Umanistica Digitale – ISSN: 2532-8816 – n. 21, 2025
M. Milazzo, G.M. Di Nunzio – *ORNARE: Toward a Digital Methodology For Onomastic Data in Medieval French Romance*
DOI: http://doi.org/10.60923/issn.2532-8816/22288

ORNARE: Toward a Digital Methodology for Onomastic Data in Medieval French Romance

Marta Milazzo

Department of Literary Studies, Philology and Linguistics, Università degli Studi di Milano Statale, Milano, Italy
marta.milazzo@unimi.it

Giorgio Maria Di Nunzio

Department of Information Engineering, Università degli Studi di Padova, Padova, Italy
giorgiomaria.dinunzio@unipd.it

Abstract

ORNARE (Onomastic Repertoire of the *Roman d'Alexandre*) addresses the lack of modern, interoperable tools for identifying, normalizing, and analysing proper names in medieval French romance. After outlining the *status quaestionis* of medieval onomastic resources, the paper explains the motivations behind the project, the criteria for constructing the research corpus, and the challenges of defining the category of *anthroponym* in a medieval context. The project implements a semi-automated pipeline, coupled with expert curation for lemmatisation, disambiguation, and variant grouping. A dual-access web interface supports corpus-level metadata queries and name-centred annotation and search. Applied to a significant corpus (the Old French *Roman d'Alexandre*), ORNARE demonstrates scalable, philologically grounded methods for recovering and visualising dispersed onomastic evidence.

Keyword: medieval onomastics, Romance philology, Roman d'Alexandre, named-entity lemmatization

ORNARE (Onomastic Repertoire of the *Roman d'Alexandre*) affronta la lacuna di strumenti moderni e interoperabili per l'identificazione, la normalizzazione e l'analisi dei nomi propri nel romanzo francese medievale. Dopo una ricognizione sullo *status quaestionis* degli strumenti dedicati all'onomastica letteraria medievale, il contributo illustra le motivazioni del progetto, i criteri adottati per la costruzione del corpus e le difficoltà connesse alla definizione della categoria di *antroponimo* nel contesto medievale. Il progetto realizza una pipeline semi-automatica, integrata con la curatela di esperti, per la lemmatizzazione, la disambiguazione e il raggruppamento delle varianti onomastiche. Una doppia interfaccia web supporta interrogazioni sui metadati a livello di corpus e ricerche/annotazioni centrate sui nomi. Al momento limitato al significativo corpus del *Roman d'Alexandre* antico francese, ORNARE, concepito come metodologicamente modulare e filologicamente rigoroso, rappresenta un innovativo strumento per il recupero e la visualizzazione di evidenze onomastiche.

Parole chiave: onomastica medievale, filologia romanza, Roman d'Alexandre, lemmatizzazione di entità nominate

Introduction

Par le non conuist an l'om — "by the name one knows the man" — was an adage cherished throughout the Middle Ages and embraced by its most accomplished novelist, Chrétien de Troyes. Rooted in classical philosophy and absorbed into Christian thought, the idea presupposes a deep correspondence between *res* and *nomen*. The proverb captures a worldview in which language, thought, and reality were perceived as intimately and necessarily connected. Consequently, the role of the proper name in medieval literature is rather prominent. From Isidore of Seville's *Etymologiae* to Dante's *Commedia*, from Thomas' *Tristan* to Antoine de la Sale's *Salade*, no literary works across the Middle Ages remain untouched by its influence. The name operates on multiple levels. It is a rhetorical device, refined by the *Artes poetriae* [5]. It serves as a narratological tool [12], capable of signalling contrasts or forging links between characters. For the historian of language, the proper name provides valuable evidence. More than other linguistic categories, toponyms and anthroponyms preserve residual and unaltered forms, often conserving traces of ancient material ([32]; [4]). Onomastics thus lies at the intersection of several disciplines: in the fullest sense of a term often misused, it is *interdisciplinary* ([15]:467-602). In the medieval context, this interdisciplinarity extends naturally to history: literary and linguistic aspects are closely intertwined with historical ones, and any comprehensive study must address all these dimensions together. It was in the Middle Ages, from the eleventh century onwards, that the rules governing Western naming were formulated and stabilised ([29]; [17]). Two developments are particularly significant. First, the emergence of the double naming system, combining a given name with a surname. Second, the convergence of personal names into a stable repertoire drawn from the heritage of saints and ancestors. At the same time, personal names began to assume what Caffarelli [6] describes as an "funzione attributiva di appartenenza" ("attributive function of belonging"): like surnames, they marked family descent, relationships, alliances, and more. To bear a name was to bear its meaning and its legacy. Names were earned and transmitted; they were both inheritance and gift.

