Historical Documents and Automatic Text Recognition: Introduction

Ariane Pinche, Peter Anthony Stokes

Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

Historical Documents and Automatic Text Recognition: Introduction

Ariane Pinche[1] **and Peter Stokes**[2]

[1]CNRS, CIHAM (UMR 4856)
[2]EPHE-PSL, AOROC (UMR 8546)

Abstract

With this special issue of *the Journal of Data Mining and Digital Humanities* (JDMDH), we bring together in one single volume several experiments, projects and reflections related to automatic text recognition applied to historical documents.

More and more research projects[1] now include automatic text acquisition in their data processing chain, and this is true not only for projects focussed on Digital or Computational Humanities but increasingly also for those that are simply using existing digital tools as the means to an end. The increasing use of this technology has led to an automation of tasks that affects the role of the researcher in the textual production process. This new data-intensive practice makes it urgent to collect and harmonise the corpora necessary for the constitution of training sets, but also to make them available for exploitation. This special issue is therefore an opportunity to present articles combining philological and technical questions to make a scientific assessment of the use of automatic text recognition for ancient documents, its results, its contributions and the new practices induced by its use in the process of editing and exploring texts. We hope that practical aspects will be questioned on this occasion, while raising methodological challenges and its impact on research data.

The special issue on Automatic Text Recognition (ATR) is therefore dedicated to providing a comprehensive overview of the use of ATR in the humanities field, particularly concerning historical documents in the early 2020s. This issue presents a fusion of engineering and philological aspects, catering to both beginners and experienced users interested in launching projects with ATR. The collection encompasses a diverse array of approaches, covering topics such as data creation or collection for training generic models, reaching specific objectives, technical and HTR machine architecture, segmentation methods, and image processing.

Keywords

ATR; eScriptorium; Kraken; HTR-United; SegmOnto

I INTRODUCTION

Automatic Text Recognition (ATR)[2] plays a central role in humanities research, where the analysis of textual information lies at the heart of scholarly inquiry. From transcribing ancient manuscripts and historical documents to exploring literary works and diverse written materials, researchers often face the laborious and time-consuming task of manual transcription. However,

[1]See the Gallic(orpor)a (Sagot et al. [2022]) or AGODA projects (Bourgeois et al. [2022]).

[2]We use ATR throughout this contribution to what is most often called HTR (Handwritten Text Recognition) but also OCR (Optical Character Recognition). This is because the distinction between OCR and HTR is not well defined and can be based on the nature of the source material (printed for OCR, handwritten for HTR), or of the method applied (character segmentation for OCR, line segmentation for HTR), and in any case the distinction has little relevance in modern systems.

with the latest advancements in technology and user interfaces, ATR has improved remarkably to the point that it has now become accessible to humanities research projects, significantly reducing the manual work involved in text acquisition.

As a result, many research projects now incorporate automatic text acquisition into their data processing pipelines, leading to an automation of tasks that profoundly impacts the researcher's role in the textual production process. This technology is also invaluable to cultural heritage institutions actively digitizing vast collections of historical documents. Integrating ATR into this digitization process enriches digital images, making them searchable and increasing their value for researchers and indeed for a broader public. Nevertheless, the challenge of Handwritten Text Recognition (HTR) for historical documents persists, given the vast variety of handwriting styles across different historical periods, as well as problems for rare or under-resourced languages and scripts, difficulties in sharing training data and models, and so on. This burgeoning data-intensive practice has therefore prompted an urgent need to collect and harmonize essential corpora for constructing training sets and making them available for exploitation. Therefore, this special issue aims to present articles that intertwine philological and technical inquiries to offer a comprehensive assessment of the use of automatic text recognition for historical documents. By exploring the outcomes, contributions, and the new practices induced by the application of ATR to the process of editing and exploring texts, this collection seeks to shed light on the ever-evolving landscape of this technology in humanities research.

The genesis of this special issue was a conference held at the École Nationale des Chartes in Paris on June 23 and 24, 2022.[3] Scholars from diverse backgrounds came together to discuss their experiences and findings in using ATR in their research. The event delved into a wide array of topics, ranging from engineering and machine learning challenges to considerations of infrastructure. As the applications of ATR continue to multiply, and an increasing number of research projects and cultural heritage institutions express interest in its potential, this issue strives to provide an overarching view of its use on historical documents during this period.

