is difficult to know exactly what they are, but the repetition of the three cuts on several folios would suggest the presence of bookmarks³⁶.

Where the subtraction of parts of the parchment is found (3r and 35r), the transcribed document exposes a new regulation to be adopted by the chapter.

On f. 18 (1 cut), ff. 41r, 44v, 53r (3 cuts) the institution of new prebends in three cases and of a new canon in one case are mentioned.

The other documents concern sales of acres or revenues and in one case on f. 13r both (this folio under the first cut identified has three cuts on the outer margin and three on the lower margin).

eScriptorium applied to the cartulary of Saint Etienne-des-Grès

Now that we have a clear picture of the use of eScriptorium and the cartulary in question, it will be easy to observe how the document was treated in eScriptorium.

The initial hypothesis was that the Regions could be applied not only to the codicological study of the manuscript but also to the diplomatic one, being able to underline in the page space the internal features of the acts such as *salutatio*, *corroboratio*, *datatio* and so on.

³⁵ For an exhaustive example look at <http://theleme.enc.sorbonne.fr/dossiers/notice20.php>.

³⁶ I would like to thank Marlène Helias-Baron, Sébastien Barret and Laurent Morelle for the valuable discussion and for pointing me to the study text: Paul Géhin, *Lire le manuscrit médiéval*, (Paris: Colin, 2005) 241, in particular he found out «signets de tête» (leather ribbons or straps, attached to the headband, cap or ear) and «signets de gouttière» or «onglets» (small rectangles of leather, parchment or strong paper, attached to one of the outer edges of the volume; sometimes these are tabs half-cut out of the leaf itself, folded outwards, passing through an incision and called: charter-tail bookmarks).

After really becoming aware of how the machine could work, it was preferred to concentrate on external features and transcription. By the way, these possibilities can be tested soon perhaps with an *ad hoc* adaptation of the NER Technology (named entity recognition), which involves the identification and categorization of important information in text known as named entities.

We previously said that description and ontology correspond somewhat to the ID card of the piece under examination. In our case, the starting point is that the document is written in Latin, left to right. Region types correspond to the areas of the document detected as important. If Kraken can attribute polygonal shapes to the different areas, this feature for the manual option is not yet available. In our case, the study led to the identification of the following areas: Catchwords, Commentary, Cut in the parchment, Document's numbering, Foliotation, Illustration, Main, Maniculae, Old numbering, Other numbering, Title.

The Line types correspond to the lines that will be transcribed and, in this case, correspond to Catchwords, Correction, Main, Numbering, Signature, Title.

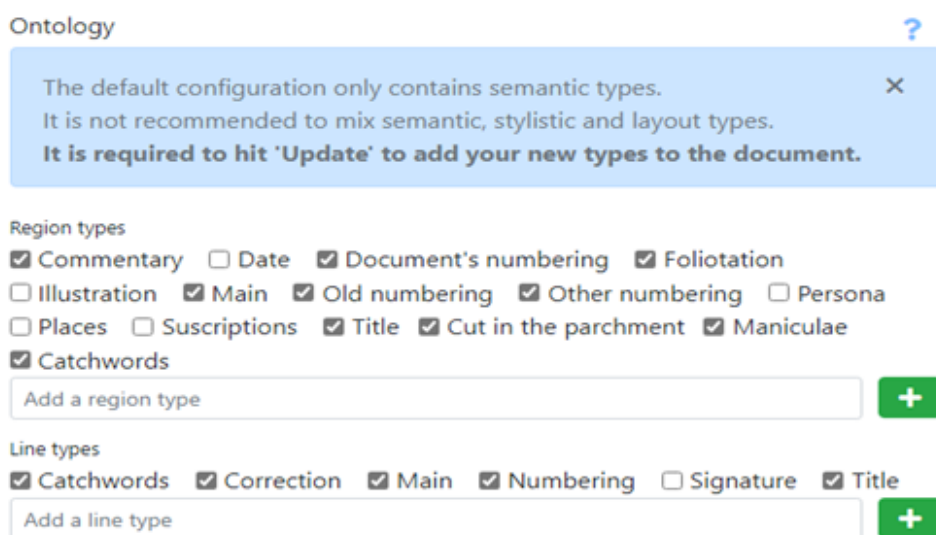


Figure 1 Region types and line types set for the analysis of the cartulary: the characteristics considered relevant for the study are ticked; it is possible to observe how the external characteristics were finally preferred to the initial hypothesis, more inherent to internal characters of the document.

The next step is importing the images.

At the Institut de Recherche et d’Histoire des Textes (IRHT) the microfilm of the Cartulaire AN LL\465 is stored, corresponding to the piece under examination with the number 19023. However, the reproduction is not always in focus and some page openings are obscured or repeated: the first digitization import in eScriptorium as well as the first model resulting from the transcription of the cartulary were based on this reproduction.

Since August 2021, digitization from this reproduction of the cartulary has been available in the IRHT's ARCA³⁷. However, although the IIF protocol is used, as the digitization belongs from a reproduction that is not of the highest quality, its use is not recommended.

A digitization of the manuscript then has been carried out using a color camera at the Archives Nationales de France and this is the reproduction on which most work has been done.

Once the images have been imported, work begins by setting the Region type and Line type on the page. It is possible to perform this step either manually or automatically, and then correct any errors detected in the layout carried out by the machine. As in the case of transcription, it is also possible to use an already created model or to create a new one for the segmentation part³⁸.

The step of correctly arranging the lines on the page is a very important one if one is not dealing with simple text, but perhaps accompanied by comments or other partitions of the page to be transcribed, and this task must be done carefully³⁹.

It is not only about finding lines, but also sorting them into a suitable order for the human being that has to read it. One must ask oneself: *which one is the good order?*

Make a decision, train the machine and it'll follow your own choice. In this case, it was preferred to put the various numberings of the folio or document's copy ahead, followed by the title and text of the document, and finally the comments.

