

Caroline KOUDORO-PARFAIT

DOCTEURE (PHD) · CHERCHEUSE EN HUMANITÉS NUMÉRIQUES (HN) ET
TRAITEMENT AUTOMATIQUE DES LANGUES (TAL)
Sorbonne Université, Paris, France

✉ caroline.parfait10@gmail.com | 🏠 carolinekoudoroparfait.github.io | 📄 These-SCAI2023

Parcours Universitaire

Projet postdoctoral : HN et TAL pour la recherche en histoire moderne

Trèves, Allemagne

UNIVERSITÉ DE TRÈVES

oct. 2025 - oct. 2027

- Bourse de recherche décernée par la Fondation Alexander von Humboldt
- Collaboration avec le Prof. Dr. Damien Tricoire

Explorer les mobilités féminines entre la métropole et les Antilles (XVI^e-XIX^e siècles) : défis de l'extraction d'informations et modélisation des données socio-historiques.

- Mots clés : Reconnaissance de l'écriture manuscrite, Reconnaissance optique de caractères, Extraction de mots clés, Reconnaissance d'entités nommées, Graphique de réseaux, Cartographie.

Thèse de Doctorat, Littérature comparée – Recherches transdisciplinaires en Humanités numériques littéraires et TAL

Paris, France

SORBONNE UNIVERSITÉ (SU)

oct. 2020 - nov. 2024

- Contrat doctoral du Sorbonne Center for Artificial Intelligence
- Directeurs de Thèse : Glenn Roe (ObTIC et CELLF) et Gaël Lejeune (STIH) ; co-encadrant : Motasem Alrahabi (ObTIC)

Des IA au service de l'espace littéraire du XIX^e siècle : évaluation et analyse des outils de reconnaissance d'entités nommées spatiales.

- Dépôt : <https://theses.hal.science/tel-05042915v1>
- Soutenance le 6 janvier 2025

Master Intelligence des données de la culture et des patrimoines

Tours, France

CENTRE D'ÉTUDES SUPÉRIEURES DE LA RENAISSANCE DE TOURS (CESR)

2018 - 2020

- Directrice de mémoire : Elena Pierazzo
- Titre du mémoire : *Inventorier, Modéliser, Concevoir, Évaluer : enjeux et pratiques des Humanités Numériques au sein d'une entreprise de développement informatique.*

Diplôme national supérieur d'expression plastique – Grade Master

Cergy, France

ÉCOLE NATIONALE SUPÉRIEURE D'ARTS DE PARIS-CERGY (ENSAPC)

2011 - 2017

- Directeur de mémoire : François Bon
- Titre du mémoire : *L'inertie, le mouvant*
- Pratiques plastiques : Sérigraphie, vidéo, creative writing et écritures numériques

Erasmus+

Oslo, Norvège

KUNSTHØGSKOLEN I OSLO (KHIO)

2014 - 2014

- Beaux-arts / arts plastiques, section textile

Maîtrise de Philosophie de l'art et esthétique

Paris, France

UNIVERSITÉ, PARIS 4, PARIS-SORBONNE

2007 - 2011

- Enseignements mineurs : Philosophie et histoire des sciences, langages logiques, communication et philosophie politique et éthique.

Parcours professionnel

2025-2027	Chercheuse Postdoctorale. Financement de la Fondation Alexander Von Humboldt , Université de Trèves, Allemagne
2023-2025	Attachée temporaire d'enseignement et de recherche (ATER) – informatique pour les sciences humaines , UFR sociologie et informatique, Sorbonne Université
2021-2023	Chargée d'enseignement vacataire : Programmation Python , Institut National des Langues et Civilisations Orientales (INALCO)
2020-2023	Doctorante contractuelle. Financement du Sorbonne Center for Artificial Intelligence , Sorbonne Université
2020	Stage d'ingénierie en Humanités numériques : Conception et développement d'une bibliothèque numérique sous Omeka S et état de l'art des outils numériques des SHS , Acatus Informatique, Orléans

Enseignements

2021 - 2026 = **559.5 heures équivalent TD** (heTDs).

