Biblissim-IA-2023. Journées annuelles du cluster 3 de l'EquipEx Biblissima+

Dominique Stutzmann

Biblissima+

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

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

Biblissim-IA-2023

Journées annuelles du cluster 3 de l'EquipEx Biblissima+

Humathèque, Campus Condorcet
Aubervilliers

20-23 mars 2023

Recueil des argumentaires édité par
Dominique Stutzmann
(IRHT-CNRS / Humboldt-Universität zu Berlin)

Table des matières

IA et ressources humaines

(Ne pas) mettre l'humain au service de la machine ? Vérité terrain, legacy data, données ouvertes et modèles entraînés

Table-ronde avec :

Laurence Bobis [1], Joanna Fronska [2], Simon Gabay [3], Arsène Georges [1], Thierry Kouamé [4], Martin Morard [2], Claudia Rabel [2]

[1] Bibliothèque interuniversitaire de la Sorbonne – Université Paris 1, Panthéon-Sorbonne, Université Paris 1 - Panthéon-Sorbonne – France
[2] Institut de recherche et d'histoire des textes – Centre National de la Recherche Scientifique – France
[3] Université de Genève – Suisse
[4] Université Bourgogne Franche-Comté [COMUE] – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. Il n'engage pas leur responsabilité scientifique.

Argumentaire
Ce qu'on appelle IA ou intelligence artificielle est aujourd'hui largement synonyme de " apprentissage machine " (*machine learning*). Dans les cas le plus favorable, le savoir est déjà formalisé ou disponible de façon semi-structurée et il est possible " d'apprendre " ou plutôt de faire apprendre à la machine à reproduire le comportement ou l'expertise humaine. Souvent néanmoins une phrase d'entraînement supplémentaire est nécessaire et peut prendre la forme d'annotations massive à l'usage de certaines tâches. Un exemple en sont les " captcha " qui ne servent pas, ou pas que, à prouver que nous ne sommes pas des robots, mais surtout apprendre à reconnaître des feux rouges, des ponts, des passages piétons à l'usage des futures voitures autonomes. Tant que la main-d'œuvre est volontaire et massive (les usager:e:s de tel ou tel moteur de recherche ou fournisseur de messagerie électronique), le tout est facilement intégré.

Dans le cas des communautés concernées par Biblissima+, les tâches qui pourraient être confiées à une machine portent sur des corpus à la fois restreints en nombre et complexes, et dont la complexité est précisément l'objet des études en SHS. Les équipes concernées peuvent vivre le besoin d'annotation pour entraîner une intelligence artificielle comme un asservissement des " humains " au service de la " machine " dans un processus qui n'est pas rentable pour les questions de recherche considérée.

En réponse aux problèmes posés par la nécessité de constituer des corpus d'entraînement pour permettre aux ordinateurs d'effectuer les tâches attendues, des solutions émergent. D'une part, le mouvement des " données ouvertes " (*open data*), qu'il s'agisse de bases de données textuelles

ou d'images ou d'éditions électroniques, permet de penser la disponibilité d'une expertise formalisée et partagée au-delà du changement et du renouvellement des outils. D'autre part, des communautés commencent à se structurer autour de " modèles pré-entraînés " et publiés et réutilisables, soit librement (publications sur Zenodo par exemple), soit dans un cadre plus restreint (modèles publics de Transkribus).

Contribuant à ces évolutions, plusieurs projets de recherche cherchent à formaliser et exploiter des données qui ont été produites dans un autre cadre et, souvent, avec un autre objectif de recherche. Dans le cas d'éditions de textes médiévaux, certaines ont pu servir à entraîner des modèles d'HTR spécifiques (projet Himanis), d'autres sont converties pour de nouvelles exploitations.

Cette table ronde permettra de considérer les choix faits pour la base iconographique Initiale (http://initiale.irht.cnrs.fr/), pour l'étude des manuscrits brûlés de Chartres (https://www.manuscrits-de-chartres.fr/) et pour l'édition électronique des manuscrits glosés de la Bible (https://gloss-e.irht.cnrs.fr/) avec les porteur:se:s de ces projets de recherche. Elle permettra aussi d'interroger les expériences et résultats de projets de mise à disposition de données et de modèles, ainsi que l'expérience en cours soutenue par Biblissima+ pour la transformation du *Chartularium Universitatis Parisiensis* et de ses suites pour alimenter la base prosopographique ORESM (*Œuvres et Référentiels des Étudiants, Suppôts et Maîtres de l'Université de Paris, des écoles et collèges parisiens*, 1200-1600).