Given this briefly outlined scenario (examined in greater detail in [25]; [26]), it is striking to note the lack of up-to-date critical tools for identifying and studying names in medieval French literature, particularly considering the prominent role of works in this language. This absence is all the more surprising since the compilation of onomastic indexes was among the earliest priorities of Romance philology. In fact, between the late 19th century and the early 20th, the founders of the discipline devoted considerable attention to onomastics. Focusing only on repertories in Latin and neo-Latin languages, we may recall, in chronological order, Alfred Franklin's *Dictionnaire des noms, surnoms et pseudonymes latins de l'histoire littéraire du Moyen Age (1100-1530)* (1875) and Ernest Langlois's *Table des noms propres de toute nature compris dans les Chansons de geste* (1904), a publication financed by the Académie des inscriptions et des belles-lettres, which in 1899, with the support of Paul Meyer and Gaston Paris, had announced the prix ordinaire with the delivery "relever les noms propres de toute nature qui figurent dans les chansons de geste imprimées antérieures au règne de Charles V" ("to record all kinds of proper names that appear in the printed chansons de geste prior to the reign of Charles V", [19]: V). It may also be important to mention two contemporary onomastic projects that were begun but never completed: the index devoted to Arthurian romances in the doctoral thesis of Fritz Seiffert (1882), a Greifswald student of Eduard Koschwitz, and the vast project of the *Onomasticon Arthurianum* (1898-1905) by Alma Blount, a Harvard student of William Henry Schofield [20].

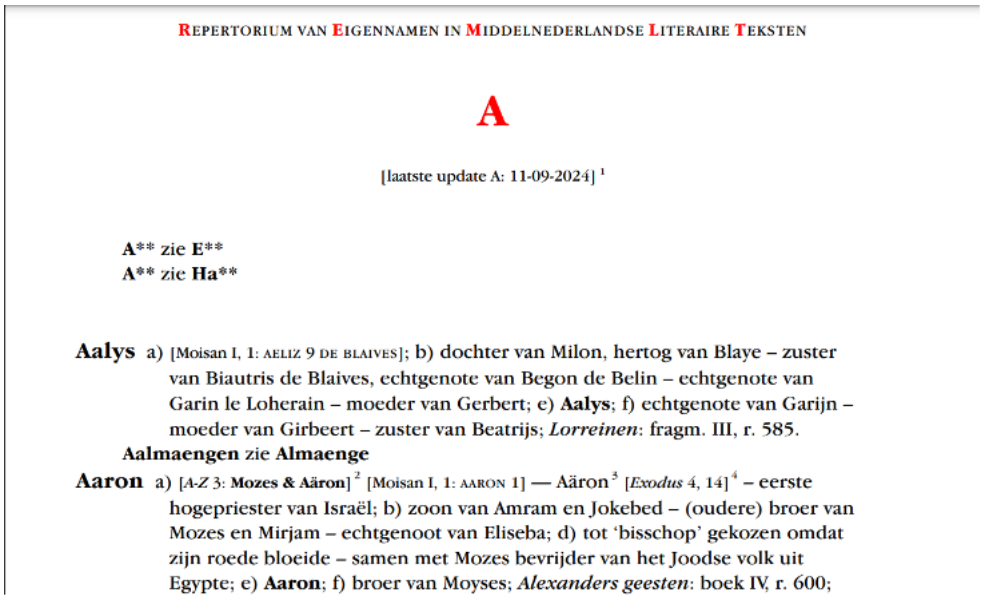
Of both works we possess preparatory material, but no definitive publication. In the same years, the academic world – in Europe and beyond – was working on similar projects. Indeed, there was a strong drive, not only scientific. Like (and perhaps more than) other disciplines, philology was, in the chronological framework in question, "al servizio delle nazioni" ("at the service of nations", [33]).

Leaving aside these broader issues, which fall outside the scope of the present study, it is worth noting that the great momentum of onomastic research suffered a sharp, though inevitable, setback after the First World War. In the decades that followed, several tools were produced, most of which are still in use today. However, they lacked the density that had characterized the *fin-de-siècle* neo-positivist philology. In the Galloromance domain, an onomastic index dedicated to medieval narrative did not appear until 1962: Louis-Ferdinand Flutre's *Table des noms propres avec toutes leurs variantes compris dans les romans français ou provençaux*. Like Langlois's earlier work, this research had been promoted by the Académie, but it was interrupted for nearly twenty years; the *prix ordinaire* awarded to Flutre bears the inauspicious date of 1943. In 1968, it was finally the turn of troubadour literature: Willhelmina M. Wiacek made available a *Lexique des noms géographiques et ethniques dans les poésies des troubadours* (12th-13th centuries). Shortly after, the Canadian scholar Gerald Derrick West made a significant contribution to Arthurian studies with his indexes dedicated to the romances—first in verse (1969), then in prose (1978). West's work was conceived in explicit continuity with similar indexes devoted to Arthurian onomastics on a European scale, ideally taking up the broad multilingual approach of Blount's *Onomasticon Arthurian*. In 1971, Frank M. Chambers updated the *langue d'oc status quaestionis* by publishing an index of proper names in the poems of the troubadours. Finally, in the 1980s, with the discreet (and never fully detailed) help of early IT resources, André Moisan replaced Langlois's *Table* with a new and substantial work devoted to the epic tradition.

Since 1986, no new onomastic indexes have appeared. Even today, anyone searching for a character's name in romances must still turn to Flutre's *Table*, or, in the case of Arthurian texts, to West's *Indexes*. It has been 64 years since Flutre's work, and 57 (for verse) and 48 (for prose) since West's. Both resources are the product of extraordinary and patient dedication, though for different reasons. Flutre provides a comprehensive mapping of names in French and Provençal romances. West, by contrast, offers not just onomastic entries but true encyclopaedic articles, made possible in part by the smaller size of his corpus compared to that of his French counterpart. These remain valuable and reliable tools, yet they are inevitably dated. In the decades since their publication, numerous new editions have appeared. Many works that were once unpublished are now available, and scientifically problematic editions have been replaced by reliable critical texts. Such developments have provided the rationale for creating a new onomastic tool for medieval French romance. This tool is not conceived as an update to previous ones, but as an original work, distinct in both scope and foundations.