Transcription

Before moving on to the fundamental part of the study, let us expose a detail. Note that there are two different types of users that eScriptorium can provide:

- on the one hand, there is a basic user (such as mine⁴⁰) for which every step that requires the machine to be trained must be requested from an administrator;
- on the other hand, there is an advanced type of user who can directly train the machine, creating and applying templates as required.

³⁷ Once known as BVMM, ARCA is the digital library of medieval manuscripts from French libraries across France, made available by the IRHT e the digitalization of the cartulary de Saint-Etienne-des-Grès, is now available at <https://arca.irht.cnrs.fr/ark:/63955/md8605742h7n>, IIF MANIFEST: <https://api.irht.cnrs.fr/ark:/63955/f7vm6nqwz8ps/manifest.json>.

³⁸ All models for both segmentation and transcription can be found in the My Models section.

³⁹ It exists a baseline segmenter that does not require binarization and is able to deal with curved lines and highly complex layout as well as deteriorated material, Benjamin Kiessling, *Op. Cit.*

⁴⁰ I would like to thank Peter Stokes for his patience, suggestions and continuous discussions during the *Humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien* aa 2020/2021 seminar, as well as my colleagues in the seminar who contributed relevant and interesting observations from time to time.

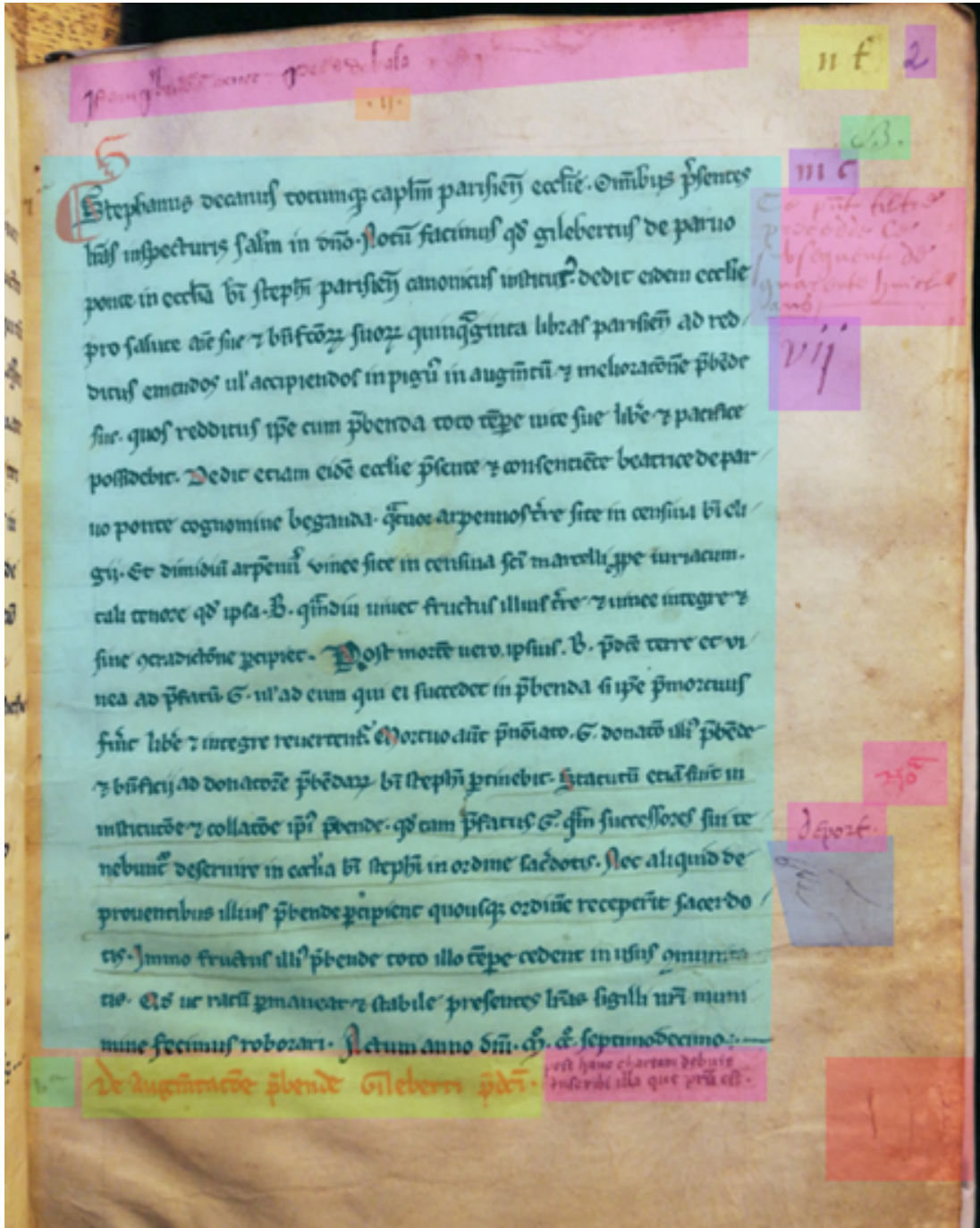


Figure 2. f. 2r as an example of Region types set, where: in light blue the main text, in pink the comments, in orange ancient numbering, in yellow numbering (top) and title (bottom), in purple other numbering, in red cut on the parchment, in blue manicura, in green other signs.

