Ancient documents and automatic recognition of handwriting
Conference on HTR, June 23 and 24, 2022, École nationale des chartes, Paris

June 23th	**Talk**	**speaker(s)**
9:15-9:30	Welcome of the participants	
9:30-9:45	**Opening speech with the presentation of the CREMMA and CREMMALAB projects**	Elsa Marguin-Hamon, d*irectrice de la recherche et des relations internationales, École nationale des chartes*
9:45-10:15	**CremmaLab projects: Transcription guidelines and HTR models for French medieval manuscripts**	Jean-Baptiste Camps, *maître de conférence, École nationale des chartes, CJM* Ariane Pinche, *post-doctorante, École nationale des chartes, CJM*
	Summary : L'étape d'acquisition du texte est première dans la plupart de nos entreprises de recherche, qu'il s'agisse d'édition de texte, d'études linguistiques, philologiques et historiques, ou de traitement massif de corpus. Pour produire des corpus textuels de qualité, il est crucial de pouvoir partager librement, en en garantissant l'interopérabilité, tant les données que nous produisons, et, in fine, de proposer à la communauté scientifique des modèles réutilisables pour qu'ils puissent être utilisés par d'autres projets. Pour ce faire, nous avons besoin de protocoles permettant la création de corpus d'entraînement et leur partage. Pour répondre à ces besoins, et plus spécifiquement aux besoins des médiévistes qui travaillent sur des sources manuscrites, le projet CREMMALAB propose des réflexions méthodologiques sur les protocoles de transcriptions des corpus afin d'optimiser des modèles d'HTR à travers la rédaction d'un guide de transcription et la mise à disposition de modèles d'HTR pour les manuscrits médiévaux. Nous présenterons les premiers résultats de ces travaux à travers le traitement de deux corpus massifs en particulier: un corpus de romans de chevalerie et un corpus de textes hagiographiques, pris en diachronie (XIIIe-XVe siècle)	
10:15-10:45	**HTR fine-tuning for medieval manuscripts models: strategies and evaluation**	Sergio Torres Aguilar, *post-doctorant, École nationale des chartes, CJM* Vincent, Jolivet, *responsable de la mission projets numériques, École nationale des chartes*
	Summary : In this presentation we intend to explore different practical questions about HTR modeling in order to determine at what point a model reaches the necessary robustness and a sufficiently broad-level of generalization to serve as a pre-trained base to raise a	

	new specialized model. For this end we use several HTR ground-truth documents from medieval cartularies and registers ranging from 12th to 15th centuries and we will evaluate two aspects: (1) the creation of robust models by trying to calculate the learning break-point and the minimum amount of ground truth necessary to achieve good generalization performances from a limited collection of documents and (2) the process of fine-tuning in the aim to quickly specialize a robust model, used here as a pre-trained base, on a type of source other than those used during training.	
10:45-11:15	Pause	
11:15-11:45	**Une cursive du 17e siècle**	Élodie Paupe, *assistante doctorante, université de Neuchâtel et chargée de projet pour les AAEB*
	Summary : Le projet « Crimes et châtiments » a pour objectif la numérisation et la transcription des procédures criminelles de l'ancien Évêché de Bâle (1461-1797). Dans le cadre de la phase pilote en cours de réalisation du projet, un modèle HTR est développé sur une série de procès de sorcellerie dont la majorité des documents sont écrits en cursive française par le prévôt Henri Farine, actif entre 1580 et 1618. Après avoir présenté les particularités de cette main et le corpus, un retour d'expérience sera donné autour des deux infrastructures utilisées (Transkribus et eScriptorium) et du recours aux méthodes de binarisation sur des documents manuscrits. Pour conclure, l'efficience du modèle « Farine » sur des documents contemporains d'autres mains sera présentée, ainsi que les pistes de développement poursuivies.	
11:45-12:15	**Un modèle ouvert pour la reconnaissance automatique des manuscrits du théâtre espagnol du Siècle d'Or**	Cuéllar Álvaro, *PhD Student, University of Kentucky*
	Summary : Le projet ETSO. *Estilometría aplicada al Teatro del Siglo de Oro* (Cuéllar et Vega García-Luengos 2017-2022) (https://etso.es/) se propose de collecter et d'analyser à travers des techniques stylométriques le plus grand nombre de pièces de théâtre espagnol du Siècle d'Or. Un nombre important de ces textes ne se retrouvent que dans des témoignages manuscrits, pour lesquels il a fallu entreprendre un processus de transcription automatique à l'aide de Transkribus. L'entraînement du modèle « Spanish Golden Age Manuscripts (Spelling Modernization) 1.0 » a nécessité 3.250.116 mots et il est capable de moderniser automatiquement le texte, obtenant un *Character Error Rate* (CER) de 10,54% dans le validation set. Grâce à ce modèle nous avons pu transcrire quelque 400 manuscrits de pièces du Siècle d'Or. Parmi tous les textes, un a retenu l'attention : La	

