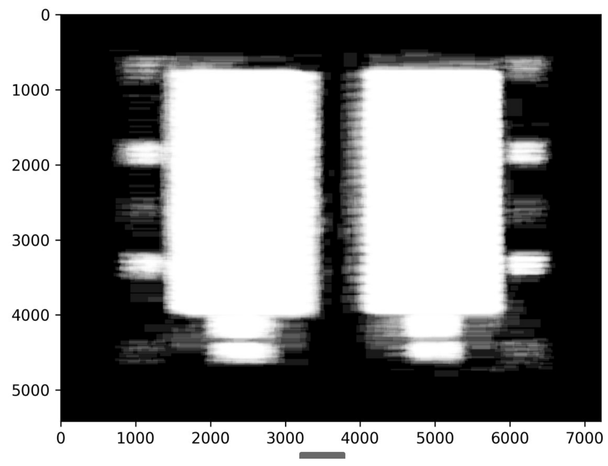
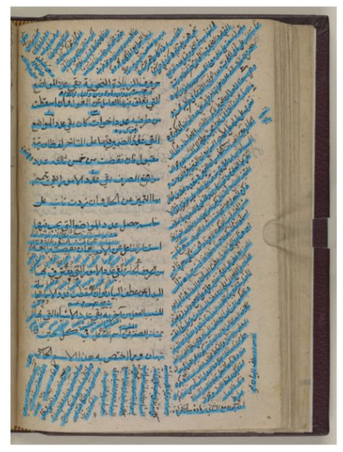
Biblissima+ Cluster 3 - Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites

Dominique Stutzmann, Daniel Stökl Ben Ezra



Biblissima+

Observatoire des cultures écrites
de l'argile à l'imprimé



Cluster 3 – Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites

Domaines d'action

Recherches sur les textes français et latins (IRHT)

- Classification des éléments graphiques (pages, zones de pages)
  - Classification iconographique des miniatures
- Reconnaissance d'entités nommées (cote, noms de personnes, titres d'œuvres)
  - Alignement des entités nommées sur des référentiels (*linking*)
- Catalogage automatique des manuscrits numérisés : identification des textes issus de HTR par comparaison avec référentiels textuels (Corpus corporum etc.)
- Référentiels de noms de lieux et de personnes dans les cartulaires médiévaux

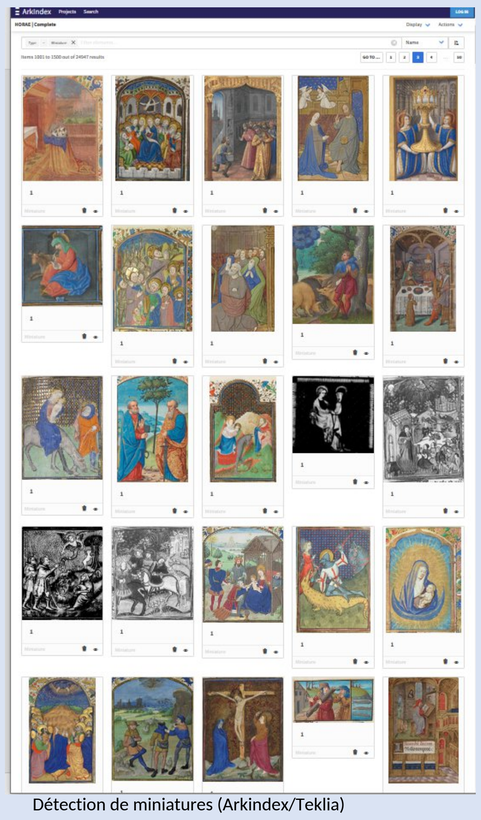
Reconnaissance d'écriture Kraken (AOrOc)

- Module HTR open source pour tous les systèmes d'écriture
  - Latin, grec, arabe, chinois, hébreu, syriaque, géorgien, geez
  - Ordre de lecture entrainable
- Développement et maintenance d'eScriptorium
  - nouvelle UI (openITI / Mellon)
  - alignement text2text avec passim, annotation de texte et d'image
  - extension de l'API, moteur de recherche, métadonnées au niveau d'image