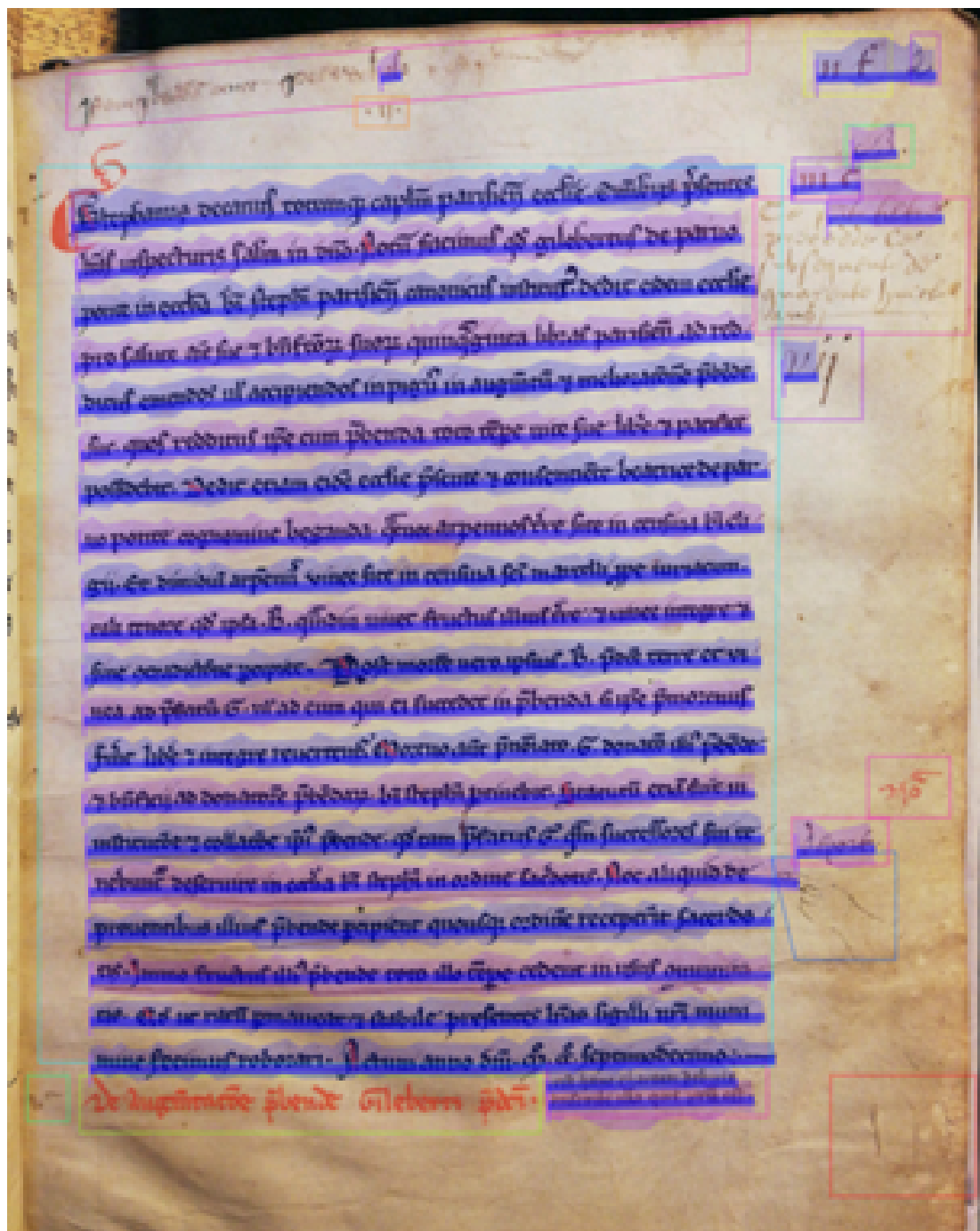


Figure 3. Authomized segmentation.

