Offre de thèse 2026-2029

Machine Learning appliqué aux sceaux byzantins

1 Information

- **Encadrants** :
  - Beatrice Caseau, UMR 8167 Orient et Méditerranée, Sorbonne Université (Prof. en Histoire)
  - Victoria Eyharabide, Laboratoire STIH, Sorbonne Université (MCF en Informatique)
- **Lieu** :
  - Bibliothèque Byzantine, Collège de France - 52 rue du Cardinal-Lemoine, 75005 Paris
  - Maison de la recherche, Sorbonne Université - 28 rue Serpente, 75006 Paris
- **Durée :** 3 ans (date de début souhaitée : janvier 2026)
- **Financement :** Contrat doctoral de l'Initiative Humanités Numériques, Sorbonne Université.
- **Salaire brut mensuel : 2300** €
- **Mots clés** : Réseaux profonds, Reconnaissance de caractères, Segmentation d'instances, Représentation des connaissances, Sigillographie byzantine

2 Méthodologie

Ce contrat inclut une recherche pluridisciplinaire en intelligence artificielle et histoire. Ce projet en humanités numériques est de proposer un programme de machine learning sur la base de photos de sceaux byzantins, afin d'améliorer la lecture des lettres, la segmentation des mots et de reconstituer les principales formules présentes sur les sceaux. Le but ultime est faciliter la lecture des sceaux dont les lettres sont endommagées ou la légende incomplète et ainsi de récupérer des sceaux laissés de côté pour une publication car de lecture trop incertaine. Si nous parvenons à améliorer les lectures, voire à lire des sceaux laissés de côté pour le moment, les informations qu'ils contiennent deviendront accessibles à la communauté scientifique.

Il s'agit de combiner des approches d'apprentissage profond [2] et des outils de traitement automatique du langage (TALN) [1] appliqués aux sceaux byzantin, afin de récupérer intégralement le texte des sceaux malgré leur forme abrégée et leurs caractères parfois endommagés ou manquants. Dans des recherches antérieures [7, 4], nous avons obtenu des transcriptions du verso des sceaux grâce à une approche neuronale en deux étapes, localisant d'abord les caractères par un détecteur d'objets profonds, puis les reconnaissant par un réseau convolutif. Nous prévoyons de poursuivre cette recherche en divisant cette tâche en plusieurs étapes [9]. Nous développerons d'abord une approche bayésienne prédisant les limites des mots (complets ou abrégés) [6, 5]. Ensuite, les hypothèses sur les mots seront éventuellement élargies en utilisant des approches de normalisation de texte et de transformation de traduction automatique. Pour l'entraînement, nous nous appuierons sur des corpus et des dictionnaires grecs [8].

3 Contexte historique

Les sceaux byzantins sont des objets importants pour la connaissance de l'administration et de l'aristocratie byzantine. On estime le nombre de sceaux byzantins découverts en fouilles à environ 80 000. Ce nombre augmente régulièrement, d'environ 1000 à 1500 nouvelles pièces par an. Les sceaux comportent du texte qui permet de reconstruire non seulement la carrière des sigillants mais aussi l'ensemble des fonctions administratives et des postes dans les provinces. C'est de loin la source la plus importante pour établir la prosopographie de l'aristocratie byzantine, puisque tous les dignitaires et les principaux fonctionnaires de l'empire byzantin (qui dure du 4e au 15e siècle) disposaient d'un sceau pour sceller leur correspondance.

Les militaires et les membres du haut clergé avaient aussi leur sceau. Les femmes avaient rarement des sceaux mais les impératrices et les membres de la très haute aristocratie avaient parfois une bulle personnelle.

Figure 1. Un exemple de sceau byzantin (Tatianos hypatos, Cheynet 2019 [3], 5.57, p. 225)

Les sceaux conservés sont très majoritairement en plomb, mesurant entre 20 et 30 mm. Les sceaux en plomb sont des objets circulaires systématiquement bi-faces à partir du 6e siècle qui portent le plus souvent de l'iconographie et toujours du texte grec (on laissera de côté les sceaux dans d'autres alphabets). Il sera possible de travailler par des méthodes de reconnaissance faciale quand l'iconographie comporte la figure des saints. Il y a enfin des monogrammes qui constituent un nom ou un titre sous la forme de lettres entremêlées et il serait utile d'avoir un programme informatique pour proposer des lectures de ces monogrammes.