Intervenant:e:s

Laurence Bobis est conservatrice générale, directrice de la Bibliothèque Interuniversitaire de la Sorbonne, porteuse avec Thierry Kouamé du projet *ECRU- Editions critiques relatives à l'Université de Paris.*

Joanna Fronska est historienne de l'art, ingénieure de recherche à l'Institut de Recherche et d'Histoire des Textes, responsable de la base de données *Initiale : Catalogue des manuscrits enluminés* et collaboratrice du projet *Renaissance virtuelle des manuscrits sinistrés de la bibliothèque de Chartres.* Ses centres d'intérêt et ses publications concernent la production, la circulation et l'usage des manuscrits de droit au Moyen Âge, l'iconographie politique et juridique et l'histoire des collections de manuscrits médiévaux.

Simon Gabay est maître-assistant à l'université de Genève auprès de la chaire de Béatrice Joyeux-Prunel et y dirige les projets *E-ditiones: corpus et outils pour l'étude du français classique* (https://github.com/e-ditiones) et *Katabase: base de données des manuscrits en circulation sur le marché privé* (https://github.com/katabase). Outre la philologie française moderne, ses principaux domaines de recherche sont le traitement automatique des langues et la reconnaissance optique de caractères.

Arsène Georges est ingénieur d'études chargé de projets numériques, impliqué dans les projet ORESM et ECRU (conception du modèle de données, traitement et publication des données textuelles).

Thierry Kouamé est professeur à l'Université de Bourgogne Franche-Comté et spécialiste de l'histoire de l'université. Il porte avec Laurent Bobis le projet *ECRU- Editions critiques relatives à l'Université de Paris.*

Martin Morard est chercheur à l'IRHT, spécialiste de l'histoire de l'exégèse de la Bible latine et directeur du projet *gloss-e - Glossae latinae omnes Scripturae -electronicae* (http://gloss-

e.irht.cnrs.fr), dans lequel il publie l'édition électronique des gloses et chaînes latines de la Bible, dont la *Catena aurea* de Thomas d'Aquin.
Claudia Rabel est historienne de l'art, spécialiste de l'iconographie et des manuscrits enluminés du Moyen Âge et responsable de la section des Manuscrits enluminés de l'IRHT et du projet projet dédié à la numérisation et à l'étude des manuscrits brûlés de Chartres. Elle est aussi responsable du séminaire de recherche Les Ymagiers, consacré à l'iconographie médiévale.

Savoirs et formations

Table-ronde avec :

Emmanuelle Bermes [1], Jean-Baptiste Camps [1], Matthieu Husson [2], Peter Stokes [3]

[1] École nationale des chartes – Université Paris sciences et lettres – France
[2] Observatoire de Paris – Université Paris sciences et lettres – France
[3] École pratique des hautes études – Université Paris sciences et lettres – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. L'argumentaire n'engage donc pas leur responsabilité scientifique.

Argumentaire

Les évolutions technologiques imposent évidemment des transformations des conditions de production et de transmission des savoirs. L'émergence des statistiques dans le champ de l'historien puis celles des humanités numériques ont suscité des débats et réflexions célèbres (l'historien programmeur...). L'intelligence artificielle impose de prendre en compte ce renouvellement peut-être plus encore que les humanités numériques qu'on serait presque tenté d'appeler " traditionnelles ". Au-delà des craintes que ChatGPT suscite pour l'évaluation des savoirs et acquis des étudiant:e:s, le modèle même de l'IA avec une couche très technique et un processus qui, par nature, est largement une boîte noire, remet en cause un modèle de savoir intégré.
Si, depuis longtemps, les recherches pouvaient être réalisées en équipe, leurs méthodes et leurs résultats étaient maitrisables par une seule personne. Encore plus que dans les "humanités numériques" des deux dernières décennies, l'intelligence artificielle (ou plutôt disons l'apprentissage machine) impose une recherche avec des tâches disséquées. En outre, elle se fonde sur l'utilisation et la création de " modèles ", tant au plan technique (les modèles " entraînés ") que dans la compréhension d'ontologie et de modélisation au sens des sciences humaines.

L'enjeu pour les formations actuelles pour les historien:ne:s qui peuvent être amené:e:s à mettre en œuvre ces techniques est grand. Faut-il créer des cursus doubles et allonger le temps d'étude ou créer des cursus spécialisés dans l'intermédiation ? Quelle est la place des apprentissages disciplinaires (aussi bien SHS que STIC) dans de nouvelles formations ? Faut-il privilégier la transversalité contre la spécialité ?