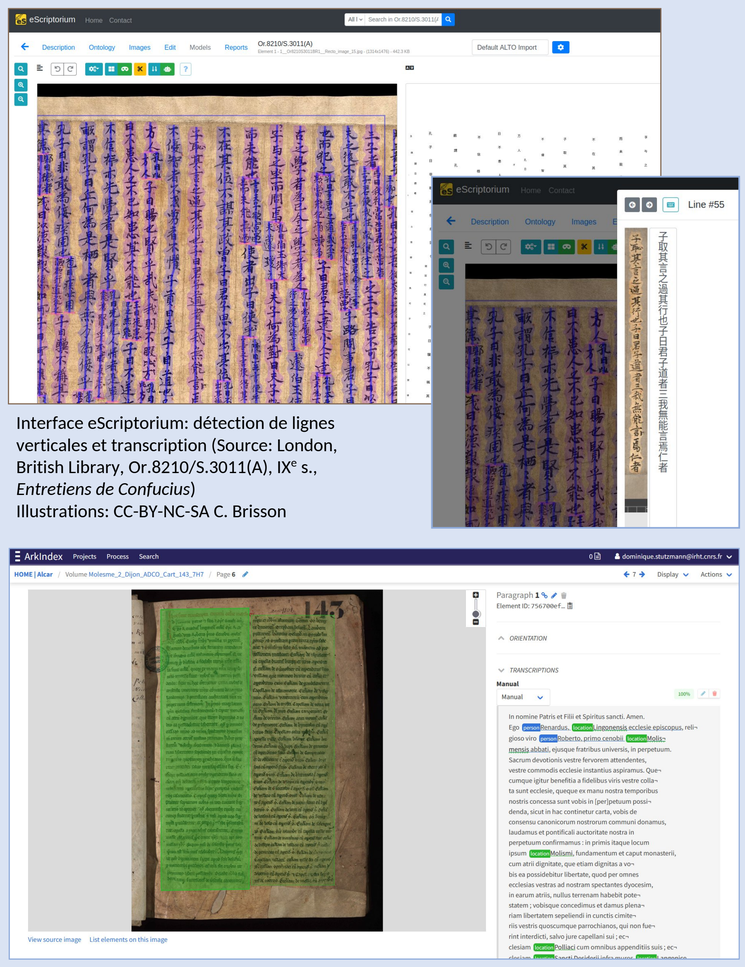
Reconnaissance des filigranes du papier et des éléments décoratifs, héraldiques, sigillographiques et numismatiques

- Répertoire de filigranes : création de métadonnées, missions photographiques (IRHT)
  - Base de filigranes : métadonnées, images ; science participative (ENC-PSL / IRHT)
- Répertoire des décors typographiques : reconnaissance automatique des formes et des fontes pour l'identification des imprimeurs
- Monnaies et données numismatiques
  - Système automatique de reconnaissance des coins monétaires antiques (AOrOc)
  - Acquisition des données en musée et réserves et mise en ligne (Craham - Caen)
  - Module Sigiscript - épigraphie du sceau (Saprat)
  - Référencement d'empreintes de monnaies sur supports céramiques

Détection de motifs dans les miniatures (personnes, livres), avec des modèles génériques ou après entraînements spécifiques avec Yolo et Segment Anything
(Sources: Paris, BnF, lat. 1408 et Baltimore, Walters Art M., W94, Livres d'heures)



Détection de miniatures (Arkindex/Teklia)



Interface eScriptorium: détection de lignes verticales et transcription (Source: London, British Library, Or.8210/S.3011(A), IXe s., *Entretiens de Confucius*)
Illustrations: CC-BY-NC-SA C. Brisson

Détection d'entités nommées. Ici: Corpus d'entraînement HOME-Alcar (doi: 10.5281/zenodo.5600883), présenté dans l'interface Arkindex de Teklia (Source: Dijon, Arch. dép., 7 H 7 Cart. 143)

Coordonner les développements, ouvrir et organiser les accès aux infrastructures, aux données et aux sources

- **Publication de logiciels open source, coordonner les développements**
  - Kraken (https://github.com/mittagessen/kraken)
  - eScriptorium (https://gitlab.com/scripta/escriptorium)
  - Arkindex (https://gitlab.teklia.com/arkindex)
- **Publication de modèles entraînés**
  - Communauté Zenodo (https://zenodo.org/communities/ocr_models/
  - Huggingface
    - https://huggingface.co/Teklia
    - https://huggingface.co/datasets/CATMuS/medieval
- **Infrastructures de calcul dédiées**
  - https://msia.escriptorium.fr/ , https://escriptorium.inria.fr/
  - Accès aux infrastructures : https://cremmacall.sciencescall.org/
- **Publication des données d'entraînement**

Faire communauté : les Journées du cluster 3 « Biblissim-IA »