ATER, SCIENCES DU LANGAGE ET INFORMATIQUE – SORBONNE UNIVERSITÉ, 2023 - 2025 = 380 heTDs
 Responsable des CM et TD de Programmation Python, Niveau L3, Sorbonne Université, 2024 - 2025.

Nom de la matière	Années	Niveau	heTDs		Supports
			CM	TD	
Total = 380 heTDs			141	239	
Programmation Python S5, S6	23/25	L3	126	84	Supports personnels : https://github.com/These-SCAI2023/L5SOPRG_2023-2024 ; supports personnels : https://github.com/These-SCAI2023/L6SOPROG/tree/main/PDF_CKP
Humanités Numériques	23/24	M2	3	–	Supports personnels : https://github.com/These-SCAI2023/M4LFHUNU
Analyse Lexicale en TAL	23/25	M1	1,5	4,5	Supports hérités
Linguistique Informatique	23/25	L2	–	6,5	Supports hérités
PIX - Programmer et résoudre des problèmes informatiques	23/25	L2	10,5	144	Supports hérités

CHARGÉE DE COURS, PROGRAMMATION PYTHON – INALCO, 2021 - 2023 = **171.5 heTDs**

Nom de la matière	Années	Niveau	heTDs		Supports
			CM	TD	
Total = 171.5 heTDs			73.5	98	
Algorithmique et structuration des données ; Programmation Python	22/23	L3	36	48	Supports personnels : https://github.com/These-SCAI2023/Algorithmique_INALCO_2021-2023
Algorithmique et structuration des données ; Programmation Python	21/22	L3	37.5	50	<i>idem</i>

ENCADREMENT DE MÉMOIRES ET STAGES

date	sujet	Statut	étudiant-e-s	Université
2025	<i>Reconnaissance d'Entités Nommées sur un corpus de poèmes du XIX^e siècle</i>	Mémoire de M1, SDL	LOPES M., ZHENG, R.	SU
2025	<i>Entraînement et évaluation d'un modèle de Reconnaissance d'Entités Nommées sur données bruitées</i>	Mémoire de M1, SDL	FIEF C., CANONGE Y.	SU
2024	<i>Évaluation des problèmes d'étiquetage de la REN.</i>	Stage de L3, SDL	FIEF, C.	SU
2023	<i>Évaluation et développement de méthodes pour la désambiguïsation sémantique d'entités nommées spatiales</i>	Stage de M2, IDCP	DORLÉANS, C.	CESR
2022	<i>Résolution de problèmes d'Entity Linking sur des sorties OCR bruitées</i>	Mémoire de M1, SDL	BUTH, P.-R.	SU
2021	<i>Les lieux de Paris dans la chanson française</i>	Mémoire de M1, SDL	LAZREG, M.	SU

SDL : Sciences du langage ; IDCP : Master Intelligences des données de la culture et des patrimoines. SU = Sorbonne Université, Paris ; CESR = Centre d'études supérieures de la Renaissance, Tours.

ACTIVITÉS D'ENSEIGNEMENTS PONCTUELLES

ENSEIGNANTE, DSAA 1 NUMÉRIQUE (GRADE MASTER) – ENSAAMA, MARS 2026 = **8 heTDs**

Nom de la matière	Années	Niveau	heTDs		Supports
			CM	TD	
Total = 8 heTDs					
Les systèmes IA pour la création littéraire : Programmation Python	25/26	M1	–	8	En cours d'élaboration

PRÉSIDENCE DE JURY DE DIPLÔME DE L'ENSEIGNEMENT SUPÉRIEUR

- 13-17 avril **Présidente du jury pour le Diplôme Supérieur des Arts Appliqués valant grade Master – Expériences numériques, en qualité d'experte du domaine des pratiques numériques**, École nationale supérieure des arts appliqués et des métiers d'art, Olivier de Serres, Paris, France
- & 3-4 juin 2026
- 4-5 juin **Présidente du jury pour le Diplôme Supérieur des Arts Appliqués – Expériences numériques, en qualité d'experte du domaine des pratiques numériques**, École nationale supérieure des arts appliqués et des métiers d'art, Olivier de Serres, Paris, France
- 2025