This paper presents the methodological foundations and first results of ORNARE (*Onomastic Repertoire of the Roman d'Alexandre*), a digital philology project that applies an innovative approach to the modeling, analysis, and management of onomastic data in medieval French romance ([25]; [26]). The initiative has two main goals. The first is to design and implement an integrated system to support scholarly analysis of medieval onomastic repertoires. The second is to create a new digital resource that combines philological precision with computational interoperability. The project has developed in two phases. The first surveyed existing onomastic resources and scholarly practices to identify key challenges and requirements. The second involved designing and prototyping a digital infrastructure tailored to the needs of Romance philology. While this article reflects on these broader aims, it focuses on the first implementation of ORNARE, dedicated to the *Roman d'Alexandre*. This initial module serves both as a proof of concept and as

a foundation for future expansions. In the following sections, we outline the project's conceptual premises, methodological choices, corpus selection, and data modeling strategy, contributing to the wider discussion on digital approaches to historical onomastics.



REPERTORIUM VAN EIGENNAMEN IN MIDDELNEDERLANDSE LITERAIRE TEKSTEN

A

[laatste update A: 11-09-2024] [1]

A** zie **E****
A** zie **Ha****

Aalys a) [Moisan I, 1: AELIZ 9 DE BLAIVES]; b) dochter van Milon, hertog van Blaye – zuster van Biautris de Blaives, echtgenote van Begon de Belin – echtgenote van Garin le Loherain – moeder van Gerbert; e) **Aalys**; f) echtgenote van Garijn – moeder van Girbeert – zuster van Beatrijs; ***Lorreinen***: fragm. III, r. 585.

Aalmaengen zie **Almaenge**

Aaron a) [*A-Z* 3: **Mozes & Aäron**] [2] [Moisan I, 1: AARON 1] — Aäron [3] [*Exodus* 4, 14] [4] – eerste hogepriester van Israël; b) zoon van Amram en Jokebed – (oudere) broer van Mozes en Mirjam – echtgenoot van Eliseba; d) tot 'bisschop' gekozen omdat zijn roede bloeide – samen met Mozes bevrijder van het Joodse volk uit Egypte; e) **Aaron**; f) broer van Moyses; ***Alexanders geesten***: boek IV, r. 600;

Figure 1 REMLT, letter A (available online at https://bouwstoffen.kantl.be/remlt/A.pdf). The last update, at the time of consultation, was made on 11-09-2024. In bold, the onomastic entries described: Aalys and Aaron.

Onomastics, Romance philology and DHs: a possible triangulation

In recent years, Digital Humanities have become an established part of philological research. Yet, in medieval philology, the field of onomastics shows a noticeable slowdown in digital development. Only two projects stand out in this respect: the *Repertorium van Eigennamen in Middelnederlandse Literaire teksten* (*REMLT*) and the *Diccionario antroponímico del ciclo amadisiano* (*DINAM*). The REMLT was launched in 1992 at the Meertens Instituut KNAW in Amsterdam and is still ongoing under the direction of Willem Kuiper. This long-running project highlights a key consideration: large-scale onomastic research can no longer be treated as a "work of a lifetime", as in West's case, but should be designed as a modular, open-ended undertaking, developed gradually and ideally by collaborative teams.The *REMLT* corpus covers all literary texts in Middle Dutch, according to a chronological span from the 13th century to the second half of the 16th century. The onomastic units analyzed include anthroponymy, toponymy, zoonomy, object names and author names. *REMLT* is conceived as a traditional dictionary: from the website, it is possible to access a .pdf file for each letter of the alphabet (see Figure 1). The onomastic entries are provided with extensive descriptions; often, links connect specific names to web pages external to the project (e.g. Wikipedia). *REMLT* remains in the groove of more

traditional onomastics, but the precision of the indexed entries and the breadth of the corpus make it an excellent tool, scientifically reliable as well as user-friendly.

DINAM adopts a more innovative approach. The project was developed at the University of Zaragoza under the coordination of Maria Coduras Bruna, a leading scholar in Spanish medieval onomastics. Drawing on the Iberian *Amadís de Gaula* corpus, it focuses exclusively on personal names (forenames). The database offers a search interface with multiple, combinable filters (see Figure 2). Each entry contains rich information and relevant hyperlinks, for example connecting related characters. While certain sections, such as the PDF family trees of the *Amadís* cycle, are valuable, the project's real strength lies in its hypertextual networks, which enable interactions between materials that would be difficult to connect in a traditional print format.

Figure 2 The DINAM search interface (available online at https://dinam.unizar.es/). At the top left, a simple search was made, typing the string 'adi'; at the bottom, in blue, the results, divided into nombre (names) and sobrenombres (nicknames).