	francesa Laura. Cette pièce de théâtre anonyme a été alignée stylométriquement avec l'ensemble du corpus du dramaturge Lope de Vega (1562-1635).	
12:15-14:00	Lunch	
14:00-14:30	**eScriptorium Project**	Benjamin Kiessling, *ingénieur de recherche, PSL* Peter Stokes, *directeur d'étude, EPHE*
	Summary :	
14:30-15:00	**De Transkribus à eScriptorium : retour(s) d'expérience sur l'usage d'outils d'HTR appliqués à un corpus d'imprimés espagnols du XIX^{e} siècle**	Élina Leblanc, *post-doctorante, unité d'espagnol, faculté des lettres, université de Genève* Pauline Jacsont, *collaboratrice scientifique, unité d'espagnol, Faculté des lettres, université de Genève*
	Summary : Dans cette communication, nous présenterons la chaîne éditoriale mise au point pour le projet *Démêler le cordel*, en vue d'élaborer une bibliothèque numérique dédiée à la collection d'imprimés éphémères espagnols du XIX^{e} siècle de la Bibliothèque universitaire de Genève. Notre chaîne éditoriale a pour particularité d'avoir eu recours à deux outils d'HTR, *Transkribus* et *eScriptorium,* dont nous proposerons une analyse en termes d'usages à différentes étapes d'un projet. Dans un premier temps, nous décrirons la collection d'imprimés, en insistant sur ses spécificités et ses enjeux dans un contexte de transcription automatique. Puis, nous reviendrons sur notre expérience avec chacun des outils d'HTR employés, sur les raisons qui nous ont conduites à passer de l'un à l'autre et sur les difficultés rencontrées. Pour conclure, nous présenterons l'exploitation des prédictions HTR sur notre site web, développé avec TEI-Publisher.	
15:00-15:30	**Lettres en lumières**	Florian Fizaine, *doctorant, archives départementales de la Côte-d'Or* Édouard Bouyé, *directeur des archives départementales de la Côte-d'Or*
	Summary : Dans le cadre du projet « Lettres en lumières » mené par les Archives départementales de la Côte-d'Or en partenariat avec le Laboratoire d'étude de l'apprentissage et du développement (LEAD, Université de Bourgogne), nous développons un outil de HTR en utilisant Mask RCNN, un algorithme de segmentation d'instance utilisé notamment dans le médical, pour la segmentation des lignes et les réseaux transformer qui ont	

	largement montré leur efficacité dans la compréhension du langage naturel, pour la transcription. Nous avons commencé ce travail sur les registres des états de bourgogne du XVIII^{e} siècle, ces données d'entrainements sont obtenues grâce à la participation de transcripteur bénévoles.	
15:30-16:00	Pause	
16:00-16:30	**Les archives inquisitoriales (Portugal) sous HTR : le projet TraPrInq (Transcribing the court records of the Portuguese Inquisition, 1536-1821)**	Baudry Hervé, *chercheur au CHAM-Centro de Humanidades (Universidade Nova de Lisboa). Responsable du projet TraPrInq.*
	Summary : Le projet TraPrInq a pour objectif de créer un modèle d'HTR. Une partie des archives inquisitoriales portugaises (Arquivo Nacional da Torre do Tombo, Tribunal do Santo Ofício, 1536-1821) est constituée de procès, au nombre de plus de 40 000. Près de la moitié de ce sous-fonds a été numérisée. Le modèle générique en cours d'élaboration sur la plateforme Transkribus par une équipe d'une dizaine de paléographes permettra la transcription à grande échelle des documents. La présente communication établit en premier lieu un état d'avancement des travaux à l'issue des cinq premiers mois d'activité : particularité du corpus, mode de travail, obstacles rencontrés et solutions adoptées, premiers résultats (données d'entraînement). En outre, comme il semble prématuré de dresser un bilan général, elle s'attache à décrire la démarche adoptée, ses évolutions, ainsi qu'à réfléchir sur les aspects techniques et humains des moyens mis en œuvre et des objectifs à atteindre.	
16:30-17:00	**Segmentation Mode for Archival Documents with Highly Complex Layout**	Stökl Ben Ezra Daniel, *directeur d'étude, EPHE* Rustow Marina, *pr*ofessor, *Princeton University* Witty Devorah, *software developper, The Research software compagny*
	Summary :	
17:00-17:30	**SegmOnto - A Controlled Vocabulary to Describe Historical Textual Sources**	Ariane Pinche, *post-doctorante, École nationale des chartes, CJM* Simon Gabay, *maître-assistant, université de Genève*
	Summary : Our initiative aims to design a controlled vocabulary for the description of the layout of textual sources: *SegmOnto*. Following a codicological approach rather than a semantic one, it is designed as a generic typology, coping with a maximised number of cases rather than answering specific needs. Systematise the layout description has a double objective: on the one	