- **Mars 2023** (https://biblissim-ia-2023.sciencesconf.org/)
  Tables-rondes thématiques sur l'IA en SHS
  Ateliers prospectifs et de formation
- **Mars 2024** (https://biblissim-ia-2024.sciencesconf.org/)
  Avancées des projets d'analyse de document avec IA

lundi 20 mars 2023

HEURES	ÉVÉNEMENT
14:00 - 14:45	Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites: principes, enjeux et potentiels (Open Space - Humathèque) - Dominique Stutzmann / Daniel Stoekl / Laurent I
14:45 - 18:00	IA et ressources humaines (Open Space - Humathèque)
14:45 - 15:45	› (Ne pas) mettre l'humain au service de la machine ? Vérité terrain, legacy data, données ouvertes et modèles entraînés - *Laurence Bobis, Bibliothèque interuniversitaire de la Sorbonne - textes - Simon Gabay, Université de Genève - Arsène Georges, Bibliothèque interuniversitaire de la Sorbonne - Thierry Kouamé, Université Bourgogne Franche-Comté [COMUE] - Martin Mo*
16:00 - 17:00	› Savoirs et formations - *Emmanuelle Bermes, École nationale des chartes - Jean-Baptiste Camps, École nationale des chartes - Matthieu Husson, Observatoire de Paris - Peter Stokes, Éco*
17:00 - 18:00	› Le coût invisible des projets interdisciplinaires - *Christopher Kermorvant, TEKLIA, Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes - Elena Pierazzo, Centre d'ét*

mardi 21 mars 2023

HEURES	ÉVÉNEMENT
09:30 - 12:30	IA et recherche en humanités : questions d'épistémologie (Open Space - Humathèque)
09:30 - 10:30	› Applicabilité de l'IA : y a-t-il des domaines où appliquer l'IA (n') est (pas) pertinent ? - *Bertrand Coüasnon, Institut de Recherche en Informatique et Systèmes Aléatoires - Laurent Hablot, É pratique des hautes études - Thierry Paquet, Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes*
10:30 - 11:30	› L'intelligence artificielle et les financements de la recherche en SHS - *Alexandre Gefen, THALIM - Théorie et histoire des arts et des littératures de la modernité - UMR 7172, Institut des S civilisations à tradition orale - Daniel Stoekl Ben Ezra, École pratique des hautes études*
11:30 - 12:30	› Intelligence artificielle, science de la donnée, néo-positivisme numérique et stratégies de recherche - *Elisa Grandi, Université Paris Cité - Torsten Hiltmann, Humboldt Universität zu Berlin des textes*
12:30 - 14:00	Déjeuner (Open Space - Humathèque)
14:00 - 17:30	Concevoir un projet de recherche avec l'IA (Open Space - Humathèque)
14:00 - 17:30	› Concevoir un projet avec de l'IA ? Brainstorming et speed dating... - *Matteo Ferrari, École pratique des hautes études - Antony Hostein, École pratique des hautes études - Jean-Philippe I Guilbert, École du Louvre - Aurélia Rostaing, Archives nationales*

mercredi 22 mars 2023

HEURES	ÉVÉNEMENT
09:30 - 13:00	Atelier n° 1 : eScriptorium (Campus Condorcet - Bât. Recherche Nord - Salle 0.010)
09:30 - 13:00	› Atelier d'initiation à l'usage d'eScriptorium - *Daniel Stoekl Ben Ezra, EPHE-PSL - Colin Brisson, EPHE-PSL - Simon Gabay, Université de Genève - Pawel Jablonski, EPHE-PSL - Benjamin H*
14:00 - 17:00	Atelier n° 2 : Sigillographie et IA (Campus Condorcet - Bât. Recherche Nord - Salle 2.001)
14:00 - 17:00	› Sigillographie et Intelligence Artificielle - *Victoria Eyharabide, Sorbonne Université - Laurent Hablot, École pratique des hautes études - Delia Préteux, Atelier national de recherche typograp*
14:00 - 17:00	Atelier n° 3 : Plans de gestion de données (PGD) (Campus Condorcet - Bât. Recherche Nord - Salle 2.001)

jeudi 23 mars 2023

HEURES	ÉVÉNEMENT
09:30 - 12:30	Atelier n° 1 : eScriptorium (Campus Condorcet - Bât. Recherche Nord - Salle 2.001)
09:30 - 12:30	› Atelier d'initiation à l'usage d'eScriptorium - *Daniel Stoekl Ben Ezra, EPHE-PSL - Colin Brisson, EPHE-PSL - Simon Gabay, Université de Genève - Pawel Jablonski, EPHE-PSL - Benjamin H*