II A BRIEF HISTORY OF ATR

The historical journey of ATR, the process of extracting textual content from images, spans over a century and has witnessed remarkable advancements.[4] Fournier d'Albe's pioneering work in 1914 on the Reading Optophone presented one of the earliest attempts at a document analysis system. It involved scanning text lines with spots of light, producing distinct musical note sequences for each letter of the alphabet (Fournier d'Albe [1914]). Although experimental, it laid the groundwork for future endeavors in document analysis. Subsequent breakthroughs came throughout the 20th century, such as Gustav Tauschek's "reading machine" (1935)[5], which proposed statistical analysis and pattern matching techniques that significantly improved the recognition capabilities and set the stage for Optical Character Recognition (OCR) systems. The 1970s and 1980s witnessed the emergence of more sophisticated OCR algorithms and document layout analysis techniques, augmenting the recognition capabilities of these systems and thereby facilitating the automatic acquisition of printed text from images (Rice et al. [1993]). The turn of the millennium ushered in a significant leap forward with the advent of machine learning

[3] All the presentations from that conference are available at the following link: https://www.canal-u.tv/chaines/enc/colloque-documents-anciens-et-reconnaissance-automatique-des-ecritures-manuscrites.

[4] For a more detailed presentation see the chapter "Document Image Analysis and Optical Character Recognition" in Kiessling [2021].

[5] Tauschek, Gustav, "Reading machine", U.S. Patent 2, 026, 329; patented 31 December 1935; filed 27 May 1929, in Austria 30 May 1928. Class 250-41. 5.

and neural network-based approaches, enabling OCR systems to achieve unprecedented levels of accuracy with the integration of deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

However, historical documents and handwritten text recognition are crucial applications of ATR that present unique challenges, and so they have inspired innovative solutions. The recognition of historical documents, with complexities such as their degraded ink and irregular layouts, can require specialized algorithms to handle such challenges. Advancements include adaptive thresholding techniques for image binarization and advanced image enhancement algorithms to improve text legibility. Progress in handwritten text recognition also involves addressing individual and varied handwriting styles, with machine learning algorithms trained on vast datasets enabling systems to transcribe handwriting with remarkable accuracy. Researchers such as Graves and Schmidhuber [2008], and Fischer [2010] have contributed significantly to advancing this field. During their early stages, these technologies were mostly experimental and required expertise in image processing and deep learning, and only projects with substantial funding and expertise could afford to use these advancements. However, recent developments in technology combined with new infrastructures and software have made these methods more and more accessible. Methods such as CRNN aim to reduce training data requirements and modern models can now achieve character error rates (CER) below 2% for manuscripts, indicating the effectiveness of these technologies (Hodel et al. [2021]).

The importance of ATR in the scientific community is also reflected in contributions to major conferences such as ICDAR (the International Conference on Document Analysis and Recognition) and HIP (Historical Document Imaging and Processing workshop) showcasing the latest developments. ATR is now also integrated into many text acquisition pipelines in the humanities[6]. The landscape of automatic text acquisition is evolving rapidly, opening up new possibilities for researchers to explore and analyze handwritten manuscripts and historical documents on an unprecedented scale.

III CONTEXT OF THIS ISSUE

From the early 2010s, the domain of humanities research has experienced a significant transformation with the emergence of ATR. Scholars, facing the need to transcribe and analyze extensive collections of historical documents, manuscripts, and texts, have sought innovative solutions to meet these challenges, and ATR has been one of these. In this context of expanding ATR adoption, groups in France and elsewhere have launched numerous initiatives to facilitate access for humanists and to optimise its application in research projects. Without claiming to be exhaustive, some of the tools that have contributed to this emergence include Tesseract (Smith [2007])[7], OCR-4all[8] and Ocropy (Breuel [2014])[9], which are Open tools operating on the command line, while Kraken operates on the same principles but is designed from the beginning to embed as few pre-assumptions about the writing-systems as possible, and therefore is intended to work with a very wide range of different scripts and writing-systems (Stokes et al. [2021]).