A further project, this time in historical rather than literary onomastics, is *NordiCon*, a spin-off of the larger *Variation and Contact in Medieval Personal Names*.[1] Looking beyond the literary field can be productive, especially where DH are applied more consistently and innovatively. *NordiCon* maps medieval Nordic forenames attested in sources outside Scandinavia, using a methodology new to historical onomastic lexicography: it combines onomastic analysis with the perspectives of material philology. As in the best DH applications, it links data that would be impossible to represent together using traditional tools. For each lemmatised entry, *NordiCon* records the name (and its referent, where known) with particular attention to its material transmission. The project foregrounds the *material facies* of the names, since all are taken from manuscript sources. Entries include the specific graphic forms (with diplomatic transcriptions) and material details such as erasures or the use of special inks. Each is linked to a reproduction of the manuscript folio where the name appears (fig. 3). *NordiCon* is thus a multi-layered resource that brings together materials otherwise inaccessible in an integrated way, a truly *digital* project, not merely a *digitised* one [34]. Despite an interface that is not immediately intuitive, it stands as a model example, capable of winning over even the most sceptical critics of DH.

[1] https://spraakbanken.gu.se/karp/tng/?mode=nordicon

Figure 3. Nordicon. The folio 87 of the ms. Reichenau, Münsterschatzkammer, Evangeliarium bearing the name Alexander, in the Alexander lemma.

The Onomastic Repertoire Corpus

ORNARE constitutes a first step in the direction of a new onomastic repertory devoted to the medieval French romance (12th-15th centuries). The project was conceived as part of one of the author's doctoral thesis, discussed in the University of Padua in June 2024, entitled "Le nom fu mie sanz raison. *Il nome proprio nel romanzo francese medievale (XII-XV sec.): studi e prolegomeni per il* Nuovo Repertorio Onomastico".[2] This *Repertorio* relates only to anthroponyms and is typologically (textual genre: romance), linguistically (Old French) and chronologically (12th-15th century) delimited. However, subsequent additions, e.g. to other onomastic units (toponyms), such as other linguistic and typological domains (romance in *langue d'oc*) are not excluded, given the modular architecture underlying the research. While some classifications, such as those based on language or chronology, are conventional yet objective, others are more complex and inherently subjective, such as deciding "quale *res* designi il termine romanzo" ("which *res* designates the term romance", [24]:32). Establishing a *regestus* of medieval romances in Old French is therefore a necessary first step in defining the onomastic corpus. This operation raises several critical issues, most of which concern the meaning of the term *roman*. In the 12th–15th centuries, what we now, problematically, call *roman* encompassed a wide range of narrative forms with highly diverse outcomes, many of which have yet to be fully studied. A unified and comprehensive definition of the romance genre remains elusive. We do not intend to engage here with the *status quaestionis* of this debate, nor could we resolve it. Instead, we have chosen to follow the methodological approach recently outlined by Piero Andrea Martina ([23]:7):

> "anziché proporre un quadro teorico più raffinato, si è preferito redigere un repertorio dei testi: questo è volutamente 'largo', con alcuni casi che fungono più da termine di paragone e che non possono essere considerati romanzi in senso stretto."[3]

[2] Tutor: Prof. Giovanni Borriero; Co-tutor: Prof. Giorgio Maria Di Nunzio.

[3] 'Rather than proposing a more refined theoretical framework, we have chosen to compile a repertoire of texts: this is deliberately "broad", including some cases that serve more as points of comparison and cannot be considered romances in the strict sense'.

Therefore, theoretical reflections on medieval genres have been set aside in favour of an empirical, operational approach. The corpus was compiled through the combined and cross-referenced use of the following repertoires:

- *Grundriss der Romanischen Literaturen des Mittelalters* (*GRLMA*), vol. IV, devoted to romances up to the end of the 13th century.
- Brian Woledge's *Bibliographie des romans et des nouvelles françaises antérieurs à 1500* ([38]; supplement 1973).
- *Nouveau Répertoire des mises en proses* (*NR*), edited by Maria Colombo Timelli, Barbara Ferrari, Anne Schoysman, and François Suard [31].
- The "Roman" section of the online *Arlima* repertory.
- The "Bibliographie" section of the online *Dictionnaire Bibliographique de l'Ancien Français* (*DEAF*).

The corpus has been restricted to romances transmitted in manuscript form. This limitation responds, in the first place, to conventional chronological parameters (according to which the medieval period is deemed to conclude in 1492). It is also consistent with a philological approach grounded in the study of manuscript transmission. Moreover, such a restriction prevents the opening of further and potentially unmanageable lines of inquiry, thereby ensuring a coherent and methodologically controlled framework. Only published texts have been included; unpublished works are excluded. Fragmentary works have also been included, from the *Tristan* fragments of Béroul and Thomas to shorter survivals such as the 144 verses of *Vallet à la cote Maltaillié*. All works indexed in *GRLMA* IV have been accepted, with two exceptions: the *Roman du Hem* by Sarrasin (*GRLMA* IV/II 448) and the *Tournoi de Chauvency* ([14]:284). The *tournoi* genre occupies a distinct position within medieval Old French literature. Unlike the *roman*, where tournaments typically serve as episodic elements within broader narrative structures, *tournoi* concentrate on the ceremonial and combative aspects of these events. This is due to their predominantly celebratory, documental, and performative nature, which contrasts with the narrative complexity and fictional scope that characterize the *roman* genre. Scholarly consensus, notably expressed by Jean Charles Payen ([14]:477), tends to exclude *tournois* from the category of medieval romances:

> "Il semble que la description de tournois ait constitué, dès la seconde moitié du xiii[e] siècle, un genre littéraire qui n'a du reste plus grand rapport avec le roman […]. L'auteur d'un «tournoi» assure en fait une sorte de reportage en vers: c'est un témoin qui entend décrire avec exactitude un faste tout théâtral".[4]

For texts from the 14th and 15th centuries, the repertories consulted are Woledge's *Bibliographie* and the *Nouveau Répertoire* (*NR*). Not all texts listed by Woledge have been accepted. Specifically, novellas, such as Laurent Primierfait's translation of the *Decameron* or the *Nouvelles de Sens*, epic prosifications like *Garin de Montglane*, and chronicles such as the *Ancienne Chronique de Pise* have been excluded. From the *NR*, all anomic texts, those that resist precise genre classification and

[4] 'It seems that, from the second half of the thirteenth century onwards, the description of tournaments came to constitute a literary genre in its own right, one that in fact bore little relation to romance. The author of a "tournament" is essentially providing a kind of verse reportage: a witness seeking to describe with accuracy a pageantry that is entirely theatrical in nature'.

are here designated as *mise en prose epico-romanesque*, have been included. In contrast, strictly epic works, as defined in the *NR*'s theoretical framework, are excluded. Although not a repertory in the strict sense, the *Bibliographie* section of the *Dictionnaire Étymologique de l'Ancien Français* (*DEAF*) provides the most comprehensive dataset currently available. For this reason, it has been consistently consulted throughout the research.

The choice of the critical reference edition requires clarification. The onomastic repertoire is constructed from the edited texts of the romances included in the corpus. The criteria guiding the selection of reference editions, however, are not entirely uniform, reflecting the diversity of editorial principles applied to the corpus texts. As a general rule, the most recent reliable critical edition has been preferred. Yet, since one of the goals of the onomastic repertoire is to document the greatest possible number of anthroponomic variants, recency alone has not always been the decisive factor. In particular, in cases where a more recent edition is based on a single manuscript, an older edition based on multiple witnesses with a comprehensive critical apparatus has been preferred. This also applies to editions aimed at non-specialist audiences that are naturally lighter in notes and apparatus, for example, Michael Zink's *Lettres Gothiques* series. Multiple reference editions are indicated when the textual tradition of a romance, such as the *roman de Thebes* or the *Vulgate* cycle, diverges significantly among different redactions. Generally, for works transmitted through a single textual tradition, the most recent reliable critical edition was chosen. In contrast, for texts with multiple witnesses, editions with a comprehensive critical apparatus were preferred.

The resulting corpus comprises 209 romances, represented by a total of 235 critical editions, many of considerable length. Different descriptive criteria may be adopted depending on the specific research objectives, provided they remain appropriate to the goals. This flexibility is especially relevant in the study of medieval anthroponymy. Before the advent of the printing press, and even afterwards, the proper name was subject to significant variation, a complex phenomenon that exceeds the scope of this discussion. Characterized by intrinsic variability far greater than that of common nouns, proper names exhibit diverse *lectiones* that render them valuable *loci critici* in textual transmission. They often deviate from grammatical norms, lack systematic capitalization, and are frequently abbreviated. These features give rise to numerous challenging questions, recently reviewed by Frédéric Duval [11]. Notably, the use of capital letters to mark proper names during the medieval period was by no means systematic. Consequently, distinguishing proper names from common nouns is often far from straightforward. In addition, the distinction in medieval romances is notably complex, due also to the frequent occurrence of combinations of common nouns functioning as proper names. Phrases like *Re Pescatore* or *Dama del Lago* function as pronominal or antonomastic collocations: their necessary (though not always sufficient) condition is a robust and precise invariability that allows them to uniquely identify and single out their referents. These fixed expressions thus qualify as proper names within the textual tradition. However, this clarity often dissolves in cases where it is risky to determine whether a phrase possesses unique referentiality. it can be difficult, if not hazardous, to determine whether a combination like *noir chevalier*, mentioned fleetingly and only once in a tournament scene, should be read as a proper name, comparable to *Blonde Esmeree*, or as a simple descriptive periphrasis denoting a knight clad in dark armor riding a black steed. Given the size of the corpus, a practical approach was necessary. It would be impossible to evaluate every potentially ambiguous personal name to determine its status in every romance. Therefore, the task has been delegated to the individual editors of each text. Accordingly, all personal names capitalized in the *restitutio textus* by the editors have been treated as proper names. While this capitalization criterion is not entirely satisfactory, it constitutes the solution with the fewest negative implications. It remains, after all, a widely shared graphic convention, at least for

anthroponyms. The final distinction between proper and common names is thus left to the discretion of the editors of the individual romances.