jeudi 7 mars 2024

HEURES	ÉVÉNEMENT
09:15 - 10:00	Acueil - petit déjeuner (Foyer)
10:00 - 11:00	Discours Introductif - Cluster 3 Biblissima+ (Auditorium 150) - D. Stutzmann, D. Stoekl Ben Ezra
11:00 - 11:30	A presentation of ManuscriptAI, an upcoming project aiming to build an AI-tool to facilitate the integration, accessibility, and usability of heterogeneous cultural heritage data on medieval manuscripts (Auditorium 150) - S. Boodts, G. Biangioni
11:30 - 12:00	Classification of Ancient Greek Coins – Challenges Using AI methods (Auditorium 150) - P. Ulrike
12:00 - 12:30	DIANET : Développement d'une interface web simplifiée pour l'utilisation d'un système intelligent de comparaison de coins monétaires (Auditorium 150) - K. Gruel, O. Masson, M. Bui
12:30 - 14:00	Déjeuner (Foyer)
12:30 - 14:00	Session poster (Foyer)
14:00 - 14:30	Vers une segmentation robuste du tracé dans les manuscrits anciens (Auditorium 150) - C. Brisson
14:30 - 15:00	Transcribathon : Results of the 1st Week of the Digital Hebrew Book (2024) (Auditorium 150) - D. Stoekl Ben Ezra, L. Bambaci, B. Kiessling
15:00 - 15:30	Low-tech AI: iconography for everyone (Auditorium 150) - D. Stutzmann
15:30 - 16:00	Pause café (Foyer)
16:00 - 16:30	AI and the paleography of Greek papyri: presentation of the new EGRAPSA project (Auditorium 150) - G. De Gregorio, I.Marthot-Santaniello
16:30 - 17:00	PapyTwin net: a Twin network for Greek letters detection on ancient Papyri (Auditorium 150) - M. Tu Vu, M. Beurton-Aimar
17:00 - 17:30	PapyroLogos: Large data and little resources (Auditorium 150) - H. Essler, R. Ast

vendredi 8 mars 2024

HEURES	ÉVÉNEMENT
09:15 - 10:00	Accueil - Petit déjeuner (Foyer)
10:00 - 10:30	Ansund: Using Machine Learning to Develop a New, Exhaustive, Open Access Corpus of Old English (Auditorium 150) - M. Faulkner , E. Magnanti
10:30 - 11:15	Keynote speech : Signs, symbols and secrets: HTR faced with ciphered texts and music scores (Auditorium 150) - A. Fornes
11:15 - 11:30	Pause café (Foyer)
11:30 - 12:00	Deep Digital oriental: A state of the art for Sanskrit oriental handwritten text recognition (Auditorium 150) - S. Thottempudi
12:00 - 12:30	Intégration de modèles Seq2Seq pour la correction et la normalisation de textes issus de l'HTR (Auditorium 150) - S. Yatsyk
12:30 - 13:00	Entraînement from scratch ou fine-tuning ? Modalités et enjeux d'un choix d'un projet HTR sur les registres de plaidoiries du Parlement de Paris (fin XIVe-XVe siècle) (Auditorium 150) - P. Spychala
13:00 - 14:00	Déjeuner (Foyer)
13:00 - 14:00	Session poster (Foyer)
14:00 - 14:30	How Do Large Language Models (LLMs) Impact Document Image Analysis? An Open Discussion. (Auditorium 150) - A. Scius-Bertrand, L. Voegtlin, N. Wegmann, A. Fakhari, A. Fischer
14:30 - 15:15	TR and LLM feedback for automatic transcription in Divergent Use Cases (Auditorium 150) - H. Miller, M. Lavee, N. Bontemps
15:15 - 15:45	Integrating Vision Transformers and Large Language Models for Enhanced Handwriting Text Recognition and Named Entity Identification in Documentary Manuscripts (Auditorium 150) - S. Aguilar Torres
15:45 - 16:00	Pause café (Foyer)
16:00 - 16:30	« Un simple OCR ? » : outils, algorithmes et méthodes de reconnaissance de document en humanité numérique. (Auditorium 150) - C. Kermorvant , Y. Schneider, M. Blanco
16:30 - 17:00	Retour(s) d'expérience(s) d'indexation de registres d'emprunteurs de bibliothèques parisiennes: chaîne de production et de diffusion des archives de PRET19 (Auditorium 150) - V. Rebolledo-Dhuin, C. Kermorvant
17:00 - 17:45	Keynote speech : Information extrac cet

Biblissima+ Observatoire des cultures écrites de l'argile à l'imprimé — cnrs IRHT Institut de recherche et d'histoire des textes — École Pratique des Hautes Études — aoroc CNRS PSL — CAMPUS CONDORCET PARIS-AUBERVILLIERS — European Research Council — OPEN ISLAMICATE TEXTS INITIATIVE — TEKLIA — Inria