[6] See the number of references to the term in the program of events such as the *Digital Humanities* conference for 2022 and 2023, and meetings of the *Text Encoding Initiative* for 2022 and 2023.

[7] https://github.com/tesseract-ocr/tesseract, accessed March 15, 2024.

[8] "Introduction — OCR4all", https://www.ocr4all.org/guide/user-guide/introduction, accessed February 2, 2024.

[9] "GitHub - ocropus-archive/DUP-ocropy: Python-based tools for document analysis and OCR", https://github.com/ocropus-archive/DUP-ocropy, accessed February 2, 2024.

OCR-D[10] is a professional version of OCR-4all for libraries, capable of processing large collections. Transkribus was the first tool to offer an ergonomic user interface for applying ATR models, presented through a centralised platform as "software as a service" which now has a large community of users (Kahle et al. [2017]). Born in 2019, eScriptorium, which runs on the Kraken ATR engine, also offers a graphical user interface to apply ATR models on both prints and manuscripts. These advancements not only present innovative solutions to text recognition challenges but also pave the way for a more accessible and efficient exploration of historical texts, unlocking new horizons for humanities research.

The "Ancient Documents and Automatic Text Recognition" event in June 2022 was conceived in the context of the emergence of eScriptorium,[11] as well as the creation of the Consortium "Reconnaissance d'Écriture Manuscrite des Matériaux" and the CREMMALab project.[12] Consequently, this special issue is significantly shaped by this context. We therefore felt it necessary to introduce in more detail the initiatives and tools referenced in the articles within this issue, particularly eScriptorium, Kraken, CREMMA and CREMMALab.

3.1 eScriptorium and Kraken

Since 2020, ATR has played an increasingly important role in the humanities and social sciences in France, and has also benefited from significant funding.[13] Amongst others, the digital humanities team at the École Pratique des Hautes Études (EPHE) has played a pioneering role in advancing Automatic Text Recognition capabilities for humanities researchers. The core team comprises Daniel Stökl Ben Ezra, Peter A. Stokes, Benjamin Kiessling, and Robin Tissot, while important contributions have also come from others at the EPHE including Marc Bui, Colin Brisson and Elhassan Gargem. eScriptorium began within the Scripta project of Université Paris Sciences et Lettres (funded by Université PSL) based at the EPHE, and it has since received the support of numerous other sources of funding including Horizon2020 and Horizon Europe, the European Research Council, and the French Agence National de la Recherche (ANR). The team's combined expertise in technical and philological domains has led to the creation of two important tools: eScriptorium and Kraken. eScriptorium is free and open-source software to provide a collaborative platform for accessible and user-friendly ATR solutions for the academic community (Kiessling et al. [2019], Stokes et al. [2021]). This platform offers a graphical interface and API for ATR tasks, including the creation of training data, model training, and ATR predictions. As eScriptorium continues to evolve, it aspires to become a key component in platforms for digital edition production utilizing ATR as their first step.

Working in conjunction with Kraken, an open-source OCR engine developed by Benjamin Kiessling, eScriptorium enables humanists to access ATR and optimize its use. In particular, Kraken and eScriptorium excel in recognizing both handwriting and printed text and support a very wide range of scripts, including languages with right-to-left, bidirectional, and vertical writings, making it valuable for working with documents in a very wide variety of writing-systems. They have successfully been applied to writing-systems including Arabic, Hebrew, Syriac, historical printed Chinese and handwritten Japanese (both written top-to-bottom), and

[10] "OCR-D", https://ocr-d.de/en/, accessed February 2, 2024.

[11] This event was funded by the LabeX Hastec (EPHE), the École nationale des chartes and the CREMMALab project.

[12] Other similar initiatives exist, such as FONDUE, the University of Geneva's HTR infrastructure, see "FoNDUE - Une infrastructure HTR pour Genève - Humanités numériques - UNIGE", [online: https://www.unige.ch/lettres/humanites-numeriques/recherche/projets-de-la-chaire/fondue], accessed February 2, 2024.