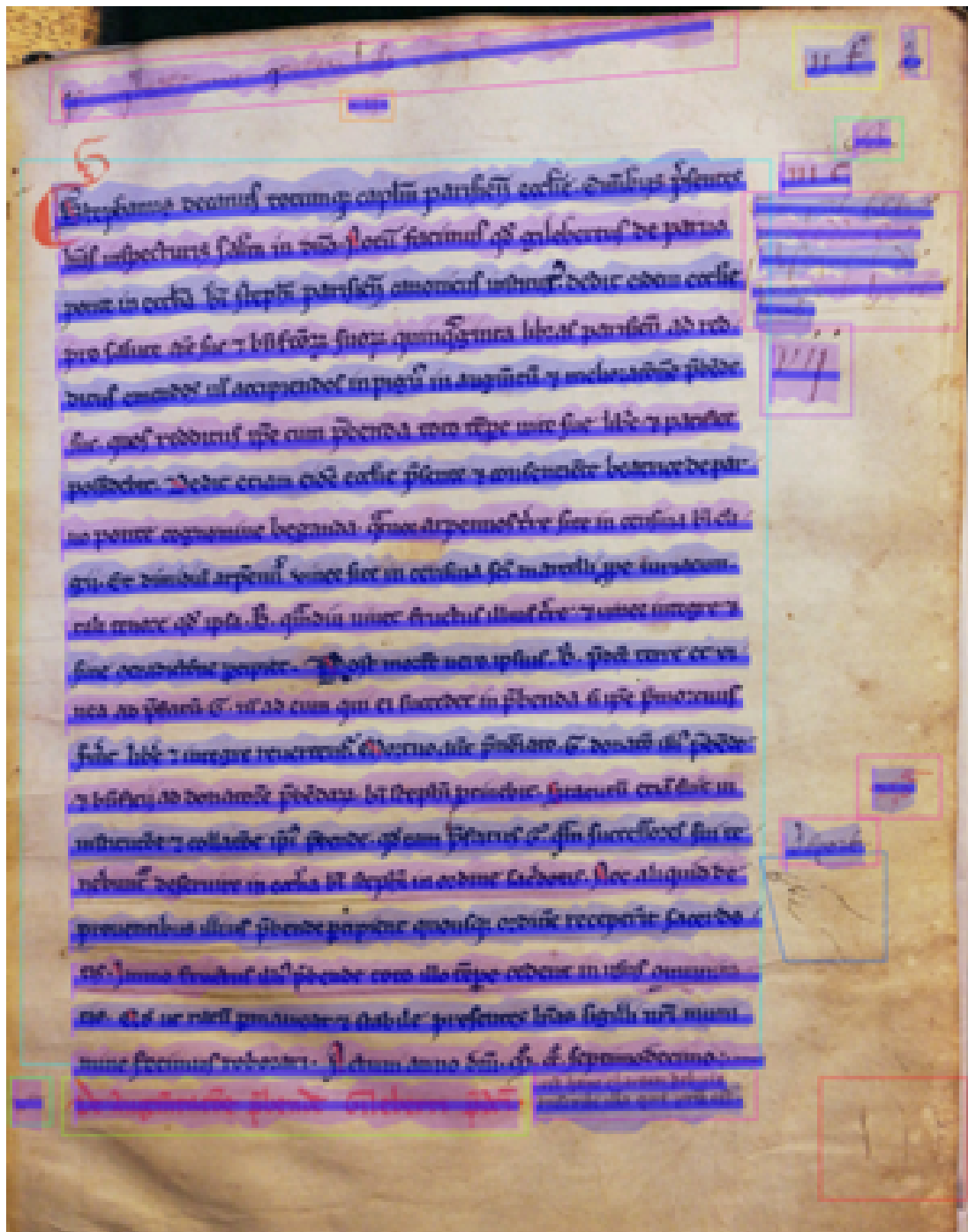


Figure 4. Manually corrected segmentation.

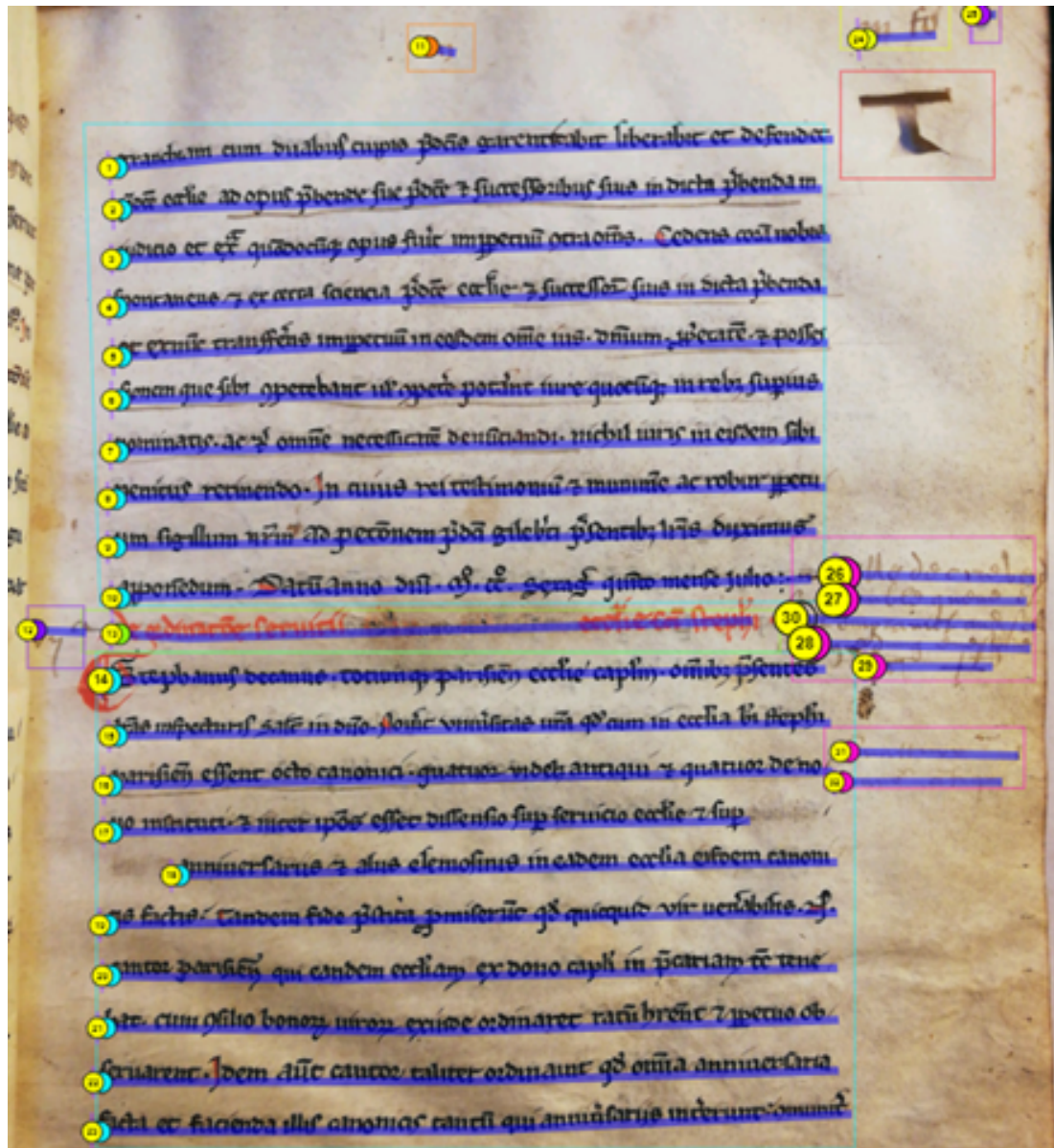


Figure 5. Automatic line's numbering.