Publications

SYNTHÈSE DES PUBLICATIONS EN DOCUMENT ANALYSIS, HUMANITÉS NUMÉRIQUES ET TAL

Catégories	Domaines	Non classées	Rang A
Total = 28	-	27	1
Conférences internationales = 8	Document Analysis	1	1
	Humanités numériques	4	-
	TAL	2	-
Revue internationale = 1	TAL	1	-
Conférences nationales = 2	Humanités numériques	1	-
	TAL/Linguistique de corpus	1	-
Revue nationale = 1	TAL/Linguistique de corpus	1	-
Chapitres de livre = 1	Humanités numériques	1	-
Workshops internationaux = 7	Document Analysis	1	-
	Humanités numériques	4	-
	TAL	2	-
Invitations = 8	Humanités numériques	8	-

Table 1: Répartition des publications selon le type de publication, le domaine et le rang (Non classées, A)

CONFÉRENCES INTERNATIONALES AVEC COMITÉ DE LECTURE

- Koudoro-Parfait, C.**, Hernandez, M., Lejeune, G., Dupont, Y. (2025). Epimethee – A Workflow from OCR to Spatial Mapping. In: Yin, X.C., Karatzas, D., Lopresti, D. (eds) Document Analysis and Recognition – ICDAR 2025. ICDAR 2025. Lecture Notes in Computer Science, vol 16025. Springer, Cham. https://doi.org/10.1007/978-3-032-04624-6_1, Conférence de **Rang A**.
- Koudoro-Parfait, C.**, Fief, C. (2025). Évaluer les erreurs d'étiquetage de la reconnaissance d'entités nommées sur des transcriptions OCR de romans français [Evaluating Labeling Errors in Named Entity Recognition on OCR Transcriptions of French Novels]. Proceedings of the Conference Humanistica 2025, Dakar, Sénégal.
- Koudoro-Parfait, C.**, Alrahabi, M., Dupont, Y., Lejeune, G., Roe, G. (2023). Mapping spatial named entities from noisy OCR output: Epimetheus from OCR to map. Scholger, W., Vogeler, G., Tasovac, T., Baillet, A., Helling, P., Collaboration as Opportunity (DH2023), ADHO Digital Humanities Conference, Graz, Austria, 154-155. <https://zenodo.org/records/8108050> (zenodo)
- Koudoro-Parfait, C.**, Lejeune, G., Buth, R. (2022). Reconnaissance d'entités nommées sur des sorties OCR bruitées: des pistes pour la désambiguïsation morphologique automatique [Named Entity Recognition on Noisy OCR Output: Approaches to Automatic Morphological Disambiguation]. Proceedings of the 29th Conference on Natural Language Processing. TAL and Digital Humanities Workshop (TAL-HN), ATALA, Avignon, France, 45-55. <https://aclanthology.org/2022.jeptalnrecital-humanum.6/> (aclanthology)

5. **Koudoro-Parfait, C.** (2022). Évaluation de la tâche de clustering pour l'alignement de formes contaminées d'entités nommées issues d'un corpus OCR bruité [Assessment of Clustering Methods for Aligning Contaminated Named Entity Variants in Noisy OCR Corpora]. Corro, C., Lejeune, G. Proceedings of the Symposium on the Robustness of NLP Systems, Paris, France, 5-8. <https://hal.science/hal-03853541v1/file/main.pdf> (hal-03853541f).
6. Alrahabi, M., Lejeune, G., **Parfait, C.**, Roe, G. (2021). Discovering Spatial Relations in Literature : What is the influence of OCR noise? Kanner, A., Mäkelä, E., Marjanen, J., Tolonen, M., Oberbichler, S., Duong, Q., Pivovarov, L., Ali, D., Verstockt, S., Ollion, É., Shen, R., Arnold, M., Brown, D., Adam, R., Balasubramanian, S., Charvat, V. M., Füllsack, M., Kleinert, J., Misera, H., ... Lomazow, S. (2021). The Book of Abstracts for What's Past is Prologue: The NewsEye International Conference. What's Past is Prologue: The NewsEye International Conference - Towards a future of interdisciplinary collaboration between Cultural Heritage, Digital Humanities, Computer Science and Data S. 20. <https://doi.org/10.5281/zenodo.5167375> (Zenodo)
7. Arruga, M., **Koudoro-Parfait, C.** (2020). Development of digital platform about stained glass restoration data in France. MSHValdeLoire, in FAIR Heritage -17 and 18 juin 2020. [Vidéo]. Canal-U. <https://doi.org/10.60527/adxq-pe72>. (vidéo consultée le 25 novembre 2025)
8. **Parfait, C.**, Arruga, M., Tigero, R. (2020). Espace Numérique du Vitrail - ENVi, Développer une plate-forme numérique sur les données de restauration du vitrail en France. Colloque Humanistica 2020.