[13] From the French national research agency (ANR) and Biblissima (ANR-21-ESRE-0005), EPHE-PSL.

Old Vietnamese inscriptions, amongst others. Notably, Kraken's output now includes bounding boxes at both line and character levels, providing finer-grained control over recognition results. Furthermore, both Kraken and eScriptorium have been developed with the goal of being fully free and open, meaning that not only is the software fully available but the trained models and data can all be exported, published and shared. Kraken's versatility and accessibility via the command line also allows it to be used alone or to be integrated into other software packages, providing a high level of customization and control over training processes (Kiessling [2019]).

The contributions of eScriptorium and Kraken, supported by the Scripta team's expertise, have significantly enriched the field of ATR, empowering humanities researchers on an unprecedented scale.

3.2 CREMMA and CREMMALab projects: infrastructure, datasets, and models for ATR

Founded by the Île-de-France region thanks to the DIM PAMIR,[14] the Consortium Reconnaissance d'Écriture Manuscrite des Matériaux Anciens (CREMMA) aims to provide a service that offers server resources, promoting enhanced access to handwriting text recognition. Partner laboratories and projects gain access to the open-source eScriptorium tool on dedicated servers, empowering them to train high-performance models. Additionally, the consortium's partners have taken on the responsibility of creating diverse datasets encompassing materials from various historical periods and languages, such as Latin and French. These datasets serve as the basis for initiating models in new projects and facilitating the utilization of ATR services. The project has also established transcription guidelines to ensure the consistency of the data produced (Chagué et al. [2022]).[15]

Building upon this initial initiative, the CREMMALab project addresses the continued challenges faced in the context of historical documents.[16] The fact is that even if the computing task of ATR may be considered solved from the point of view of informatics, significant difficulties remain when dealing with historical documents, particularly because of the wide variety of materials and scripts across time. Therefore, the CREMMALab project seeks to provide open training data and Handwritten Text Recognition (HTR) models specifically tailored to Western medieval manuscripts from the 12th to the 15th century. This endeavor aims to assist medievalists in accelerating the transcription phase of their research, while enabling access to vast corpora, such as universal histories or extensive chivalric romances in prose from the 13th century. Alongside this work, the process of learning algorithms for ATR has been examined in order to assess how the training corpus influenced the models' robustness and adaptability, thereby facilitating researchers in mastering ATR tools for their specific research needs and enabling them to train their own models.

In conclusion, the development of eScriptorium and Kraken marks a significant advancement in Automatic Text Recognition (ATR) for humanities researchers. These tools have provided accessible and user-friendly ATR solutions, empowering scholars to explore vast collections of historical documents. Moreover, the establishment CREMMA has provided the beginning of

[14] The DIM Patrimoines matériels – innovation, expérimentation et résilience (PAMIR) aims to support new research around questions on heritage collections and issues. It brings together an interdisciplinary network of scientists specializing in archaeology, paleontology, art history, history, archives, and heritage conservation-restoration, but also in natural sciences and data sciences. See further https://www.pamir.fr/en/the-dim/.

[15] https://gist.github.com/alix-tz/6f89444521bf1cab0522da520f7e4ff4.

[16] See in this special issue: Ariane Pinche, "Generic HTR Models for Medieval Manuscripts The CREMMALab Project", available at https://doi.org/10.46298/jdmdh.10252.

an infrastructure in France to improve access to ATR and an impulse to share training data and generic models, such as the CREMMALab project for medieval documents. These initiatives have also highlighted new challenges, such as making ATR accessible to more researchers, sharing data, and building new generic models.

IV NEW CHALLENGE OF ATR: SHARING DATA AND PRACTICES

As this discussion has shown, new challenges have emerged in the wake of the extensive adoption of ATR, prompting researchers to address crucial aspects of its implementation. As the demand for ATR continues to grow, the need to facilitate its use and support the initiation of projects incorporating ATR components becomes paramount. To tackle these challenges, pedagogical initiatives like the *Harmonising ATR* project,[17] led by Anne Baillot and Mareike Koenig and funded by DARIAH,[18] are coming to the fore. This project aims to provide a roadmap and tutorials to humanists, offering guidance on how to integrate ATR effectively into their research endeavors. In addition – and this is the focus of this section – the growing accumulation of data raises concerns about data management, research continuity and ATR model development. Researchers need to consider how to streamline data collection and build on existing work rather than starting ATR projects and models from scratch, in order to obtain efficiency and sustainable progress. To that end, several other projects have been central to numerous contributions in this special issue, and so the projects are further described here.

