Rapido, interopérabilité et fouille de textes : vers un alignement des publications scientifiques en archéologie

Justine Revol[1] Agnieszka Halczuk[2] Lucas Anki[1] Pascal Cuxac[1]
(1) Inist-CNRS (UAR 76), 2 rue Jean-Zay, 54500 Vandœuvre-lès-Nancy, France
(2) Persée (UAR 3602), ENS de Lyon, 15 parvis René Descartes, 69342 Lyon, France
(1) [prénom].[nom]@inist.fr,
(2) [prénom].[nom]@persee.fr

RÉSUMÉ

Le projet RAPIDO vise à enrichir les publications scientifiques en reliant automatiquement les toponymes archéologiques à des référentiels d'autorité grâce à des outils de reconnaissance d'entités nommées. Il s'appuie sur l'annotation manuelle et l'apprentissage automatique (Flair, BERT) pour extraire et aligner ces toponymes. L'article présente cette méthode, les résultats obtenus et les perspectives d'amélioration.

ABSTRACT

Rapido, interoperability and text mining : towards an alignment of scientific publications

The RAPIDO project aims to link archaeological publications with referential using named entity recognition techniques. The approach relies on manual annotation and the training of a model based on Flair and BERT. The algorithm extracts and aligns toponyms from articles published in the Bulletin de Correspondance Hellénique (BCH), calculating a confidence score to validate matches. In this article, we present this processing pipeline, analyze the results of the training phase and discuss potential improvements.

MOTS-CLÉS : entités nommées, référentiels, OCR, fouille de textes, apprentissage automatique.

KEYWORDS: named entity recognition (NER), authority files, OCR, text mining, machine learning.

ARTICLE : **Accepté à** CORIA.

Le développement de NER est un élément important pour l'analyse des corpus textuels en archéologie comme le montrent les travaux de (Brandsen *et al.*, 2020, 2022; Brandsen, 2024). Les travaux de (Vincent, 2024) montrent que l'interopérabilité et la gestion des référentiels restent également des défis importants. Le projet RAPIDO, mené par Persée [1], le réseau EFE [2], l'INIST-CNRS [3] et l'Abes [4], vise à relier automatiquement les publications scientifiques aux données de recherche validées. Il expérimente une chaîne de traitement intégrée pour enrichir les publications des Écoles françaises de Rome et d'Athènes via la plateforme Persée. L'objectif est de faciliter la navigation entre articles et données certifiées, en offrant un accès thématique enrichi aux résultats de la recherche. Le projet mobilise archéologues, éditeurs et professionnels de l'information scientifique.

1. Persée : https://www.persee.fr/
2. Réseau des Ecoles françaises à l'étranger : https://www.resefe.fr/fr/reseau-efe
3. Institut de l'Information Scientifique et Technique : https://www.inist.fr/
4. Agence bibliographique de l'enseignement supérieur : https://abes.fr/

La revue Bulletin de Correspondance Hellénique (BCH), fondée en 1877 par l'École française d'Athènes, a été choisie par les archéologues comme corpus d'apprentissage. Quinze volumes représentatifs ont été sélectionnés, soulignant des difficultés telles que la complexité éditoriale évolutive et la qualité variable de l'OCR selon l'ancienneté des volumes. Le corpus comprend 8432 pages numérisées et océrisées, ainsi que 1457 pages issues de PDF vectoriels. Au total, 7 223 toponymes ont été validés par des archéologues pour garantir la qualité.

Notre modèle de reconnaissance d'entités nommées (NER) est appliqué pour extraire les entités

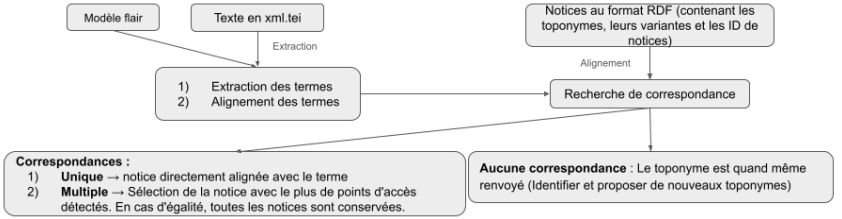

FIGURE 1 – Schéma simplifié de la phase 2

archéologiques qui sont ensuite alignées avec les notices du référentiel selon deux scénarios (figure1) :

— Si une seule notice est associée à une entité, elle est directement retenue.
— Si plusieurs notices correspondent à une même entité, une désambiguïsation est nécessaire.

Dans notre évaluation (Tableau 1), la précision atteint 92%, signifiant que les entités extraites coïncident bien avec les toponymes annotés. Le rappel s'élève à 80%, montrant que la majorité des toponymes présents dans les données sont bien détectés. L'exactitude de 69% montre des résultats plus modérés sur l'ensemble des annotations. Ce résultat s'explique par un nombre important de faux négatifs présents dans des contextes pauvres. Enfin, la F-mesure atteint 85%, traduisant un bon équilibre entre précision et rappel. La méthode est robuste malgré des erreurs d'OCR. Pour évaluer

Métrique	**Valeur**
Rappel (*Recall*)	0.80
Précision (*Precision*)	0.92
F-mesure (*F1-score*)	0.85
Exactitude (*Accuracy*)	0.69

TABLE 1 – Tableau des métriques d'évaluation du modèle pour l'extraction

l'alignement, nous utilisons la métrique de précision qui mesure la justesse de notre modèle. Elle atteint 73%, mais on note que les alignements considérés comme faux sont des propositions multiples de notices et que dans 90% des cas, au moins l'une d'elles est correcte.

Parmi les pistes d'amélioration, une classification multi-labels est envisagée pour enrichir l'extraction en associant les entités à plusieurs types (ex. localisation et archéologie). L'élargissement du périmètre des notices, amorcé avec la Grèce, nécessite un retour d'experts. Le modèle est disponible sur Hugging Face [5]. À terme, la plateforme Persée offrira une navigation enrichie, reliant automatiquement publications et données validées pour faciliter l'exploration et la réutilisation des contenus.

5. https://huggingface.co/INIST-CNRS/rapido

Remerciements

Nous souhaitons exprimer notre profonde gratitude au Fonds National pour la Science Ouverte (FNSO) pour son soutien financier essentiel au projet Rapido. Nous remercions également nos tutelles, le CNRS et l'ENS de Lyon, pour leur accompagnement et leur engagement dans la réussite de cette initiative. Enfin, un immense merci aux directions et équipes impliquées – Abes, Inist, Persée, EFA et EFR – dont l'expertise et la collaboration sont déterminantes pour mener à bien ce projet.

Références

BRANDSEN A. (2024). Archaeology specific BERT models for English, German, and Dutch. Amsterdam. DOI : 10.5281/zenodo.10650835.

BRANDSEN A., VERBERNE S., LAMBERS K. & WANSLEEBEN M. (2022). Can BERT Dig It ? Named Entity Recognition for Information Retrieval in the Archaeology Domain. *J. Comput. Cult. Herit.*, **15**(3), 51 :1–51 :18. DOI : 10.1145/3497842.

BRANDSEN A., VERBERNE S., WANSLEEBEN M. & LAMBERS K. (2020). Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4573–4577, Marseille, France : European Language Resources Association.

VINCENT J. (2024). Les alignements de référentiels. ISSN : 3000-8506, DOI : 10.58079/w1qm.