REVUES INTERNATIONALES AVEC COMITÉ DE LECTURE

1. **Koudoro-Parfait, C.**, Petkovic, L., Roe, G. (2023). Analyse multilingue de l'impact de la correction automatique de la ROC sur la reconnaissance d'entités nommées spatiales dans des corpus littéraires [Multilingual Analysis of the Impact of OCR Correction on Spatial Named Entity Recognition in Literary Corpora]. Corro, C., Lejeune, G., Niculae, V. Revue TAL, Vol. 64 - no2 / 2023, 43-67. https://www.atala.org/index.php/content/tal_64_2_2

CONFÉRENCES NATIONALES AVEC COMITÉ DE LECTURE

1. **Koudoro-Parfait, C.**, Lejeune, G., Petkovic, L. (2023). Correction automatique des interférences OCR dans la reconnaissance d'entités nommées spatiales : réel gain ou perte de l'information ? [Automatic Correction of OCR Interference in Spatial Named Entity Recognition: Real Gain or Loss of Information?]. Gugliotta, E., Pallanti, L., Kraif, O., Fabry, I., Barletta, M. Journée d'étude Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques [Symposium—Background Noise or Added Value? Managing Noise in the Computational Processing of Linguistic Corpora]. Grenoble, France, 2-5. <https://hal.science/hal-05176056> (hal-05176056)
2. **Koudoro-Parfait, C.**, Lejeune, G. (2022). De la création des corpus numériques à leur utilisation : les interférences dans la reconnaissance d'entités nommées spatiales [From Building Digital Corpora to Using Them: Challenges and Interference in Spatial Named Entity Recognition]. Gherdevich, D., Mermet, É. Séminaire Des sources aux systèmes d'information géographique : des outils pour la cartographie dans les humanités numériques [Seminar: From Sources to Geographic Information Systems — Tools for Mapping in Digital Humanities], Paris (online), France, 1-2. <https://hal.science/hal-05176021> (hal-05176021)

REVUES NATIONALES AVEC COMITÉ DE LECTURE

1. Petkovic, L., **Koudoro-Parfait, C.**, Desmarest, M.-S., Lejeune, G. (2025). Quelle solution pour améliorer les performances de la reconnaissance d'entités nommées sur des données bruitées, corriger l'entrée ou filtrer la sortie ? [Improving Named Entity Recognition on Noisy Data: Should We Correct the Input or Filter the Output?]. Gugliotta, E., Pallanti, L., Kraif, O., Barletta, M., Fabry, I. Revue Corpus n°26, 2-19. <https://doi.org/10.4000/1364s> (DOI)

CHAPITRES DE LIVRE AVEC COMITÉ DE LECTURE

1. **Koudoro-Parfait, C.**, Lejeune, G. (2023). Reconnaissance des Entités Nommées spatiales sur un corpus littéraire bruité : des entités à la carte [Spatial Named Entity Recognition in a Noisy Literary Corpus: A Customized Entity Mapping Ap-

proach]. Gherdevich, D., Mermet, É. Book Chapter, Des sources aux systèmes d'information géographique : des outils pour la cartographie dans les humanités numériques. Publishers, Matériologiques dans la collection Modélisations, simulations, systèmes complexes. (déclaration d'acceptation de l'éditeur)