4.1 HTR-United

Continuing the legacy of the CREMMA project, T. Clérice and A. Chagué have initiated the ambitious HTR-United project (Chagué and Clérice).[19] Recognizing the pressing need for diverse ground truth datasets, HTR-United invites the user community to contribute and open up data from different ATR platforms, striving to minimize production costs and enabling by doing so the scientific community to train or fine-tune models even with small corpora.

Its primary objective is to enhance the discoverability of open datasets, encompassing a wide range of periods, scripts, and languages. HTR-United aims to create a public catalog of dataset descriptions, contributed by individuals or groups volunteering their own content (Chagué and Clérice [2023]). By doing so, it helps researchers to locate available training datasets for the creation of transcription or segmentation models. Beyond serving as a central catalog of available ground truth, HTR-United further helps to enable coherence in dataset descriptions by providing a schema for standardizing key information, such as the type of script, period, language, and tools used for data creation, because data produced with different tools are not always directly compatible and need to be adjusted to fit. Additionally, the initiative emphasizes the importance of providing transcription guidelines, ensuring data conformity and adherence to standardized rules. To facilitate collaboration and data management the project provides recommendations on how to organize a data repository, and it also includes tools for quality control and continuous documentation which are available locally or through Github Actions and continuous integration. This includes character control within the dataset, validation of ATR output verification of

[17] https://harmoniseatr.hypotheses.org/.

[18] "The Digital Research Infrastructure for the Arts and Humanities (DARIAH) aims to enhance and support digitally-enabled research and teaching across the arts and humanities. DARIAH is a network of people, expertise, information, knowledge, content, methods, tools and technologies from its member countries. It develops, maintains and operates an infrastructure in support of ICT-based research practices and sustains researchers in using them to build, analyse and interpret digital resources." See https://www.dariah.eu/about/dariah-in-nutshell.

[19] https://htr-united.github.io/

Page or ALTO XML, and detection of empty lines, all of which help to streamline the research process and enhance the reliability of the outcome.

By promoting data sharing, standardization, and collaboration, this initiative plays a major role in propelling the field of ATR by not only reducing production costs, but also speeding up the production of increasingly efficient models.

4.2 SegmOnto

Sharing large amounts of data also presents a significant challenge in maintaining coherence, since researchers and groups tend to have their own approaches, systems and vocabularies which are not necessarily compatible with others. To help address this issue in the case of segmentation, there is a pressing need to standardize document descriptions to ensure data compatibility. Recognizing this major problem, the SegmOnto initiative was launched as part of the CREMMA project, in collaboration with the University of Geneva and INRIA.

SegmOnto aims to design a controlled vocabulary for describing the layout of textual sources, focussing on the layout rather than the content (Gabay et al. [2023]). This approach aims to create a generic typology capable of accommodating a wide range of cases, rather than catering to specific needs. SegmOnto's controlled vocabulary operates under the assumption that most textual sources, be they historical prints or manuscripts, can be described in a similar manner when adopting a perspective focused on the page and its layout. However, numerous challenges arise, particularly concerning the desired level of granularity that each project aims to adopt. Accommodating diverse philological ambitions is essential, but the difficulty is not to let this impede the creation of common guidelines, and so one must ensure that they are designed with sufficient flexibility to cater to a broad array of situations. Harmonizing layout descriptions serves a dual purpose: first, it facilitates the mutualization of annotated data, enabling the training of more effective image segmentation models — an essential preliminary step for successful text recognition. Secondly, it supports the development of a shared post-processing workflow and pipeline, allowing the transformation of ALTO or PAGE files into formats that are more standard in the Digital Humanities such as TEI while preserving the critical link between the extracted information and the digital facsimile.