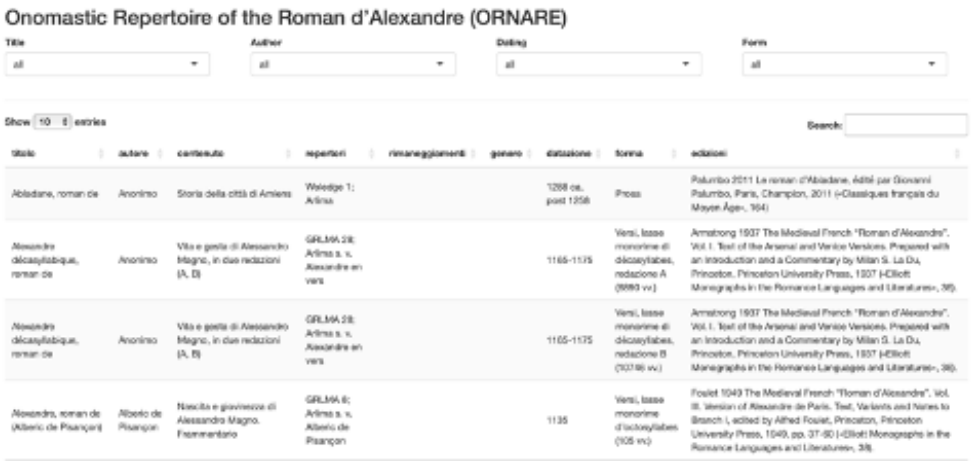


Figure 4 Screenshort of the ORNARE Corpus Web application.

A Digital Record for Preliminary Searches

To facilitate exploratory access to the corpus beyond strictly onomastic analysis, a dedicated web application has been designed and implemented using the R Shiny framework.[5] The platform is available online[6] and serves as a structured digital registry that allows users to navigate and filter the surveyed corpus independently of the onomastic layer. The Web application supports cross-searching through multiple filters and offers a flexible interface for preliminary investigations, as shown in Figure 4. The romances included in the corpus are listed in alphabetical order by title, omitting articles and introductory phrases for consistency. Each work is described through a standardized card, which functions as the minimal unit of the *regestus* ('schedatura'). These cards are concise and uniformly structured, containing key bibliographic and contextual metadata: title, dating, author, content, form, editions, repertoire, genre, and re-arrangements. The *regestus* is thus conceived as a lightweight yet robust access point for scholars, enabling both targeted queries and exploratory browsing within the broader corpus.

1) Title. The title provided corresponds to the conventional one in use. Two titles may be found, if there is more than one vulgate title or if the reference critical edition has a different title, e.g. *Didot-Perceval* (*Perceval en prose*) or Yvain (*Chevalier au lion*).

2) Dating. The dating is inferred from the critical edition. In the case of multiple editions, the dating is taken from the most recent one. If the critical edition is rather old and the dating has been more recently reconsidered, the updated dating is provided in

[5] https://shiny.posit.co/

[6] https://shiny.dei.unipd.it/ornare/schedatura/

parentheses alongside the bibliographic reference where the discussion occurs, e.g., *Barlaam et Josaphat*, Version 'champenoise'; Dating: 1199–1229 [9].

3) Author. The author is inferred from the critical edition. If the author's name conventionally has multiple forms, the name according to the reference edition is given and the name according to the alternative is enclosed in brackets, e.g. Alexandre de Paris (Alexandre de Bernay).

4) Content. A short description of the content of the text is given: e.g. *Bel inconnu*, Content: Adventures of the son of Gawain and the White-Handed Fairy; *Fergus*, Content: Adventures of Fergus, unknown knight who wins a place at the Round Table. It is indicated whether the text is a translation: *Apollonius de Tyr*, Content: Translation of the *Historia Apollonii Regis Tyri*. It is indicated whether the text is fragmentary. It is indicated whether the text is part of a cycle: *Lancelot, Lancelot propre*, Content: *Loves and Adventures*. Lancelot of the Lake. Third romance of the *Vulgate* cycle. All information is taken from the reference edition(s). Concerning redactions and the question of whether different versions of a text constitute distinct works, the *GRLMA* has served as a model. For example, separate records have been created for *Floire et Blancheflor* (popular version and aristocratic version), while only one record exists for the *Estoire del saint Graal*, with further details provided under *Content* as needed.

5) Form. It is specified whether the text is in verse or prose. If the text is in verse, the verse form and number of verses (if available) is specified. The information is taken from the critical reference edition.

6) Editions. Partial editions are indicated only if complete editions are not available. Multiple reference editions are indicated when the textual tradition, such as in the case of the Vulgate, presents significantly different redactions. In such cases, preference is given to acknowledging more than one edition to account for the diversity within the manuscript tradition.

7) Repertories. For each text, the entry number or identifier used in repertories is provided. The repertories are cited in the following order: *GRLMA*, Woledge, *NR*, and finally Arlima. Not all texts appear in all repertories. For Arlima, the entry is explicitly noted only if the text is registered under a different title than the one used here (e.g., *Amis et Amiloun*, Arlima s.v. *Ami et Amile*).

8) Genre. As a general rule, the genre is not specified, since the repertory is assumed to have already defined it. However, in cases of hybrid texts, the collocation *mise en prose epico-romanesque* is employed.

9) Re-arrangements. It is indicated whether the text has undergone reworking, with cross-references to the edition of the reworked text and its record, e.g. a *mise en prose* of Chrétien de Troyes' *Cligés* has been produced (cf. *Alixandre empereur de Constentinoble et de Cliges son filz, le livre de*).