WORKSHOPS/ATELIERS INTERNATIONAUX AVEC COMITÉ DE LECTURE

1. Boubehziz, T., **Koudoro-Parfait, C.**, Lejeune, G. (2025). Assessing the Impact of Image Resolution on OCR Transcription Accuracy. Vyshnevetska, V. 3rd International Workshop on Digital Language Archives de la conférence ACM/IEEE Joint Conference on Digital Libraries 2025.
2. Bezançon, J., Boubehziz, T., Chutaux, C., Zine, O., Acensio, L., **Koudoro-Parfait, C.**, Briglia, A., Lejeune, G. (2023). SPQR@Deft2023: Résolution automatique de QCM médicaux à partir de corpus de domaine et de mesures de similarité [SPQR@Deft2023: Automatic Solving of Medical Multiple-Choice Questions Using Domain Corpora and Similarity Measures]. Bazoge, A., Daille, B., Dufour, R., Labrak, Y., Morin, E., Rouvier, M. TALN2023 Conference Proceedings, Workshop Défi Fouille de Textes@TALN2023, Text Mining Challenge@TALN2023, ATALA, Paris, France, 11-22. https://coria-taln-2023.sciencesconf.org/data/proceedings_DEFT.pdf (hal-04131847)
3. Ben Ltaifa, I., Boubehziz, T., Briglia, A., Chutaux, C., Dupont, Y., González-Gallardo, C., **Koudoro-Parfait, C.**, Lejeune, G. (2022). Stylo@ DEFT2022: Notation automatique de copies d'étudiant·e·s par combinaisons de méthodes de similarité [Stylo@ DEFT2022: Automated Assessment of Student Responses via Hybrid Similarity Approaches]. Proceedings of the 29th Conference on Natural Language Processing. Atelier DÉfi Fouille de Textes (DEFT) [Text Mining Challenge], ATALA, 11-22. <https://hal.sorbonne-universite.fr/TALN-RECITAL2022/hal-03705873v1> (hal-03705873)
4. **Koudoro-Parfait, C.** (2022). Spatial Named Entities Recognition in Literary Texts: the Crossroads of Digital Humanities and Natural Language Processing. Baillot, A., Magué, J.-P. Natural Language Processing meets Digital Humanities, Lyon (online), France, 1-2.
5. **Koudoro-Parfait, C.**, Lejeune, G., Roe, G. (2021). Spatial Named Entity Recognition in Literary Texts: What is the Influence of OCR Noise? Moncla, L., Brando, C., McDonough, K. GeoHumanities@SIGSPATIAL 2021: Proceedings of the 5th ACM, SIGSPATIAL International Workshop on Geospatial Humanities, Beijing (online), China, 2 - 5 Novembre, 13-21. <https://dl.acm.org/doi/10.1145/3486187.3490206>
6. **Koudoro-Parfait, C.**, Lejeune, G., Alrahabi, M., Roe, G. (2021). Reconnaissance d'entités nommées spatiales dans des transcriptions OCR bruitées de textes littéraires [Spatial Named Entity Recognition in Noisy OCR Transcriptions of Literary Texts]. Brando C., Moncla L. SAGEO 2021, Atelier Humanités Numériques Spatialisées [Workshop on Spatialized Digital Humanities,], Paris (online), France, 1-2. <https://hal.science/hal-05175990> (hal-05175990).
7. **Koudoro-Parfait, C.**, Lejeune, G., Tanguy, J.-B. (2021). Réutiliser les corpus numériques constitués par OCR : bruit et silence dans les data papers [Reusing OCR-Generated Corpora: Noise and Silence in Data Papers]. Bardiot, C., De Maeyer, J. DHNord2021 : Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux [DHNord2021: Publishing, Sharing, and Reusing Research Data: Data Papers and Their Challenges]. <https://hal.science/hal-04029006> (hal-04029006)