La table ronde réunira des chercheurs et des responsables de formation. Parmi les premiers, Matthieu Husson, à la fois historien des sciences et spécialiste des transformations disciplinaires dans les humanités numériques, et Peter Stokes, dont l'un des chantiers est la constitution d'un savoir formalisé et transmissible face au modèle ancien de la "paléographie oraculaire" où

l'expertise humaine et le *connoisseurship* fonctionnent comme une black box. Parmi les responsables de formation, Jean-Baptiste Camps et Emmanuelle Bermès interviendront tous les deux sur la modification des apprentissages et l'offre actuelle de formation.

Intervenant:e:s

Emmanuelle Bermès est responsable pédagogique du master " Technologies numériques appliquées à l'histoire " à l'Ecole nationale des Chartes (PSL). Archiviste paléographe et conservatrice générale des bibliothèques, elle a publié plusieurs livres sur les technologies numériques des bibliothèques.

Jean-Baptiste Camps est responsable pédagogique du master " Humanités numériques " de PSL. Ses recherches portent sur la littérature médiévale d'oc et d'oïl et concernent également, sur un plan méthodologique, les débats et pratiques de l'ecdotique (notamment l'édition électronique), ainsi que les méthodes computationnelles pour la production (reconnaissance optique de caractères, annotation linguistique, collation) ou l'analyse de données manuscrites (paléographique quantitative, scriptométrie, stylométrie, stemmatologie).

Matthieu Husson est chercheur au CNRS et ses recherches portent sur les pratiques mathématiques en astronomie aux XIIIe-XVe siècles et les Humanités numériques et intelligence artificielle. Il dirige également plusieurs projets internationaux (*EIDA Editing and analysing historical astronomical diagrams with Artificial Intelligence*, ANR-22-CE38-0014-01 ; *ALFA Shaping a European scientific scene,* ERC Consolidator Grant 2016 ; *DATA Digital Analysis Tools for the history of Astral sciences*).

Peter Stokes est directeur d'études à l'Ecole Pratiques des Hautes Etudes (PSL), titulaire de la chaire d'humanités numériques et computationnelles appliquées à l'étude de l'écrit ancien et chargé de mission pour les humanités numériques à l'EPHE.

Le coût invisible des projets interdisciplinaires

Table-ronde avec :

Christopher Kermorvant [1], Elena Pierazzo [2], (sous réserve) Florent Goiffon[3]

[1] TEKLIA – TEKLIA – France
[2] Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes – Université de Rouen Normandie
[3] Centre d'études supérieures de la Renaissance UMR 7323 – Ministère de la Culture et de la Communication, Université de Tours, Centre National de la Recherche Scientifique – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. L'argumentaire n'engage donc pas leur responsabilité scientifique.

Argumentaire

La bibliographie sur les disciplines, leur définition, leur utilité et l'interdisciplinarité est immense depuis le début des années 1970. Une publication précoce et largement diffusée, affirmait " l'interdisciplinarité ne s'apprend ni ne s'enseigne, elle se vit " (cité et critiqué par G. Palmade en 1977). Des contributions anciennes n'ont guère perdu en actualité, comme une interrogation sur les rapports entre théorie et pratique dans l'interdisciplinarité parue en 1983, donnant la primauté à une pratique de questionnement scientifique avant la théorie (pluri-, inter-, ou transdisciplinaire) et évoquant la labilité et la fluidité dans la mise en place de " appareils théoriques ou méthodologiques spécifiques et partiellement périssables ". Parmi les publications plus récentes, des chercheurs britanniques proposent une approche intégrant les structures de pouvoir (" *states of rest* "), les relations asymétriques et une approche phénoménologique par une analyse des sentiments développés au cours du travail interdisciplinaire (" *feeling fuzzy* ", Callard-Fitzgerald, 2015). Certaines de leurs analyses critiques sont très pertinentes et, au fond, rappellent et actualisent celles de G. Palmade : l'injonction à l'interdisciplinarité crée celle-ci avec ses difficultés matérielles, et, notamment dans un contexte d'emploi universitaire précaire, les mobilités géographiques et thématiques. Aussi importants sont les bonheurs et les souffrances de jeunes gens qui, pour les profils les plus " humanités numériques ", éprouvent des difficultés dans leurs choix de carrière et dans leurs perspectives de recrutement. Ces obstacles, signalés depuis les débuts de la réflexion sur le sujet, existent bien et l'on peut insister sur la charge morale que la pratique interdisciplinaire peut faire peser.
De l'amont à l'aval, de la conception à la réalisation du projet, mais aussi pour l'avenir des intervenant:e:s, les projets interdisciplinaires ont un coût en ce qu'il faut programmer " en plus " pour des recherches de cette nature : apprendre à se connaître, trouver un terrain commun pour définir les modalités de collaboration (par ex. prestation de service, recherche dans plusieurs champs en parallèle), identifier les bons acteurs, rédiger des projets, apprendre les bonnes méth-

odes du travail en interdisciplinarité et les mettre en œuvre, communiquer avec des équipes nouvelles, interpréter en commun les résultats des différentes équipes, préparer les projets suivants et préserver l'employabilité des personnes impliquées dans le projet.