Aiming to strike a balance within this broad continuum of possibilities, SegmOnto endeavors to craft a generic rather than specific controlled vocabulary, emphasizing the layout over the content of textual regions to promote gathering and standardized segmentation datasets, but also for the creation of re-usable pipelines for textual acquisition.

4.3 Transcription Guidelines

The final step in data harmonization is transcription. Achieving compatibility between datasets requires a thoughtful approach to transcribing texts, as any text can always be transcribed in multiple ways. The CREMMA projects have therefore spearheaded initiatives and reflective discussions on establishing a more generic transcription method.

A seminar focused on French medieval manuscripts in 2022 resulted in the development and publication of transcription guidelines advocating for graphematic transcription,[20] preserving original punctuation, abbreviations, and spellings (Pinche [2022]). The aim was to support the creation of training data and to optimize the machine learning process for HTR models. The challenge lies in finding a way to translate the original text on its original medium into a format that a machine can interpret and learn from. The solutions provided are inherently reductive

[20]For the graphematic approach see Stutzmann [2011].

and interpretative, as attempting to render the full variety of handwriting with a limited number of characters is impossible for a computer. It is also essential to note that these proposed transcription methods are not meant to build a definitive or final edition, but rather present pragmatic approaches that try to stay close to the source material. For instance, preserving abbreviations enhances the models' adaptability to different genres, periods, linguistic variations, or languages, and this in turn increases the usefulness of models and reduces the need for retraining and producing new datasets. All CREMMA guidelines adhere to those general rules, with slight specificities introduced for different periods or languages:

- Modern French manuscripts (Chagué and Clérice [2022]);
- French Medieval manuscripts (Pinche [2022]);
- Latin Medieval manuscripts (Clérice et al. [2023]).

V CONTENT OF THE ISSUE

As noted above, this special issue on ATR is dedicated to providing a comprehensive overview of the use of this technology in the humanities, and especially the state of the art in the early 2020s when applied to historical documents. This issue presents a fusion of engineering and philological aspects, catering to both beginners and experienced users interested in launching projects with ATR. The collection encompasses a diverse array of approaches, covering topics such data creation or collection for training generic models and reaching specific objectives technical and HTR machine architecture, segmentation methods, image processing.The content of the issue is structured into two main axes: research project, corpus, or model building, and technical improvements in ATR.

5.1 ATR and Research Projects, Corpus, and Model Building

This section features presentations of research projects and articles that delve into datasets and models building, whether they are general on the challenges to be met in the creation of the corpus or more targeted on a particularity about dataset building and its exploitation to train the most efficient model in the context of the objectives of an object.

a. Research Projects. The first subgroup features includes contributions from various projects, each presenting their methodology.

- COUTURE, Béatrice, VERRET, Farah, GOHIER, Maxime [et al.], "The challenges of HTR model training: Feedbacks from the project Donner le goût de l'archive à l'ère numérique", https://jdmdh.episciences.org/12556;
- CALVELLI, Lorenzo, BOSCHETTI, Federico and TOMMASI, Tatiana, "EpiSearch. Identifying Ancient Inscriptions in Epigraphic Manuscripts", https://doi.org/10.46298/jdmdh.10417;
- ROMEIN, C. Annemieke, HODEL, Tobias, GORDIJN, Femke, [et al.], "Exploring Data Provenance in Handwritten Text Recognition Infrastructure: Sharing and Reusing Ground Truth Data, Referencing Models, and Acknowledging Contributions. Starting the Conversation on How We Could Get It Done", https://doi.org/10.46298/jdmdh.10403;
- PERDIKI, Elpida, "How to (Auto) Collate Big Manuscript Data with Minimal HTR Training", https://doi.org/10.46298/jdmdh.10419.

b. Corpus and Model Building. The second subgroup centers on articles that elaborate on experiments in model training, emphasizing specific corpus objectives or challenges.

- GILLE LEVENSON, Matthias, "Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR)", https://doi.org/10.46298/

jdmdh.10416;

- PINCHE, Ariane, "Generic HTR Models for Medieval Manuscripts: The CREMMALab Project", https://doi.org/10.46298/jdmdh.10252.