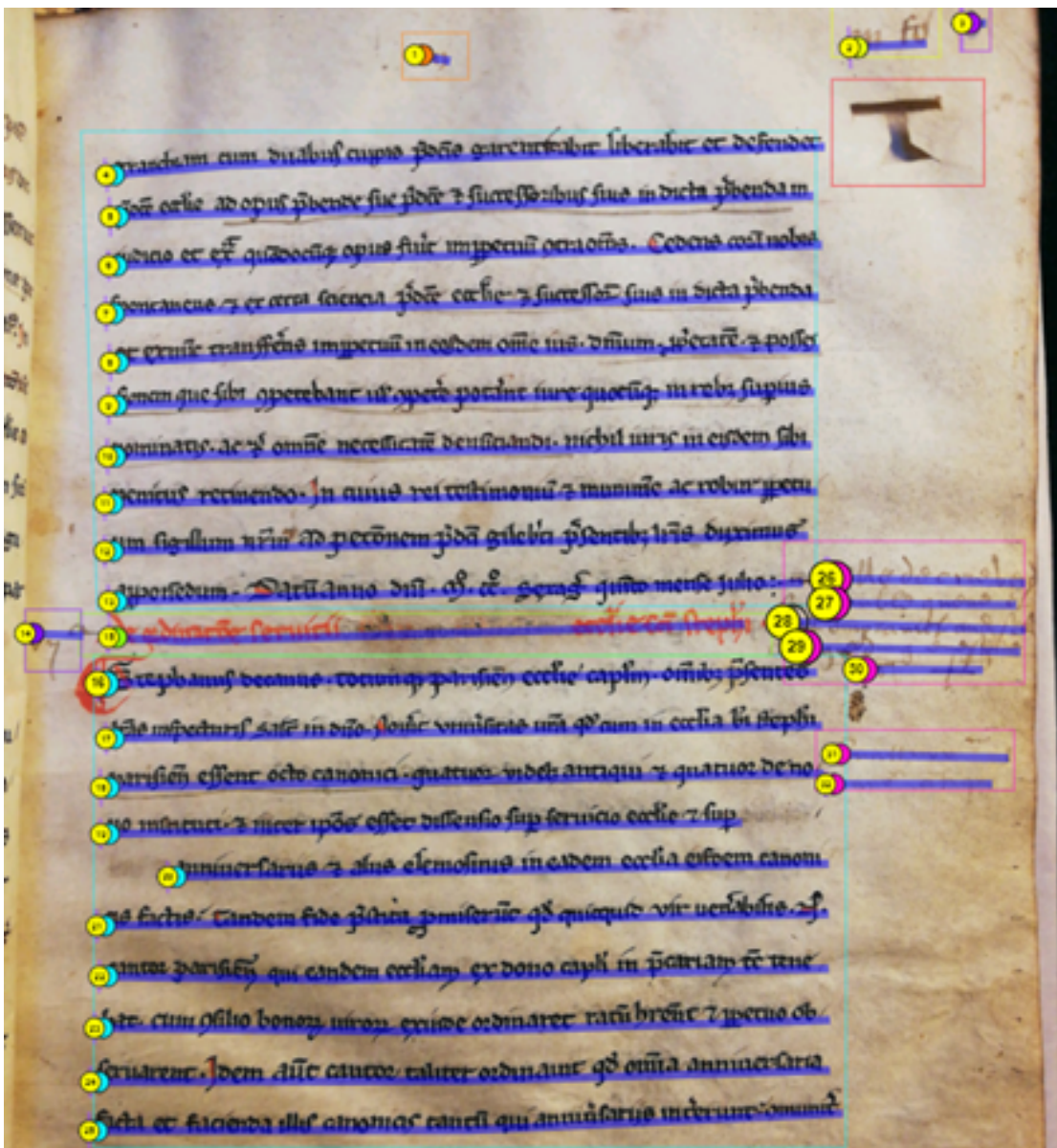


Figure 6. Manual line correction according to your needs.

Transcription can be done manually, semi-automatically or automatically from an existing model. To create a writing model, one will need to transcribe a maximum of pages of a document, train the machine, then correct any errors and repeat the transcription process until the result is satisfactory⁴¹.

One of eScriptorium’s strengths is certainly its ergonomic transcription.

To transcribe a line of text, simply select it and the transcription window will automatically appear on the screen.

⁴¹ Whole scripts and certain graphs not represented in Unicode can be encoded through use of the Private Use Area, Benjamin Kiessling, *Op. Cit.*

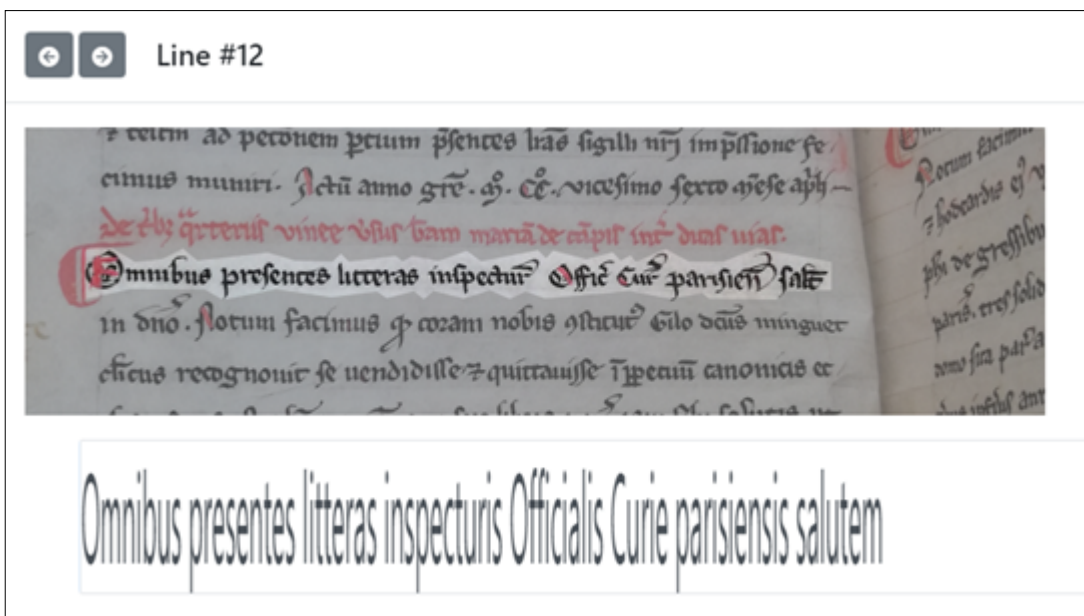


Figure 7. Line transcription window from the transcription panel.



Figure 8. Full screen showing the uploaded image (left), the binarized images (center) and the transcription (right).