Intervenants

Christopher Kermorvant est président fondateur de la société TEKLIA, spécialiste de reconnaissance automatique de documents par Intelligence Artificielle. La plateforme Arkindex développée par Teklia pour la compréhension automatique du document (classification, analyse de mise en page, reconnaissance de texte imprimé et manuscrit, extraction d'entités ommées), est issue de recherches et développements dans le cadre de projets interdisciplinaires et au service des institutions patrimoniales ou des équipes de recherche en humanités.

Elena Pierazzo, professeur des universités, directrice de l'UMR CESR, a dirigé le master humanités numériques au KIng's College de Londres, avant d'enseigner les humanités numériques en France. Présidente de la TEI (2012-15) et co-présidente de la conférence DH 2019 à Utrecht, elle dirigé, évalué et accompagné de nombreux projets interdisciplinaires.
Sous réserve : Florent Goiffon, PCN Horizon Europe "Cluster 2 & intégration des SHS" au Ministère de l'enseignement supérieur, de la recherche et de l'innovation

IA et recherche en humanités : questions d'épistémologie

Applicabilité de l'IA : y a-t-il des domaines où appliquer l'IA (n') est (pas) pertinent ?

Table-ronde avec :

Bertrand Coüasnon [1], Laurent Hablot [2], Benjamin Kiessling [3], Thierry Paquet [4]

[1] Institut de Recherche en Informatique et Systèmes Aléatoires – Institut National de Recherche en Informatique et en Automatique (INSA) de Rennes – France
[2] École pratique des hautes études – Université Paris sciences et lettres – France
[3] École pratique des hautes études – Université Paris sciences et lettres – France
[4] Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes – Université de Rouen Normandie – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. L'argumentaire n'engage donc pas leur responsabilité scientifique.

Argumentaire
Le terme " Intelligence artificielle " est très largement utilisé à l'heure actuelle et remplace dans bien des cas les termes plus précis d'apprentissage automatique (machine learning) qui donne la capacité d'apprendre à partir de données et, en particulier, de l'apprentissage supervisé, dans le cas où les données sont déjà annotées.

Pourtant, la plupart des domaines couverts par le périmètre d'intérêt de Biblissima+ sont considérés comme des domaines à faibles ressources (*low-resource domain*). Non seulement les corpus, même quand ils sont exhaustifs, y sont d'une taille souvent très inférieure à ceux utilisés dans les grands développements des dernières années, mais en outre les annotations disponibles, nécessaires à l'entraînement, sont en plus faible nombre encore. La production et la validation de nouvelles annotations dans ces domaines spécifiques nécessitent enfin des personnes qualifiées à haut niveau d'expertise.

Ici peut-être plus qu'ailleurs, le désir d'une " intelligence artificielle hybride " se fait sentir, qui combinerait les avantages de l'IA symbolique, fondée sur des règles, et de l'IA sub-symbolique, peu transparente et gourmande en données. L'enjeu pourrait être d'intégrer la logique d'ensemble formalisée par des générations de chercheur:se:s, avec leur connaissance implicite de leur domaine de spécialité, tout en étant assez souples pour accueillir l'infinie diversité des situations constatées dans les sources patrimoniales.

Pour discuter de l'applicabilité de l'intelligence artificielle dans les domaines des SHS couverts

par Biblissima+, quatre intervenants seront interrogés.

Intervenants

Bertrand Coüasnon est maître de conférences (HDR) en informatique à l'INSA Rennes et membre de l'IRISA. Il travaille sur la formalisation des connaissances, les méthodes génériques de reconnaissance de la structure des documents, la reconnaissance de l'écriture manuscrite, l'interaction utilisateur, l'apprentissage avec peu de données annotées, la combinaison apprentissage profond et méthodes syntaxiques, les systèmes auto-adaptatifs, avec des applications sur des documents d'archives et des documents complexes.

Benjamin Kiessling, ingénieur de recherche PSL, développe *kraken*, le module HTR (handwritten text recognition) au coeur des développements de la plateforme de transcription automatique des documents manuscrits *eScriptorium*.