5.2 ATR, Technical Improvement, and Tools: Image Enhancement, Segmentation, ATR Engine Architecture, etc.

This technical section focuses on advances in ATR engine technology.

a. Improvement of Segmentation and ATR Engine. The first subsection deals with segmentation issues and improvements :

- AGUILAR, Sergio Torres and JOLIVET, Vincent, "Handwritten Text Recognition for Documentary Medieval Manuscripts", https://doi.org/10.46298/jdmdh.10484;
- CLÉRICE, Thibault, "You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine", , https://doi.org/10.46298/jdmdh.9806.

b. Source Pre-treatment and ATR Improvement. The second subsection covers the preprocessing of source images and advances in ATR enhancement.

- JACSONT, Pauline and LEBLANC, Elina, "Impact of Image Enhancement Methods on HTR Trainings with eScriptorium", https://doi.org/10.46298/jdmdh.10262;
- WEST, Graham, SWINDALL, Matthew I., KEENER, Ben, [et al.], "An Approach for Noisy, Crowdsourced Datasets Utilizing Ensemble Modeling, 'Human Softmax' Distributions, and Entropic Measures of Uncertainty", https://doi.org/10.46298/jdmdh.10297.

Overall, this special issue aims to foster the increased utilization of ATR in the humanities field, promoting a better understanding of machine learning and the importance of generating sustainable data to be shared within the scientific community. We hope that the diverse topics and methodologies presented in this issue will contribute to empowering scholars in their endeavors to explore and analyze historical texts with greater efficiency.[21]

5.3 List of the Datasets and Models Cited in the Issue

- Boschetti F., episearch-htr. Published online November 23, 2022. https://github.com/vedph/episearch-htr
- Clérice T. YALTAi: Segmonto Manuscript and Early Printed Book Dataset. Published online July 10, 2022. https://doi.org/10.5281/zenodo.6814770
- Hodel T, Schoch D, Dängeli P. Handwritten Text Recognition Ground Truth Set: StABS Ratsbücher O10, Urfehdenbuch X. Published online August 2, 2021. https://doi/org/10.5281/zenodo.5153263
- Jacsont P. Toponomasia: edition of cod. 174 of Bern Burgerbibliothek. Published online July 26, 2022. https://doi.org/10.5281/zenodo.7026585
- Levenson MG. Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR). Datasets and scripts. Published online December 1, 2022. https://doi.org/10.5281/zenodo.7389195
- Perdiki E. List of manuscripts containing John Chrysostom's Homilies and the relevant manual transcriptions. Published online February 27, 2023. https://doi.org/10.5281/zenodo.7681133

[21] This article was written primarily by Ariane Pinche, although it represents the combined view and work of both authors on this Special Issue of JDMDH, and both share responsibility for its content.

- Pinche A, Gabay S, Leroy N, Christensen K. Données HTR incunables du 15e siècle. Published online March 22, 2023. https://github.com/Gallicorpora/HTR-incunable-
- Pinche A, Gabay S, Leroy N, Christensen K. Données HTR manuscrits du 15e siècle. Published online March 22, 2023. https://github.com/Gallicorpora/HTR-MSS-15e-Si
- Pinche A. Cremma Medieval. Published online June 2022. https://github.com/HTR-United/cremma-medieval
- Torres Aguilar S, Jolivet V. Dataset and evaluation for HTR models for Latin and French Medieval Documentary Manuscripts. Published online January 10, 2023. https://doi.org/10.5281/zenodo.7401833
- Torres Aguilar S, Jolivet V. HTR model for Latin and French Medieval Documentary Manuscripts (12th-15th). Published online January 18, 2023. https://doi.org/10.5281/zenodo.7547438

References

Nicolas Bourgeois, Fanny Lebreton, Aurélien Pellet, Marie Puren, and Pierre Vernus. Le projet AGODA. Annoter et publier les débats parlementaires français de la fin du XIX e siècle : défis et solutions. In *Présentation des projets AGODA et Gallicorpora, Bibliothèque nationale de France*, Paris, France, June 2022. URL https://hal.science/hal-03762957.