The transcription and visualization panels are fully synchronized so the user can always look at the work done/in doing.

My exploration into transcribing the cartulary of the Collegiale of Saint-Etienne-des-Grès began with manually transcribing the images from the IRHT microfilm. The initial transcription was challenging due to the low quality of the images, leading to the development of a writing template based on the first thirty pages, called cart_30p. This model, once utilized, resulted in an accuracy level of 63.3%.

While this accuracy rate may seem modest, various factors need to be considered when evaluating the success of such a model.

1. Material aspects and physical characteristics of a manuscript can significantly impact how the engine analyzes the image. Factors like rashes or holes in the parchment can affect the engine's ability to investigate the document.
2. The image resolution also plays a crucial role in the engine's capabilities. Working with poor-quality images can be thus a beneficial exercise for future transcriptions, as the machine may learn to recognize even vague shapes.
3. Difficult-to-read images are matched by inefficient transcriptions. This raised an important question regarding uncertainty in manuscript interpretation, such as abbreviations or unclear words: how to indicate uncertainty? We observed that it would be preferable for the machine to leave a blank space or use an apostrophe rather than using a question mark.
4. Finally, it is essential to note that verifying the machine's work is always beneficial. Sometimes, the machine's pixel scanning abilities can detect details that the human eye might miss.

Two existing models were then applied to the cartulary: one resulting from the transcription of some gospels in Latin, *latin-gospels-text-168-193*, which produced a degree of accuracy of 68.0%, and another derived from the model which was trained on Boccaccio manuscripts, which are much closer in date to the cartulary, *fr412_mod_best*, which produced a degree of accuracy of 70.9%.

- The *fr412_mod_best* model has delivered the best results so far, displaying both strengths and weaknesses.
- Challenges in distinguishing red ink titles: for those solely interested in text transcription rather than the document's material aspect, it is advisable to use black and white images in order to facilitate Kraken's job.
- Difficulty in differentiating similar letters, like "t," "c," "r" (Copyist 1 and 2) or the sequences "in" and "ni" (Copyist 1 and 2).
- Conversely, the machine excels in deciphering abbreviations and demonstrates a high proficiency in identifying a vast number of words

Work is still in progress on integrating the *cart_30p* model with the new transcriptions made and creating a basic segmentation model specifically for cartularies. If the prospects for using the latter seem good to me, the transcription model leaves me with some doubts. Those who are working on a corpus with several scripts must ask themselves whether it is appropriate to have only one more generic model to apply to all or the creation of different models more suitable for different scripts.

Although a cartulary appears as a simple codex, it is the result of a collaborative work unfolding over several decades and the documents it contains are copied by at least eleven copyists, not counting the various different hands who wrote the notes.

All of them belong to the sphere of gothic writing, but there are also very different types of scripts, including set, semi-set and cursive. My hypothesis, still to be verified, is that, since we are dealing with a single manuscript, it is more likely to achieve a satisfactory degree of accuracy in automatic transcription by using several different models, which can probably be reused on other 12th-14th century Parisian cartularies and maybe even outside the Parisian or even French area.

This particular work is still on the way, but to conclude we can say that automatic transcriptions can be over 90% accurate and as high as 98-99%.

Conclusions

The main focus was on utilizing eScriptorium to work on a 13th century cartulary: a document that is complex by its nature, in terms of form, content and scripts, and which opens up multiple reflections.

The starting hypothesis was to be able to deal with both the internal and external forms of the documents transcribed in the codex with the software eScriptorium, and also to find, if possible, a handwritten recognition model that would synthesise the various hands - three identified as praegothic, three as gothic documentary, four as gothic cursive and one as semi-cursive documentary - recognized in the document.

As we observed, the predictions about external forms were confirmed with the machine's ability to identify Region Types defined as valuable (Catchwords, Commentary, Cut in the parchment, Document's numbering, Foliotation, Illustration, Main, Maniculae, Old numbering, Other numbering, Title) after the training. Additionally, the automatic transcription achieved over 90% accuracy, likely due to the scripts all belonging to the gothic background.

Regarding internal forms, this aspect may see improvements when eScriptorium incorporates features like text and image annotation or NER Technology. So, going back to the original question: can eScriptorium be considered a valuable tool for Diplomatics? Is it beneficial to utilize it? In my view, the answer is indeed affirmative.

Presently, eScriptorium already represents a useful tool, as it is open-source and customizable, making it ideal for transcribing extensive corpora. When image and text annotation features aligned with Archetype principles and TEI are integrated, the software is expected to undergo significant advancement. This evolution will position eScriptorium not just as a tool for online editions of diverse documents in various scripts and materials but also as a means for in-depth Diplomatics analysis of internal and external forms. Archetype will enable detailed paleographic examinations, while TEI will facilitate text markup, making it an optimal tool for processing documentary resources. Moreover, if hopefully NER Technology will be integrated, eScriptorium will be able to identify and extract specific entities such as names of persons, places, dates, specific words and other relevant information from documents, giving the possibility to create structured data.

The combination of these enhancements, the software's open-source nature, the capability to work with Region Types, and its flexibility collectively establish eScriptorium as one of the top digital tools for Manuscript Studies and related disciplines. In comparison to other platforms utilizing AI for processing written resources, eScriptorium stands out as one of only two projects which received European funding (it has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement No. 871127 (RESILIENCE), and from the Initiatives de Recherches Interdisciplinaires et Stratégiques of Université PSL (Scripta-PSL))⁴² together with Transkribus⁴³, widely used in the academic and non-academic fields, which has been transformed into a commercial product, which result of research funded as part of the Recognition and Enrichment of Archival Documents (READ) project⁴⁴.

Talking about other platforms, a recent one is AMAP platform (Advanced Manuscript Analysis Portal) developed in 2019 by iXMan_Lab, Research Group Image Processing Department of Informatics at the University of Hamburg, which would allow the user to manipulate images via a visual programming language with compatible tools for pre-processing, segmentation and recognition⁴⁵ ⁴⁶. Currently, the official page is often unavailable.