From OCR to a Full Digital Onomastics Tool: The Case Study of The *Roman d'Alexandre*

The first phase of the ORNARE project, presented by Milazzo and Di Nunzio [27] at TPDL 2023,[7] laid the theoretical and technical foundations for the development of a digital onomastic repertoire for medieval French romance. The study focused on the methodological challenges encountered during the construction of the initial dataset, highlighting the limitations of OCR technologies when applied to critical editions, whose paratextual complexity, such as apparatus notes, line numbers, manuscript sigla, etc., generates significant textual noise. A further core issue was the variability and ambiguity of name spellings, which complicates the identification and normalization of anthroponyms, as well as their association with specific narrative characters. Cases of homonymy, uncertain equivalences between variant forms, and the recurrence of similar names across different texts revealed that manual, expert-driven interpretation remains essential in building reliable onomastic entries.

Beyond identifying these challenges, the paper also outlined key functionalities expected by domain experts from a digital tool designed to support onomastic and literary analysis. These functionalities were grouped into two main axes: filtering and browsing romances based on metadata (e.g., subject, form, dating) and querying proper names across their various spellings and narrative contexts. The envisioned interface would enable users to access enriched character profiles, including typological and geographical categorizations, and track orthographic variants through lemmatized entries. This initial work thus provided both the conceptual groundwork and user-centered design requirements for the more targeted application later developed around the *Roman d'Alexandre*, a renowned medieval French romance recounting the legendary exploits of Alexander the Great. Despite its importance, only one edition of the *Alexandre* exists, which is problematic: it is incomplete and lacks an *Index nominum* ([1]; [2]). This limitation was the main reason for choosing such a significant corpus as the prototype for the digital onomastic repertoire. The entire corpus consists of 532 pages and 28,827 verses, which have been scanned and subjected to OCR. Additionally, there are 265 pages of variants and notes on different branches, currently under analysis.

From OCR to Structured Onomastic Data: A Semi-Automated Pipeline

To process the critical editions of the texts in the corpus and extract meaningful data for onomastic analysis, we developed a semi-automated pipeline built entirely in R, using a coordinated set of packages that span from image rendering to text curation and web deployment. This pipeline supports the full transition from scanned page images to structured textual data, and it also underpins the first prototype of the ORNARE web application (see Figure 5).

[7] *Theory and Practice of Digital Libraries*, http://www.tpdl.eu/.

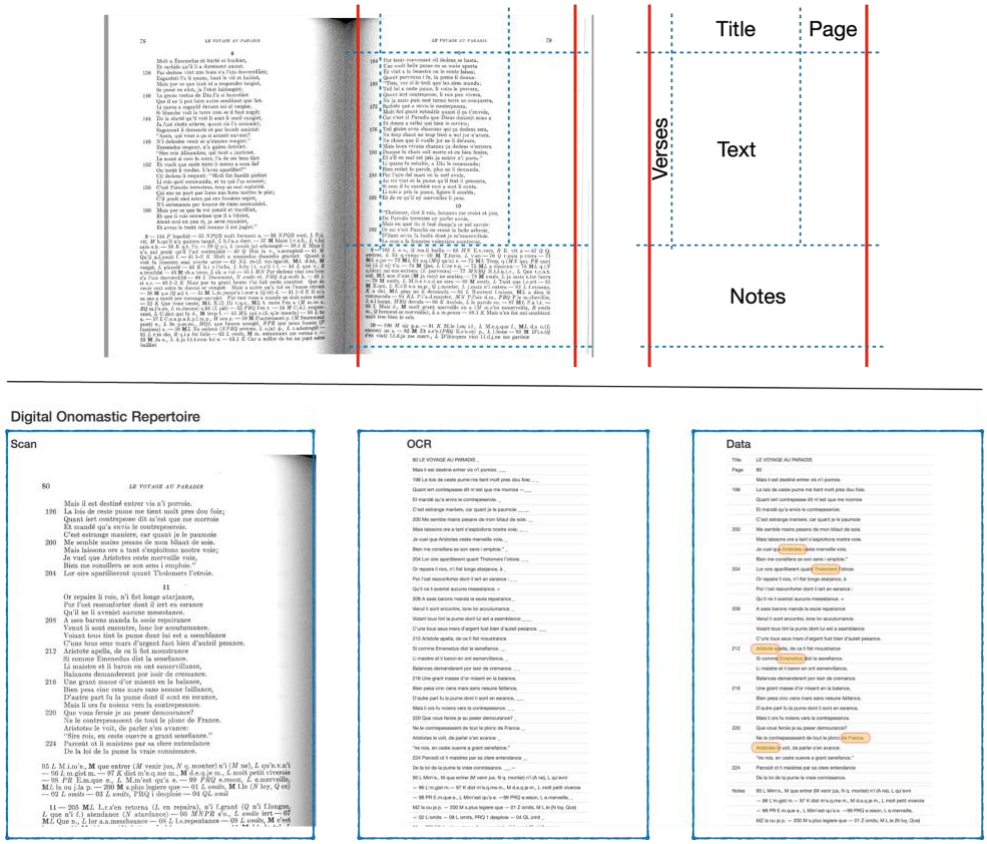


Figure 5 (Top) An example of the organization of a text in the Roman d'Alexandre corpus. On the left side of the figure, we highlighted the (right) page of a book (red line) and the different parts of the text (dashed blue line). On the left side of the figure, the meaning of the different parts of the text that we need to process. (Bottom) A screenshot of the Web application during the process of the conversion of the pdf (right) to OCRed text (center) to structured data (right).