Laurent Hablot est directeur d'études à l'Ecole Pratique des Hautes Etudes (PSL) et titulaire de la chaire d'emblématique occidentale. Il développe ses recherches sur les différents systèmes de signes - armoiries, cimiers, supports, cris - et les pratiques héraldiques dans leurs divers aspects artistiques, juridiques, symboliques et sociaux. Il dirige les projets de recherche et le développement des bases de données ARMMA (https://armma.saprat.fr/) et *Sigilla* (http://www.sigilla.org/) et est l'un des coordinateurs du cluster 3 de Biblissima+.
Thierry Paquet est professeur à l'université de Rouen et membre du laboratoire LITIS (EA 4108). Il est spécialiste en reconnaissance de l'écriture manuscrite, en reconnaissance de formes pour données séquentielles, en analyse et reconnaissance d'images de documents et des modèles probabilistes et réseaux de neurones.

L'intelligence artificielle et les financements de la recherche en SHS

Table-ronde avec :

Alexandre Gefen [1,2], Alexis Michaud [3], Daniel Stoekl Ben Ezra[4]

[1] THALIM - Théorie et histoire des arts et des littératures de la modernité - UMR 7172 – Université Sorbonne Nouvelle - Paris 3, Centre National de la Recherche Scientifique, Département Arts - ENS Paris – France
[2] Institut des Sciences Humaines et Sociales – Centre National de la Recherche Scientifique - CNRS – France
[3] Langues et civilisations à tradition orale – Université Sorbonne Nouvelle - Paris 3 : UMR7107, Institut National des Langues et Civilisations Orientales : UMR7107, Centre National de la Recherche Scientifique : UMR7107, Université Sorbonne Nouvelle - Paris 3, Institut National des Langues et Civilisations Orientales, Centre National de la Recherche Scientifique – France
[4] École pratique des hautes études – Université Paris sciences et lettres – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenantes listé:e:s ci-dessus. L'argumentaire n'engage donc pas leur responsabilité scientifique.

Argumentaire

Pour permettre aux collègues des communautés "humanités" de se projeter correctement dans des recherches incluant l'intelligence artificielle, ces journées du cluster 3 abordent les thématiques de discussion et les points bloquants, parmi lesquelles on identifie les problématiques de ressources humaines (l'humain au service de la machine, gestion du dialogue interdisciplinaire, coût invisible de la rédaction de projet) et des raisons d'épistémologie (domaines d'applicabilité et volumétrie des données, néo-positivisme de la data science, dislocation du savoir).

Cette table ronde abordera un thème à la jonction des deux : celui du financement et pilotage de la recherche et du fléchage de ressources. Elle permettra de clarifier le paysage actuel des financements, notamment en tenant compte de leurs impacts sur les SHS, en lien avec l'intelligence artificielle.

Les intervenants proposeront une vue large, dépassant le périmètre de Biblissima+, tout en restant connecté aux problématiques disciplinaires, pour expliciter comment des équipes de recherche perçoivent la capacité de l'IA à apporter du nouveau et comment les différents objectifs scientifiques sont combinés ou en concurrence, pour les crédits, les recrutements ou les

thématiques.

Intervenants

Alexandre Gefen est directeur de recherche au CNRS et Directeur Adjoint Scientifique à l'INSHS (Institut des Sciences Humaines et Sociales) du CNRS. Ses recherches portent sur des questions de théorie littéraire appliquées à la littérature française contemporaine et il a développé en parallèle une activité de recherche consacrée aux humanités numériques (philologie numérique et ses enjeux épistémologiques, les écritures en réseau) et aux cultures numériques. Il est le porteur du projet ANR CulturIA, visant à proposer une approche culturelle de l'intelligence artificielle (IA), de sa "préhistoire" jusqu'aux développements contemporains du deep learning, en combinant les méthodes de l'histoire des sciences, de l'histoire des idées et des imaginaires collectifs avec des analyses de terrain.

Alexis Michaud est directeur de recherche CNRS et directeur du LACITO. Il est spécialiste de phonétique/phonologie expérimentale et des langues tibéto-birmanes. Son expertise propre dans ce domaine de low resource languages et son action forte dans les opérations de documentations des langues en danger (notamment via le projet ANR-DFG CLD2025 Computational Language Documentation by 2025) se conjoignent avec son intérêt tant pour l'IA que pour la science ouverte.
Daniel Stökl Ben Ezra est directeur d'études à l'Ecole pratique des hautes études (EPHE-PSL), sur la chaire Langue, littérature, épigraphie et paléographie hébraïque et araméenne. Ses recherches portent sur les manuscrits de la mer Morte, la littérature rabbinique ancienne, et les humanités numériques. Ses publications électroniques incluent, entre autres, l'édition numérique de la Mishna (codirigé avec H. Lapin), ainsi que la plateforme open-source eScriptorium pour la transcription automatique de manuscrits. Il est l'un des porteurs du projet ERC Synergy *MIDRASH - Migrations of Textual and Scribal Traditions via Large-Scale Computational Analysis of Medieval Manuscripts in Hebrew Script.*

Intelligence artificielle, science de la donnée, néo-positivisme numérique et stratégies de recherche

Table-ronde avec :

Elisa Grandi [1], Torsten Hiltmann [2], Dominique Stutzmann[3],

[1] Université Paris Cité – France
[2] Humboldt-Universität zu Berlin – Allemagne
[3] Institut de recherche et d'histoire des textes – Centre National de la Recherche Scientifique – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. L'argumentaire n'engage donc pas leur responsabilité scientifique.