Equally unavailable is MONK, a tool developed at the University of Groningen's Artificial Intelligence and Cognitive Engineering (ALICE) group which «consists of two major components: (1) a setup for the storage and web-based annotation of scanned page images and parts thereof; (2) a set of (handwriting and text) recognition algorithms as well as retrieval and search methods⁴⁷».

While we do not yet know which way AMAP will go, both Transkribus and MONK have become commercial products.

These platforms offer a wide range of features for handling manuscript documents. However, it's important to note that AI has been extensively utilized in the transcription field, often leading to the creation of commercial products such as

⁴² <https://cordis.europa.eu/project/id/871127>.

⁴³ <https://readcoop.eu/it/transkribus/>.

⁴⁴ This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 674943. This research was previously funded as part of the tranScriptorium project. This project received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement No. 600707. Major details here: <https://cordis.europa.eu/project/id/674943>.

⁴⁵ Vinodh Rajan, H. Siegfried Stiehl, "Making DIA Accessible to Non-Experts: Designing a Visual Programming Language for Document Image Analysis", in 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, IEEE, 2019, p. 23–27. [Online] [Consult. 18 February 2024] Available at <https://doi.org/10.1109/ICDARW.2019.20048>.

⁴⁶ Vera Isabell Schwarz-Ricci, "Il riconoscimento automatico di scrittura per documenti storici: rapporto tecnico" Istituto di Scienze del Patrimonio Culturale, Consiglio Nazionale delle Ricerche, Sede di Napoli. 2022 [Online] [Consult. 18 February 2024] Available at: https://openportal.ispc.cnr.it/data/2022/465029/2022_465029.published.pdf?id=people_____%3A%3A4b1167634bb7c9512338f055242e9fe8.

⁴⁷ <https://www.ai.rug.nl/~lambert/Monk-collections-english.html>.

Quartex by Adam Matthew Digital⁴⁸ and Aletheia⁴⁹ and a series of OCR engines such as Tesseract 4⁵⁰, anyOCR⁵¹, OCRopus⁵², Corpusbuilder⁵³, OCR4all⁵⁴, Tensorflow⁵⁵ developed by Google and Pytorch⁵⁶ developed by Facebook. These OCR engines, though developed by reputable tech companies, are somewhat limited in features and lack a user-friendly web interface.

Considering these factors, eScriptorium emerges as the most accessible platform available to date, as it is completely open-source and shows promise, making it suitable for application in Manuscript Studies and as a valuable tool for Diplomatics based on the aforementioned reasons.

References

ADAM MATTHEW DIGITAL, Quartex [Consult. 25 January 2023]. Available at <https://www.amdigital.co.uk/create/am-quartex>

Aletheia [Consult. 25 January 2023]. Available at <https://www.primaresearch.org/tools/Aletheia>

AMBROSIO, Antonella (2020). La Diplomatica e il digitale. Il Fondo della Biblioteca della Società Napoletana di Storia Patria online. RTH – Education & Philosophy. Riconoscimenti, trasformazioni e misteri. Culture a confronto attraverso secoli di Autor [Online]. Vol.7. [Consult. 15 September 2021]. Available at <https://doi.org/10.6093/2284-0184/7330>

ANSANI, Michele (1999). Diplomatica (e diplomatisti) nell'arena digitale. *Scrineum*, v. 1, p. 1-11. Available at <http://dohc.unipv.it/scrineum/biblioteca/ansani.htm>

ANYOCR [Consult. 25 January 2023]. Available at <https://anyline.com/free-ocr-font>

ARCHETYPE software [Consult. 15 September 2021]. Available at <https://github.com/kcl-ddh/digipal/wiki>

ARCHIVES NATIONALES DE FRANCE, AN LL\465 IIIF MANIFEST [Consult. 15 September 2021]. Available at <https://bvmm.irht.cnrs.fr/iiif/33551/manifest>

CORPUSBUILDER [Consult. 25 January 2023]. Available at <https://github.com/berkmancenter/corpusbuilder>

⁴⁸ <https://www.amdigital.co.uk/create/am-quartex>.

⁴⁹ <https://www.primaresearch.org/tools/Aletheia>.

⁵⁰ <https://github.com/tesseract-ocr/>.

⁵¹ <https://anyline.com/free-ocr-font>.

⁵² <https://github.com/ocropus/ocropy>.

⁵³ <https://github.com/berkmancenter/corpusbuilder>.

⁵⁴ <https://www.ocr4all.org/>.

⁵⁵ https://www.tensorflow.org/lite/examples/optical_character_recognition/overview.

⁵⁶ <https://github.com/courao/ocr.pytorch>.

DEROLEZ, Albert (2003) –The Palaeography of Gothic Manuscript Books: From the Twelfth to the Early Sixteenth Century. Cambridge: Cambridge University Press.

DURANTI, Luciana (2009). From Digital Diplomatics to Digital Records Forensics. *Archivaria*, v. 68, p. 39–66. Available at <https://archivaria.ca/index.php/archivaria/article/view/13229>

eSCRIPTA blog [Consult. 15 September 2021]. Available at <https://escripta.hypotheses.org/>

eSCRIPTORIUM software [Consult. 15 September 2021]. Available at <https://gitlab.com/scripta/escriptorium/>

eSCRIPTORIUM webpage [Consult. 15 September 2021]. Available at <https://escriptorium.fr/>

EUROPEAN COMMISSION (2015). “tranScriptorium”, Community Research and Development Information Service. [Consult. 25 January 2023]. Available at [available at: https://cordis.europa.eu/project/rcn/106843_en.html](https://cordis.europa.eu/project/rcn/106843_en.html)

EUROPEAN COMMISSION (2016). “Recognition and enrichment of archival documents”, Community Research and Development Information Service, [Consult. 25 January 2023]. Available at: https://cordis.europa.eu/project/rcn/198756_en.html

EUROPEAN COMMISSION (2019). “Religious Studies Infrastructure: tooLs, Experts, conNectiOns and Centers”, Community Research and Development Information Service, [Consult. 25 January 2023]. Available at <https://cordis.europa.eu/project/id/871127>

GÉHIN, Paul (2005) – Lire le manuscrit médiéval. Paris: Colin.