INVITATIONS DANS DES CONFÉRENCES ET SÉMINAIRES

1. **Parfait, C.** (2025). Explorer la collection Française d'ELTeC avec l'outil de cartographie textuelle Epiméthée. Dumenieu, B., Abadie, N., Brando, C., Moncla, L. Ateliers MAGIS, Humanités numériques spatialisées, SAGEO 2025. <https://sageo2025.sciencesconf.org/resource/page/id/3>
2. Bouchard, A., **Koudoro-Parfait, C.**, Solnon, C. (2024). Table ronde : L'impact de l'IA sur les pratiques documentaires : regards croisés entre chercheurs et bibliothécaires. Bouju, P. ENSSIB.
3. **Parfait, C.**, Nascimento, L. (2024). Table ronde : s'approprier sa recherche. La posture des doctorantes et doctorants ..., axe : « ... face aux Intelligences artificielles (IA) ». Courtois, M. et Collectif doctoral. Sorbonne Actuelle 2024. <https://lettres.sorbonne-universite.fr/sites/default/files/media/2024-05/Programme%20-%20S%27approprier%20sa%20recherche.pdf>
4. **Koudoro-Parfait, C.** (2023). Garder trace de son travail et le partager en traitement automatique des langues. Massot M.-L., Tricoche, A., Schultz, E. Atelier Digit_Hum 2023, Manipuler des données en SHS : R, Python et les carnets interactifs. École normale supérieure. <https://explore.psl.eu/fr/le-magazine/videos/garder-trace-de-son-travail-et-le-partager-en-traitement-automatique-des-langues>

5. Lejeune, G., **Koudoro-Parfait, C.** (2023). OCR impact on Named Entity Recognition on historical documents : Problem Presentation ...and Epistemological consequences. Summer School on Digital Humanities, UNAM/SCAI, Université nationale autonome du Mexique.
6. **Parfait, C.** (2022). Reconnaissance d'entités nommées spatiales dans des textes bruités : vers une cartographie littéraire. Alrahabi, M., Roe, G. Journée d'étude organisée par ObTIC : Analyse spatiale des textes littéraires. <https://obtic.sorbonne-universite.fr/actualite/je-analyse-spatiale-des-textes-litteraires/>
7. **Koudoro-Parfait, C.** (2022). Des formats de données pour mes sources : un casse-tête ?!. Grison, T., Bordier, C. Séminaire doctoral CERES.
8. **Koudoro-Parfait, C.** (2021). Reconnaissance des Entités Nommées (REN) spatiales dans un corpus littéraire : robustesse des systèmes existants ?. Roe, G. Séminaire Humanités Numériques Littéraires, ObTIC.

PARTICIPATIONS AUX SÉMINAIRES DES LABORATOIRES DE RATTACHEMENT

1. **Koudoro-Parfait, C.** (2025). Exploration de la collection Française d'ELTeC avec l'outil de cartographie textuelle Épiméthée. Couffignal, G., Lejeune, G. Séminaire de l'équipe Variation Linguistique et Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.
2. **Koudoro-Parfait, C.** (2024). Annoter un corpus bruité quelles difficultés et perspectives. Couffignal, G., Lejeune, G. Séminaire de l'équipe Variation Linguistique et Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.
3. **Koudoro-Parfait, C.**, Lejeune, G. (2023). Regrouper des Entités Nommées de Lieux issues d'océrisation : Épiméthée de l'OCR à la carte [Mapping spatial named entities from noisy OCR output: Epimethee from OCR to map]. Couffignal, G., Lejeune, G. Séminaire de l'équipe Variation Linguistique et Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.
4. **Koudoro-Parfait, C.** (2023). Groupe de Travail OCR. Couffignal, G., Lejeune, G. Séminaire de l'équipe Variation Linguistique et Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.
5. **Koudoro-Parfait, C.**, Petkovic, L. (2023). Correction automatique des interférences OCR dans la reconnaissance d'entités nommées spatiales : réel gain ou perte de l'information ?. Couffignal, G., Lejeune, G. Séminaire de l'équipe Variation Linguistique et Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.
6. **Koudoro-Parfait, C.** (2023). Cybernetics and Society - The Human Use of Human Beings. N. Wiener, 1950. Club de lecture ML4DH, CERES, SCAI, ObTIC.
7. **Koudoro-Parfait, C.** (2022). Évaluation des résultats de NER sur des transcriptions OCR bruitées par élaboration de supports cartographiques. Millour, A., Lejeune, G. Séminaire de l'équipe de Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.
8. **Koudoro-Parfait, C.** (2021). Reconnaissance d'entités nommées dans des textes de littérature française du 19ème : le poids du bruit ?. Millour, A., Lejeune, G. Séminaire de l'équipe de Linguistique Computationnelle. Laboratoire Sens Texte Informatique Histoire.