Thomas M. Breuel. Ocropy: Python-based tools for document analysis and OCR, 2014. URL https://github.com/tmbdev/ocropy.

Alix Chagué and Thibault Clérice. HTR-United: Ground Truth Resources for the HTR and OCR of patrimonial documents.

Alix Chagué and Thibault Clérice. Règles générales de transcription pour les corpus cremma. 2022. URL https://gist.github.com/alix-tz/6f89444521bf1cab0522da520f7e4ff4.

Alix Chagué and Thibault Clérice. "I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data. In *Digital Humanities 2023: Collaboration as Opportunity*, Graz, Austria, July 2023. Alliance of Digital Humanities Organizations and University of Graz. URL https://inria.hal.science/hal-04094233.

Alix Chagué, Thibault Clérice, and CREMMA. Règles générales de transcription pour les corpus CREMMA, September 2022. URL https://gist.github.com/alix-tz/6f89444521bf1cab0522da520f7e4ff4.

Thibault Clérice, Malamatenia Vlachou-Efstathiou, and Alix Chagué. CREMMA Medii Aevi: Literary manuscript text recognition in Latin. *Journal of Open Humanities Data*, 9:4, April 2023. doi: 10.5334/johd.97. URL https://enc.hal.science/hal-03828353.

Andreas Fischer, Emanuel Indermühle, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. Ground truth creation for handwriting recognition in historical documents. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, pages 3–10, New York, NY, USA, June 2010. Association for Computing Machinery. ISBN 978-1-60558-773-8. doi: 10.1145/1815330.1815331. URL https://doi.org/10.1145/1815330.1815331.

E. E. Fournier d'Albe. On a Type-Reading Optophone. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 90(619):373–375, 1914. URL https://www.jstor.org/stable/93525.

Simon Gabay, Ariane Pinche, Kelly Christensen, and Jean-Baptiste Camps. SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles. December 2023. URL https://hal.science/hal-04343404.

Alex Graves and Jürgen Schmidhuber. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://papers.nips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html.

Tobias Hodel, David Schoch, Christa Schneider, and Jake Purcell. General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, 7(0):13, July 2021. doi: 10.5334/johd.46. URL http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.46/.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference*

on Document Analysis and Recognition (ICDAR), volume 04, pages 19–24, November 2017. doi: 10.1109/ICDAR.2017.307.

B. Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–24, September 2019. doi: 10.1109/ICDARW.2019.10032.

Benjamin Kiessling. Kraken - an Universal Text Recognizer for the Humanities. In *Digital Humanities 2019 Book of Abstracts*, Utrecht, July 2019. CLARIAH. URL https://dev.clariah.nl/files/dh2019/boa/0673.html.

Benjamin Kiessling. *Avancées en Reconnaissance Optique des Caractères pour les Documents Arabes Historiques*. Doctoral dissertation, Université Paris sciences et lettres, April 2021. URL https://www.theses.fr/2021UPSLP023.

Ariane Pinche. Guide de transcription pour les manuscrits du Xe au XVe siècle. June 2022. URL https://hal.archives-ouvertes.fr/hal-03697382.

Stephen Rice, Junichi Kanai, and Thomas Nartker. An evaluation of OCR accuracy. *Information Science Research Institute*, 9:20, 1993.

Benoît Sagot, Laurent Romary, Rachel Badwen, Pedro Ortiz Suárez, Jean-Baptiste Camps, Simon Gabay, and Ariane Pinche. Gallic(orpor)a: extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue, September 2022. URL https://gallicorpora.github.io/. original-date: 2022-09-12T11:10:29Z.

Ray Smith. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2822-8. URL https://storage.googleapis.com/pub-tools-public-publication-data/pdf/33418.pdf.

Peter A. Stokes, Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and Gargem El Hassane. The eScriptorium VRE for Manuscript Cultures. *Classics Journal*, 2021. URL https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/.

Dominique Stutzmann. Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? In *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2*, pages 247–277. Books on Demand, 2011. URL https://halshs.archives-ouvertes.fr/halshs-00596970.