Argumentaire
La production massive de données et leur exploitation à grande échelle permet aux historiens et autres humanistes, quel que soit le domaine considéré, un discours scientifique argumenté et fondé sur la "science de la donnée". Faisant primer le document sur la théorie, ou, comme le formule O. Poncet au sujet de l'École des chartes, " se plaç(ant) à la tête d'un combat pour une histoire plus scientifique au contact des sources ", la science des données qui propose une vue plus claire des corrélations fait naître ce qui a été appelé un " néo-positivisme numérique " (Mosco 2014). Elle n'est pourtant ni neutre du point de vue axiologique, ni toujours menée de façon correcte pour le domaine historique. Les modélisations et calculs mathématiques oblitèrent une réalité historique plus complexe, des biais de source où les échantillonnages et les corpus de taille restreinte font toujours le risque que " sous l'apparence de rigueur et de scientificité que donnent les calculs de corrélation se trouvent des données chiffrées non probantes, au regard des standards d'administration de la preuve en histoire, y compris en histoire quantitative " tel qu'il a été formulé récemment dans une critique cinglante d'un article de data science (Anglaret et al., 2021).

À chaque étape du traitement des données, depuis la formalisation (" *capta* " plutôt que " *data* " selon J. Drucker, 2011) jusqu'à la production des éléments probants et au commentaire de graphiques, le risque est grand de surinterpréter ou de transformer. Quelle place ou espace peuvent prendre les approches théoriques et les réflexions ou sensibilités des chercheur.e.s dans un discours historique fondé sur la donnée de la science et comment formuler l'approche hypothético-déductive dans le discours historique ?

L'intelligence artificielle, et en particulier l'apprentissage automatique, ajoute une couche d'obscurité potentielle dans le traitement des données, en particulier par le biais d'apprentissage difficile à

maitriser. Que faire des données produites par les traitements automatiques quand il est impossible de rendre compte du processus de décision ? Quels changements d'approche imposent-ils ou occasionnent-ils ?

Outre l'administration de la preuve, des changements interviennent dans l'ordre du discours et de l'évaluation. Comment les historien.ne.s peuvent apprendre à travailler et à argumenter avec des résultats de traitements dont la performance est évaluée (par exemple, avec un algorithme dont la performance est évaluée sur un autre corpus à 80%) ? Les mesures habituelles du domaine des sciences dures (précision, rappel, F1-score) peuvent-elles aider les historien.ne.s à comprendre les données et la réalité du monde ? Les chercheur.e.s en humanités sont tenté.e.s d'affirmer la primauté de corpus maîtrisés à 100 %, mais peut-être jamais parfaitement saisis, et d'oublier les conséquence des erreurs ou lacunes de dépouillements. Comment considérer l'apport et les risques de l'intelligence artificielle dans ce contexte ?

Intervenants

Torsten Hiltmann est professeur de *digital history* à la Humboldt-Universität de Berlin, et aussi spécialiste de l'histoire de la communication visuelle au Moyen Âge. Ses recherches et publications portent sur l'impact de la datafication en histoire et les conséquences épistémologiques des changements de médialité.

Dominique Stutzmann est directeur de recherche au CNRS. Spécialiste de l'histoire de l'écriture au Moyen Âge, et faisant porter ses recherches sur des grands corpus issus de transcription automatique, il explore les conséquences des traitements automatiques et de la massification dans les sciences de l'érudition.

Concevoir un projet de recherche avec l'IA

Concevoir un projet avec de l'IA ? Brainstorming et speed dating...

Table-ronde avec :

Matteo Ferrari [1], Antony Hostein [1], Jean-Philippe Moreux [2], Anne Ritz-Guilbert [3], Aurélia Rostaing [4]

[1] École pratique des hautes études – Université Paris sciences et lettres – France
[2] Bibliothèque nationale de France – France
[3] École du Louvre – France
[4] Archives nationales – France

Pour information

L'argumentaire ci-dessous a été rédigé par les organisateurs de la conférence pour introduire la table ronde et non par les intervenant:e:s listé:e:s ci-dessus. L'argumentaire n'engage donc pas leur responsabilité scientifique.

Argumentaire
Le but de cet atelier est de permettre à des porteurs de projets en humanités numériques de soumettre à des spécialistes de l'IA des objectifs techniques ne pouvant *a priori* pas être pris en charge par les solutions actuelles.

