



HAL
open science

Epimethee – A Workflow from OCR to Spatial Mapping

Caroline Koudoro-Parfait, Marceau Hernandez, Gaël Lejeune, Yoann Dupont

► **To cite this version:**

Caroline Koudoro-Parfait, Marceau Hernandez, Gaël Lejeune, Yoann Dupont. Epimethee – A Workflow from OCR to Spatial Mapping. Document Analysis and Recognition - ICDAR 2025 - 19th International Conference, Wuhan, China, September 16-21, 2025, Proceedings, Part III, Xu-Cheng Yin; Dimosthenis Karatzas; Daniel Lopresti, Sep 2025, Wuhan China, China. <10.1007/978-3-032-04624-6>. <hal-05620669>

HAL Id: hal-05620669

<https://hal.science/hal-05620669v1>

Submitted on 19 May 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-SA 4.0 - Attribution - Non-commercial use - ShareAlike - International License

EPIMETHEE - a workflow from OCR to spatial mapping

Caroline Koudoro-Parfait^{1,2,3}, Marceau Hernandez^{2,4},
Gaël Lejeune^{2,4}, Yoann Dupont⁵

(1) ObTIC, Sorbonne Université, 1 Rue Victor Cousin, 75005 Paris, France

(2) STIH, Sorbonne Université, 1 Rue Victor Cousin, 75005 Paris, France

(3) SCAI, Campus Pierre et Marie Curie, 4 place Jussieu 75005 Paris, France

(4) CERES, Sorbonne Université, 28 rue serpente, 75006, Paris, France

(5) Lattice, UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge, France

`caroline.parfait@sorbonne-universite.fr`

`marceau.hernandez;gael.lejeune@sorbonne-universite.fr`

`yoann.dupont@sorbonne-nouvelle.fr`

Abstract. We present the elaboration of EPIMETHEE, a text-processing pipeline that goes from Optical Character Recognition (OCR) to Named Entity Recognition (NER) and the cartographic representation of places mentioned in ancient literary texts. We will present the difficulties encountered when using off-the-shelf tools for the NER and Map stages in noisy data and the methods used to overcome them. One involves grouping different versions of NEs, for example *Besançon*, *Besangon* or *besanqon*, using a clustering algorithm. We present the assessment for several clustering algorithms. The analysis of spatial NEs with EPIMETHEE lead researchers to better understand, represent and deepen the stakes of a novel by observing the diegetic landscapes proposed by literary authors.

Keywords: Named entity recognition · Optical character recognition · Noisy Documents · Robustness.

References