INSTITUT DE RECHERCHE EN HISTOIRE DES TEXTES, CartuIr [Consult. 15 September 2021]. Available at <https://telma-repertoires.irht.cnrs.fr/cartulr/page/presentation>

INSTITUT DE RECHERCHE EN HISTOIRE DES TEXTES, RegeCart webpage [Consult. 15 September 2021]. Available at <http://regecart.irht.cnrs.fr/>

INTERNATIONAL IMAGE INTEROPERABILITY FRAMEWORK [Consult. 15 September 2021]. Available at <https://iiif.io/>

InterPARES Project [Consult. 15 September 2021]. Available at <http://www.interpares.org/>

KIESSLING, Benjamin (2019). Kraken – A Universal Text Recognizer for the Humanities. DH 2019 [Online]. [Consult. 20 September 2021]. Available at <https://dev.clariah.nl/files/dh2019/boa/0673.html>

KIESSLING, Benjamin *et al.* (2021). The eScriptorium VRE for Manuscript Cultures. *Classic@Journal* [Online]. v.18, n. 1. [Consult. 17 September 2021] Available at: <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>

KRAKEN software [Consult. 15 September 2021]. Available at <http://kraken.re/>; resource code github.com/mittagessen/kraken

KÖLZER, Theo (2009). Diplomatik, Edition, Computer. *In*: VOGELER Georg. Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden, Archiv für Diplomatik – Beiheft 12; Köln-Weimar-Wien, p. 13-27.

MILLER, Matthew Thomas; ROMANOV, Maxim G.; BOWEN SAVANT Sarah (2018). Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans. *International Journal of Middle East Studies*, v. 50, n.1, p. 103–109. Available at <https://doi.org/10.1017/S0020743817000964>

MUEHLBERGER, Guenter *et al.* (2019). Transforming Scholarship in the Archives Through Handwritten Text Recognition: Transkribus as a Case Study. *Journal of Documentation*, v. 75, n.5, p. 954-976. Available at <https://doi.org/10.1108/JD-07-2018-0114>

OCR4all [Consult. 25 January 2023]. Available at <https://github.com/ocropus/ocropy>

OCROPUS [Consult. 25 January 2023]. Available at <https://www.ocr4all.org/>

PYTORCH [Consult. 25 January 2023]. Available at <https://github.com/courao/ocr.pytorch>

RAJAN, Vinodh; STIEHL, H. (2019). AMAP: A Visual Programming Language Based System to Support Document Image Analysis. *MuC'19: Proceedings of Mensch und Computer 2019*, p. 881-884. [Consult. 25 January 2023]. Available at <https://doi.org/10.1145/3340764.3345372>

RAJAN, Vinodh; STIEHL, H. (2019). Making DIA Accessible to Non-Experts: Designing a Visual Programming Language for Document Image Analysis, p. 23-27. [Consult. 25 January 2023]. Available at <https://doi.org/10.1109/ICDARW.2019.20048>

SCHWARZ-RICCI Vera Isabell (2022). “Il riconoscimento automatico di scrittura per documenti storici: rapporto tecnico”. Istituto di Scienze del Patrimonio Culturale, Consiglio Nazionale delle Ricerche, Sede di Napoli. [Consult. 25 January 2023] Available at: https://openportal.ispc.cnr.it/data/2022/465029/2022_465029.published.pdf?id=people_____%3A%3A4b1167634bb7c9512338f055242e9fe8

SCRIPTA PSL Project [Consult. 15 September 2021]. Available at <https://scripta.psl.eu/>

STOKES, Peter (2020). eScriptorium : un outil pour la transcription automatique des documents.” *ÉpheNum: Veille, agenda et actualités des humanités numériques à l'EPHE* [Online]. [Consult. 14 September 2021] Available at <https://ephenum.hypotheses.org/1412>

TENSORFLOW [Consult. 25 January 2023] Available at: https://www.tensorflow.org/lite/examples/optical_character_recognition/overview

TESSERACT 4.0. [Consult. 25 January 2023] Available at: <https://github.com/tesseract-ocr/tesseract>

TEXT ENCODING INITIATIVE [Consult. 15 September 2021]. Available at <https://tei-c.org/>

TRANSKRIBUS [Consult. 25 January 2023]. Available at <https://readcoop.eu/it/transkribus/>

VOCABULAIRE INTERNATIONAL DE LA DIPLOMATIQUE [Online] [Consult. 13 August 2021] Available at <https://www.cei.lmu.de/VID/>

VOGELER, Georg (2014). Digital Diplomatics: What Could the Computer Change in the Study of Medieval Documents? *Initial. A Review of Medieval Studies* [Online]. Vol. 2, p. 163-185. [Consult. 15 September 2021]. Available at: https://www.academia.edu/11893608/digital_diplomatics_what_could_the_computer_change_in_the_study_of_medieval_documents

VOGELER, Georg (2018). Digital Diplomatics: The Evolution of a European Tradition or a Generic Concept? *In*: CUBELIC Simon, ZOTTER Axel, MICHAELS Astrid. *Studies in Historical Documents from Nepal and India* [Online] Heidelberg: Heidelberg University Publishing. [Consult. 15 September 2021] p. 85–109. Available at <https://heiup.uni-heidelberg.de/reader/download/331/331-69-80416-1-10-20180223.pdf>

ZENODO'S "OCR/HTR model repository" [Consult. 15 September 2021]. Available at https://zenodo.org/communities/ocr_models/search?page=1&size=20

ÉCOLE FRANÇAISE DE ROME, Medieval European Cartularies [Consult. 15 September 2021]. Available at <https://www.efrome.it/en/meca>