Les contributeurs, à l'occasion d'une courte intervention (10 mn), présenteront sommairement leur projet pour se concentrer sur un aspect précis de leurs besoins et/ou d'un objectif à soumettre à des solutions relevant de l'IA. Les trois projets qui permettront de lancer la discussion sont choisis dans le domaine graphique, héraldique et numismatique, créant une cohérence sur les monuments, objets 3D et l'exploitation de leurs représentations.

Le public est invité à participer, à présenter ses questions et défis. Des groupes de discussion seront organisées pour faciliter les échanges.

Intervenant:e:s

Matteo Ferrari, docteur en histoire de l'art médiéval et ancien élève de l'École Normale Supérieure de Pise, interviendra comme représentant de la base *ArmmA: Armorial monumental du Moyen Âge* (http://base-armma.edel.univ-poitiers.fr/), qui vise, sur le long terme, à offrir une couverture la complète possible et à soutenir une étude approfondie des représentations héraldiques monumentales produites dans la France médiévale.
Antony Hostein, directeur d'études à l'EPHE, est historien et numismate, spécialiste de l'histoire de Rome, avec un intérêt particulier pour la Gaule, l'Orient romain ainsi que la période de la " crise du IIIe siècle ".

Jean-Philippe Moreux, ingénieur diplômé INSA Toulouse (informatique, 1990) est l'expert scientifique de Gallica à la Bibliothèque nationale de France. Il travaille sur les programmes de valorisation du patrimoine numérique de la BnF et participe à des projets de recherche sur ces sujets.

Aurélia Rostaing est archiviste paléographie, docteure en histoire de l'art et diplômée de l'École pratique des hautes études. Elle est conservatrice du patrimoine au département du Minutier central des notaires de Paris aux Archives nationales. A ce titre, elle a dirigé les travaux du projet Lectaurep au Minutier central, projet visant à la mise en place d'outils de transcription collaborative et à la reconnaissance automatique de l'écriture manuscrite des sources modernes.

Anne Ritz-Guilbert, historienne de l'art, rendra compte de son expérience et réflexions dans le cadre du projet COLLECTA, consacré à la reconstitution et à l'étude de la collection documentaire de François-Roger de Gaignières (1642-1715). La reconstitution virtuelle de la collection rend compte à la fois de son classement originel, de ses sources et de ses usage (https://www.collecta.fr/index.php).

Atelier n∘ 1 : eScriptorium

Atelier d'initiation à l'usage d'eScriptorium

Atelier pratique avec :

Daniel Stoekl Ben Ezra [1,2], Colin Brisson [3,4], Simon Gabay [5], Pawel Jablonski [6,7], Benjamin Kiessling [6,7], Svetlana Yatsyk [8]

[1] EPHE, PSL University – EPHE, PSL University – France
[2] U.M.R.8546-Laboratoire AOROC, 4 Rue Lhomond, 75005 Paris, France – U.M.R.8546-Laboratoire AOROC, 4 Rue Lhomond, 75005 Paris, France – France
[3] EPHE, PSL – CRCAO – France
[4] Centre de recherche sur les civilisations de l'Asie Orientale – Ecole Pratique des Hautes Etudes, Collège de France, Centre National de la Recherche Scientifique, Université Paris Cité – France
[5] Université de Genève – Suisse
[6] EPHE, PSL – UMR 8546 AOROC, Paris, France. – France
[7] UMR 8546 AOROC, Paris, France. – UMR 8546 AOROC, Paris, France. – France
[8] CIHAM UMR 5648 – UMR 5648 – France

Cet atelier initialisera les participants durant deux matinées à l'usage d'eScriptorium, logiciel libre et open source pour l'analyse de documents manuscrits et imprimés, y compris des inscriptions. Après une initiation à des aspects théoriques nous aborderons la segmentation et la transcription automatiques et manuelles, l'alignement automatique à des transcriptions existantes ainsi que les bases de la gestion de projet.
Nous pouvons accueillir un maximum de 20 participants. Inscription obligatoire. Nous conseillons fortement d'étudier de près le tutoriel : https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium. Dès que vous auriez reçu le lien d'accès à l'infrastructure msIA (manuscriptologIA), il vaudra bien de faire toutes les étapes détaillez dans ce tutoriel.