Organisation d'évènements scientifiques _____

ORGANISATION DU SÉMINAIRES **NER FOR HISTORICAL DOCUMENTS**

<https://ner-for-historical-docs.github.io/>

Clavaud, F., Graiff, C. (30 mai 2024). NER4Archives : reconnaître, réconcilier et identifier les entités nommées dans les instruments de recherche archivistiques au format EAD. **Koudoro-Parfait, C.**, Chazalon, J., Lejeune, G., Dumenieu, B., Abadie, N., Kermorvant, C. NER for historical documents.

Tarride, S. (28 mars 2024). Extraction d'informations structurées dans des documents historiques. Kermorvant, C., **Koudoro-Parfait, C.**, Chazalon, J., Lejeune, G., Dumenieu, B., Abadie, N. NER for historical documents.

Paquet, T., Coquenot, D., Constum, T. (16 novembre 2023). Approches pour la reconnaissance d'écriture et la reconnaissance d'entités nommées. Chazalon, J., **Koudoro-Parfait, C.**, Lejeune, G., Dumenieu, B., Abadie, N., Kermorvant, C. NER for historical documents.

Poibeu, T. (19 octobre 2023). Analyser les entités dans les romans pour détecter des tendances historiques (grâce au projet BookNLP). Chazalon, J., **Koudoro-Parfait, C.**, Lejeune, G., Dumenieu, B., Abadie, N., Kermorvant, C. NER for historical documents.

Brando Lebas, C., Mélanie-Becquet, F. (17 mars 2023). Annuaire de propriétaires et des propriétés de Paris (1898, 1903, 1913, 1923) : du papier à la carte. **Koudoro-Parfait, C.**, Chazalon, J., Lejeune, G., Dumenieu, B., Abadie, N., Kermorvant, C. NER for historical documents.

Doucet, A., Boros, E. (16 septembre 2022). Impact de la reconnaissance optique des caractères sur la reconnaissance des entités nommées. Lejeune, G., **Koudoro-Parfait, C.**, Chazalon, J., Dumenieu, B., Abadie, N., Kermorvant, C. NER for historical documents.

Abadie, N., Dumenieu, B., Chazalon, J., Kermorvant, C., **Koudoro-Parfait, C.**, Lejeune, G. (7 juillet 2022). Session d'introduction. NER for historical documents.

ORGANISATION D'ATELIERS

Koudoro-Parfait, C., Hernandez, M. (2024). De l'OCR à la visualisation de données. Atelier ObTIC, BnF.

Koudoro-Parfait, C., Petkovic, L. (26-30 juin 2023). Utilisation des techniques de transcription automatique des textes pour l'analyse de documents maritimes dactylographiés du XXe siècle. École d'été d'Analyse Numérique de l'Histoire de la Mer et des Océans (ANHIMO 2023). Institut de l'Océan de l'Alliance Sorbonne Université et Sorbonne Center for Artificial Intelligence (SCAI), Paris (France).

Koudoro-Parfait, C., Tangy, J.-B. (2021). Workshop ROC & NER. Maison de la recherche, Sorbonne Université, Paris.

Formations

ETAL 2025 : École d'été en Traitement Automatique des Langues (1-5 sept. 2025). Lejeune, G., Piwowarski, B., Soulier, L. Roscoff (France).