	hand it facilitates the exchange of annotated data and therefore the training of better models for image segmentation (a crucial preliminary step for text recognition), on the other hand it allows the development of a shared post-processing workflow and pipeline for the transformation of ALTO or PAGE files into DH standard formats such as RDF or TEI	
17:30-17:45	Conclusion of the day	Pinche Ariane, *post-doctorante, École nationale des chartes, CJM*

June 24th	**Talk**	**speaker(s)**
9:15-9:30	Welcome of the participants	
9:30-10:00	**FoNDUE - A Lightweight HTR Infrastructure for Geneva**	Simon Gabay, *maître-assistant, université de Genève*
	Summary : Recognising text on an image is becoming increasingly important for scholars working with textual sources. Because institutions have to address the needs ot their members, the University of Geneva has decided to offer a free of charge and user-friendly solution based on *eScriptorium*. The specificity of our instance is that it relies only on local infrastructures to minimise its cost and offer additional services, such as training models directly with command lines. Therefore, it promotes a double empowerment: the one of the institution, that does not depend on external private solutions, but also the one of scholars, who gain new digital skills. On top of a theoretical reflexion on this empowerment, we propose a first feedback on how to deploy an efficient HPC-based instance of *eScriptorium*.	
10:00-10:30	**From HTR to Critical Edition: A Semi-Automatic Pipeline**	Stoekl Ben Ezra Daniel, *directeur d'étude, EPHE* Hayim Lapin, p*rofessor, University of Maryland, College Park* Bronson Brown-Devost, *post-doctoral researcher, Scripta Qumranica Electronica* Pawel Jablonski, *PhD student, EPHE*
	Summary :	
10:30-11:00	Pause	
11:00-11:30	**Analyse, Reconnaissance et Indexation des manuscrits CHAM**	Anne-Valérie Schweyer, c*hercheuse CNRS, Centre Asie du Sud-Est (CASE-EHESS-INALCO)*, Jean-Christophe Burie, *professeur des universités, Université de La*

		Rochelle Tien Nam Nguyen, *doctorant, Université de La Rochelle*
	Summary :	
11:30-12:00	**Expérimentations pour l'analyse automatique de sources chinoises anciennes**	Bizais-Lillig Marie, *maître de conférences, université de Strasbourg,* Vidal-Gorène Chahan, *doctorant, École nationale des Chartes et EPHE.*
	Summary :	
12:00-14:00	Lunch	
14:00-14:30	**Sharing HTR dataset with standardized metadata: the HTR-United initiative**	Chagué Alix, d*octorante, EPHE, Université de Montréal, Inria* Clérice Thibault, r*esponsable pédagogique du master TNAH, École nationale des chartes, CJM*
	Summary : Since some scholars adopted Ocropy in the mid-2010s, production of HTR or OCR ground truth has seen an impressive and steady growth. However, few projects share their gold dataset, and when they do, they are scattered across many different hosting options (Github, zenodo, gitlab, institutional repository, etc.) making them very hard to find. For reuse, when they are "discovered", their description is often lacking crucial details. The HTR-United initiative is an answer to this problem: with a standardized metadata schema, a curated catalogue and tools focusing on helping them through every step, owners can now easily publish and make their dataset findable.	
14:30-15:00	**EpiSearch. Recognising Ancient Inscriptions in Epigraphic Manuscripts**	Boschetti Frederico, *researcher; Institute for Computational Linguistics "A. Zampolli" – CNR, Pisa / VeDPH, Ca' Foscari University of Venice* Tommasi Tatiana, *MA student; Ca' Foscari University of Venice*
	Summary : The project focuses on epigraphic codices as a proof of concept for putting digital tools at the test, thus defining new ways for the integration of large epigraphic collections. As a sample, we use the epigraphic manuscript composed by the Venetian abbot Giovanni Antonio Astori (1672-1743) and preserved in the Marciana National Library in Venice: Marc. lat. XIV, 200 (4336). In the first part of our talk, we analyse the life of the author and the characteristics of his manuscript. In the	