Regarder au moins une partie des vidéos d'explication en amont va faciliter grandement l'apprentissage :
https://vimeo.com/user130532566

MERCREDI, 22 mars 2023:

a) présentation rapide des participants (15 min)
b) introduction générale (navigation UI, bases IA) (60 min)

15 min pause

c) import d'images, segmentation automatique, transcription automatique et correction de transcription manuelle (45 min)

15 min pause

d) correction manuelle de la segmentation / session pratique sur un document commun (60

min)

JEUDI, 23 mars 2023:

e) import xml, zip, iiif / export (30 min)

10 min pause

f) entrainement de modèles avec eScriptorium et smart bootstrapping (50 min)

10 min pause

g) alignement automatique (30 min)

10 min pause
h) session pratique des participants sur leurs propres documents (40 min)

Atelier n∘ 2 : Sigillographie et IA

Sigillographie et Intelligence Artificielle

Séance et discussion avec :

Victoria Eyharabide [1], Laurent Hablot [2], Philippe Jacquet , Catherine Kasteleiner [3], Delia Préteux [4], Alessio Sopracasa [1]

[1] Sorbonne Université – Sorbonne Université – France
[2] École pratique des hautes études – Université Paris sciences et lettres – France
[3] Université de Strasbourg – France
[4] Atelier national de recherche typographique (ANRT) – Ministère de la Culture – France

Plusieurs programmes de bases de données consacrées aux sceaux occidentaux ou byzantins se penchent plus spécifiquement sur le traitement des données épigraphiques ou la reconnaissance automatique d'images et explorent les possibilités de l'IA.

Les inscriptions des sceaux offrent en effet un répertoire considérable d'attestations épigraphiques datées et localisées. Leur traitement numérique reste pourtant délicat tant en raison de leur mise en forme circulaire qu'à cause des innombrables variantes des types d'écritures.

La reconstitution automatique d'empreintes complètes à partir de fragments issus d'une même matrice se heurte à de multiples écueils (usures de différentes natures, perturbation du modelé après décollement de la matrice, variété des cires, etc.) et interroge les possibilités numériques.

La reconnaissance de certaines images au sein de l'iconographie du sceau, notamment les armoiries, reste un sujet complexe sur lequel travaillent aujourd'hui plusieurs projets de recherches.

Cet atelier sera l'occasion de présenter les projets et cours, les solutions retenues et les perspectives attendues par les différentes bases sigillographiques.

Y participeront notamment le projet ANR BHAI - Artificial intelligence applied to Byzantine Sigillography (https://anr.fr/Projet-ANR-21-CE38-0001/), présenté par Victoria Eyharabide (Sorbonne Université), le projet ANR Digibyzseal - base des sceaux byzantins - présenté par Alessio Sopracasa (Sorbonne Université)), la base SIGILLA - base numérique des sceaux conservés en France présentée par Laurent Hablot, Philippe Jacquet et Catherine Kasteleiner (EPHE-PSL) ; le projet de recherche sur la normalisation de l'épigraphie des sceaux par Delia Préteux (ANRT)

Liste des sponsors

EquipEx Biblissima+

Biblissima+ (Observatoire des cultures écrites, de l'argile à l'imprimé) fait partie des équipements structurants pour la recherche ÉquipEx+ sélectionnés en 2020 dans le cadre des Investissements d'avenir. Il prend le relais de l'ÉquipEx Biblissima (Bibliotheca bibliothecarum novissima : observatoire du patrimoine écrit du Moyen Âge et de la Renaissance, 2012-2021). Biblissima+ est porté par le Campus Condorcet, piloté par Anne-Marie Turcan-Verkerk (AOROC, EPHE-PSL) avec François Bougard (IRHT, CNRS), Marie-Agnès Lucas-Avenel (CRAHAM, Université de Caen) et Emmanuelle Morlock (HiSoMa, CNRS).

Biblissima+ fédère 16 établissements et une entreprise privée, réunis en un consortium et agissant par l'intermédiaire de leurs services ou d'unités de recherche ou de service placées sous leur tutelle. Sur le plan opérationnel, la mise en œuvre du programme repose sur l'établissement coordinateur (EPCC Campus Condorcet) et sur les équipes fondatrices de Biblissima+.

IRHT-CNRS

L'Institut de recherche et d'histoire des textes, fondé en 1937, est une unité propre du CNRS. Il se consacre à la recherche sur les manuscrits, principalement médiévaux (livres et archives sur parchemin, papier, papyrus), avec des prolongements vers les périodes antiques et modernes. À travers eux, c'est la transmission des textes qu'ils contiennent comme « témoins » qui est en jeu, de l'Antiquité à la Renaissance, c'est-à-dire jusqu'au premier livre imprimé, en ce qu'il a recueilli l'héritage du manuscrit.

Humathèque Condorcet (Campus Condorcet)

Bibliothèque et archives du CampusCondorcet, l'Humathèque est née de la mutualisation de plus de 50 bibliothèques, fonds documentaires et services d'archives.

Liste des intervenant:e:s