Outils et méthodes pour la thèse, Humanités numériques : principes, outils et exemples (7 mars 2024). Bibliothèque Universitaire de Jussieu, Sorbonne Université.

Éthique et intégrité scientifique - Ateliers de sensibilisation aux exigences de l'intégrité scientifique (5-12 nov. 2024). Collège doctoral, Sorbonne Université.

Préparer le devenir professionnel des doctorants dans le secteur public comme dans le secteur privé. Service Historique de la Défense à Vincennes (15 mai 2023). Inquimbert, A.-A.

Atelier : Extraction et analyse des réseaux de personnages dans les romans (13 juin 2023). Labatut, V., Amalvy, A., atelier de l'ObTIC et BnF DataLab.

Sciences participatives, sciences citoyennes (19 oct. 2021). Collège doctoral, Sorbonne Université.

Comment passer d'une présentation vers les pairs à une présentation grand public ? (14 sept. 2021). Sorbonne Université.

Modules de formation à la création d'entreprise. Concours Créa Campus : Étude de marché Positionnement, marketing. (2018). Université de Tours.

Atelier de recherche *Objets Poétiques* (2016). Broqua, V. Master de Création Littéraire, Université Paris 8.

Atelier – Rencontre avec des acteurs de l'édition et de l'écriture (2016). Message, V. Master de Création Littéraire, Université Paris 8.

Activités associatives

Adhérente de l'Association pour le Traitement Automatique des Langues, ATALA. 2022 - Présent.

Participation à la création et adhérente de l'association des étudiants en IA de Sorbonne Université, SOPHIA. 2021 - 2025.

Adhérente de l'association du Nombril du monde. 2014 - 2023.

Adhérente de l'association Cent soleils – D'autres images de cinéma. 2019 - 2021.

Outils et ressources développés

ÉPIMÉTHÉE : DE L'OCR À LA CARTE

Épiméthée a été développé dans le cadre des recherches doctorales de Caroline Koudoro-Parfait, en collaboration avec Yoann Dupont (Lattice, Sorbonne Nouvelle) et Marceau Hernandez (CERES, Sorbonne Université).

Épiméthée est un outil conçu pour :

- extraire des entités nommées spatiales à partir de textes numérisés (potentiellement bruité) ;
- appliquer des méthodes d'aide à la prise de décision pour les utilisateur·ice·s afin de lever les ambiguïtés liées à l'identification des entités sur des textes bruités :
 - clustering ;
 - combinaison de plusieurs systèmes de reconnaissance d'entités nommées ;
- proposer une interface interactive permettant de visualiser les clusters d'entités spatiales et d'écarter les faux positifs ;
- offrir une interface conviviale facilitant l'exploration et l'interprétation des résultats.

<https://github.com/These-SCAI2023/EPIMETHEE?tab=readme-ov-file>

CONTRIBUTIONS MÉTHODOLOGIQUES EN HUMANITÉS NUMÉRIQUES

Notebook reprenant la méthodologie pour l'évaluation de l'impact du bruit induit par la reconnaissance optique de caractères sur la reconnaissance d'entités nommées.

https://github.com/These-SCAI2023/EXPE39_DHENS-2023_13092023

Compétences et centres d'intérêt

LANGAGES ET OUTILS DE PROGRAMMATION :

- Python, R
- Matplotlib, Seaborn, Pandas, OpenCv
- PyCharm, Spyder, Jupyter notebook, Rstudio
- Spacy, Stanza, Bert, Flair, CasEn, SEM
- Unitex, TXM, Gephi
- XML-TEI/MEI/EAD, Dublin Core, RDF, RDFs, CIDOC-CRM
- QGIS, Unity, inDesign, Photoshop, illustrator

- $\LaTeX$

INTÉRÊTS DE RECHERCHE

- Variabilité des données
- Applicabilité des systèmes d'intelligence artificielle à des données bruitées
- Conception centrée sur l'utilisateur
- Analyse computationnelle de données historiques et archivistiques en sciences humaines et sociales
- Traitement multimodal de sources textuelles et visuelles (OCR, analyse d'images, traitement automatique du langage – TAL)