	second part, we focus on the following tasks: a) evaluating the accuracy of eScriptorium on epigraphic manuscripts with training sets of different size, in order to estimate the best trade-off between the human effort to prepare the training sets and the human effort to correct the results; b) mapping legacy manual transcriptions on the manuscript facsimile; c) improving the layout analysis for epigraphic manuscripts.	
15:00-15:30	**HTR of Handwritten Paleographic Greek Text as a Function of Chronology**	Platanou Paraskevi, *postgraduate student, Athens University of Economics and Business*
	Summary : Today classicists are provided with a great number of digital tools which, in turn, offer possibilities for further study and new research goals. In this paper we explore the idea that old Greek handwriting can be machine-readable and consequently, researchers can study the target material fast and efficiently. The overall aim of this paper is to assess HTR for old Greek manuscripts. To address this statement, we study and use images of the Oxford University Bodleian Library Greek manuscripts. By manually transcribing images, we have created and present here a new dataset for Handwritten Paleographic Greek Text Recognition. The dataset instances have been organized by establishing as a leading factor the century to which the manuscript and hence the image belongs. In this way, the HTR performance can reveal century-specific challenges when it comes to Handwritten Paleographic Greek Text Recognition.	
15:30-16:00	Pause	
16:00-16:30	**Reconnaissance et extraction d'informations dans des tableaux manuscrits historiques : vers une compréhension des recensements de Paris de l'entre-deux guerre**	Constum Thomas, *doctorant, LITIS EA4108, université Rouen Normandie*
	Summary : Le projet POPP, *Projet d'Océrisation des Recensements de la Population Parisienne* (S. Brée et al, 2022) vise à constituer une vaste base de données à partir des recensements nominatifs de Paris de l'entre-deux guerres, composés chacun d'environ 100 000 pages simples manuscrites sous forme de tableaux. Nous avons à ce jour traité les recensements de 1926, 1931, et 1936, ce qui représente un total d'environ 9 millions d'individus. Ce corpus est une source d'information primordiale pour les historiens, les démographes, les économistes ou les sociologues. L'objectif de notre communication est de décrire un système complet pour l'extraction d'informations de recensements historiques de la population. *POPP* est un	

	projet qui a réuni des chercheurs en vision par ordinateur, en reconnaissance de formes et en démographie historique.	
16:30-17:00	**Retour d'expériences sur l'utilisation comparée de plusieurs de dispositifs de transcription numérique d'archives de fouilles archéologiques**	Tufféry Christophe, *ingénieur de recherche à l'Institut national de recherches archéologiques préventives, doctorant à CY Cergy Paris Université, en partenariat avec l'Institut national du patrimoine.*
	Summary : Dans le cadre d'une thèse de doctorat engagée depuis 2019, nous proposons une étude historiographique et épistémologique des effets du numérique sur l'archéologie et sur les archéologues sur les cinquante dernières années, une période pendant laquelle l'archéologie a vu ses méthodes modifiées par l'introduction progressive de la micro-informatique dès le terrain . Cette recherche s'appuie sur notre expérience comme archéologue depuis la fin des années 1970 et sur notre activité à l'Inrap depuis 2010. Nous avons exploité plusieurs archives de chantiers de fouilles dont celles d'un chantier sur lequel nous avons été fouilleur bénévole entre 1980 et 1988. Nous avons procédé à la numérisation de deux cahiers de fouille. Nous avons ensuite procédé à leur transcription numérique avec trois solutions techniques différentes et complémentaires, dont eScriptorium, qui présentent des avantages et des limites techniques et méthodologiques. Nous avons pu ensuite exploiter les résultats de la transcription avec diverses méthodes et outils numériques.	
17:00-17:15	Conclusion of the day	