4 Profil du/de la candidat·e

Les candidat·e·s doivent répondre aux critères suivants :

- Être titulaire d'un **master en informatique** ou équivalent.
- Avoir une forte appétence pour les sciences humaines numériques et l'histoire.
- Maîtriser de manière avancée le langage de programmation **Python** (obligatoire).
- Posséder une solide expérience en **apprentissage automatique et apprentissage profond** appliqués aux images et/ou aux textes, avec l'usage des bibliothèques correspondantes (TensorFlow, PyTorch, etc.).
- Capacité à travailler en équipe pluridisciplinaires (chercheures en histoire, sigillographes, archivistes)
- Notions en graphes de connaissance et ontologies seront appréciées.
- Avoir une **excellente maîtrise de l'anglais** écrit et oral (indispensable).
- Des compétences en communication en **français** constituent un atout mais ne sont pas obligatoires.

5 Modalités de candidature

Les personnes intéressées sont invitées à envoyer un dossier complet par e-mail aux deux encadrants :

- Beatrice Caseau — *beatrice.caseau@sorbonne-universite.fr*
- Victoria Eyharabide — *maria-victoria.eyharabide@sorbonne-universite.fr*

Le dossier doit comprendre :

1. Une **lettre de motivation (1 page max.)** détaillant l'intérêt pour le sujet et les compétences pertinentes pour la thématique proposée.
2. Un **CV complet** (formations, expériences, publications).
3. Un **relevé de notes** (Master).
4. Deux **lettres de recommandation** (*optionnel*).

Dates importantes

- **Date limite pour candidater : 20 octobre 2025**
- **Début souhaité de la thèse : janvier 2026**

5. La bourse est déjà disponible : le/la doctorant·e pourra **commencer immédiatement** si possible.

6 References

[1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases, 2022.

[2] Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. Restoring and attributing ancient texts using deep neural networks. Nature, 603(7900):280–283, 2022.

[3] Jean-Claude Cheynet. Les sceaux byzantins de la collection Yavuz Tatis. Izmir. Privately published, Izmir, 2019.

[4] Victoria Eyharabide, Laurence Likforman-Sulem, Lucia Maria Orlandi, Alexandre Binoux, Theophile Rageau, Qijia Huang, Attilio Fiandrotti, Beatrice Caseau, and Isabelle Bloch (2023). Study of historical Byzantine seal images: the BHAI project for computer-based sigillography. In ICDAR 2023, 7th International Workshop on Historical Document Imaging and Processing (HIP '23). Association for Computing Machinery ACM, New York, NY, USA, 49–54. https://doi.org/10.1145/3604951.3605523

[5] Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. A bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112(1):21–54, 2009.

[6] Shu Okabe, Laurent Besacier, and Francois Yvon. Weakly supervised word segmentation for computational language documentation. In Annual meeting of the Association for Computational Linguistics, 2022.

[7] Théophile Rageau, Laurence Likforman-Sulem, Attilio Fiandrotti, Victoria Eyharabide, Béatrice Caseau, Jean-Claude Cheynet (2025). Character recognition in Byzantine seals with deep neural networks. Journal of Digital Applications in Archaeology and Cultural Heritage, Elsevier. Volume 37, pp e00403, https://doi.org/10.1016/j.daach.2025.e00403

[8] Francois Yvon. Rewriting the orthography of SMS messages. Natural Language Engineering, 16(2):133–159, 2010.

[9] Victoria Eyharabide (2024). Artificial Intelligence Applied to Byzantine Sigillography: Current Research, Challenges, and Future Perspectives. In Special issue "Digital Approaches to Medieval Sigillography," of the journal Digital Medievalist (DM), 17(1): 1–16. https://doi.org/10.16995/dm.15119