The process begins with scanned editions in PDF format. Using the pdftools package,[8] each page is rendered and subsequently converted into a bitmap image via png[9] and magick,[10] allowing for the detection of spatial zones such as titles, line numbers, main text, and critical notes. These geometric boundaries are crucial for isolating relevant content before applying OCR. Text recognition is handled by tesseract,[11] after which the results undergo a series of cleaning and

[8] https://doi.org/10.32614/CRAN.package.pdftools

[9] https://doi.org/10.32614/CRAN.package.png

[10] https://doi.org/10.32614/CRAN.package.magick

[11] https://doi.org/10.32614/CRAN.package.tesseract

transformation operations using tidyverse[12] tools to extract and normalize the desired textual segments.

Once OCR and geometric parsing are complete, the enriched data enters a validation phase supported by an R Shiny application. This interface displays, side by side, the original scan, the OCRed text, and the resulting structured data in tabular form. Users can inspect and interact with the extracted content, particularly focusing on the detection of proper names. Preliminary automatic identification highlights potential anthroponyms (e.g., Aristotes, Aristote, Emenedus, France, Tholomers), which may represent either spelling variants or distinct entities—cases that require scholarly judgment.

A crucial phase of the pipeline thus involves expert curation: human intervention is needed to

resolve ambiguities, validate names, group variants, assign typological and geographic labels (e.g., category: "king", "philosopher"; location: "France", "Africa"), and establish links across the corpus. The goal is to produce a rich, searchable set of records that preserve both the textual specificity and the interpretative layers essential to onomastic research. This curated dataset, which captures both occurrences and their contextual attributes, forms the backbone of the ORNARE repository and is currently being expanded in tandem with the ongoing development of the application's backend.

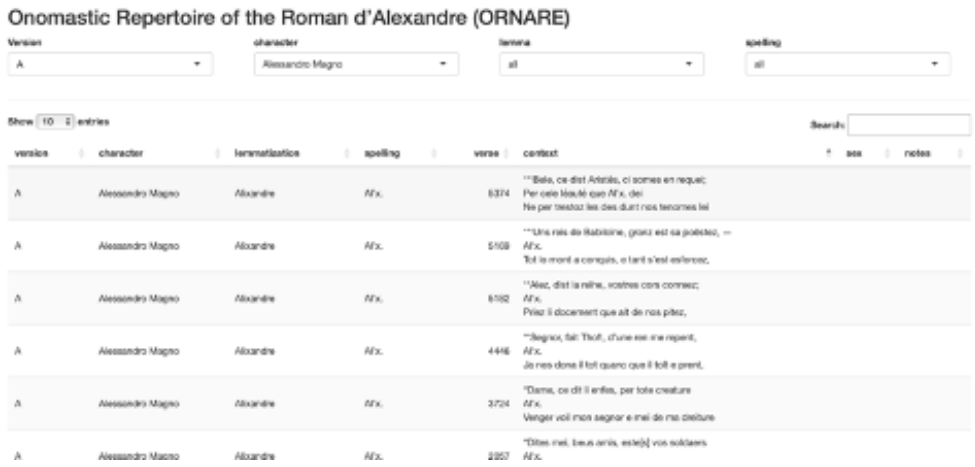


Figure 6 https://shiny.dei.unipd.it/ornare/indice/. Screenshot of the ORNARE web application interface. The main panel displays the selected edition of the Roman d'Alexandre, with options to search for normalized names, specific spellings, or character roles. Users can interactively annotate, filter, and edit verses. The layout facilitates both philological precision and exploratory research.

ORNARE: A Web Environment for Onomastic Analysis in the Roman d'Alexandre

The ORNARE system constitutes the core digital application developed to support the analysis and annotation of proper names in the *Roman d'Alexandre* ([27];[28]). It represents the first

[12] https://doi.org/10.32614/CRAN.package.tidyverse

operational outcome of the broader project, offering an integrated interface for managing the onomastic data extracted from the corpus.

The underlying dataset includes approximately 1,500 pages of critical editions and textual variants, amounting to over 33,000 verses. An additional 180 pages of variant readings are currently being manually redacted and incorporated into the digital environment. As part of the ongoing annotation effort, we annotated:

- 3 versions of the *Roman d'Alexandre*;
- 8,535 annotations overall;
- 1,336 distinct spellings of names were identified;
- 270 characters were distinguished.

However, determining the precise formalized forms of names remains a philological challenge, due to the presence of homonyms and ambiguous variants that require expert validation.

The system was developed using the R Shiny framework, extending a previous prototype [27], and integrates additional packages such as DT[13] for interactive tables and textplot[14] for text-centered visualizations. The interface is designed for ease of use by linguists and literary scholars: users can load or write text, perform targeted searches for names (whether normalized forms like "Alixandre" or specific spellings like "Al'x"), and filter results according to various criteria, such as character titles (e.g., "King of...") or geographic references.

One of ORNARE key features is the ability to update onomastic annotations directly within the interface. Once a search has been performed, users can apply modifications or additions to the selected verses, linking occurrences to a particular character entry and assigning attributes such as category (e.g., "Saracen", "Queen", "Philosopher") or geographical origin. All modifications are stored in a structured format, facilitating downstream research and data export.[15]

The web application also allows for the on-the-fly correction and refinement of the source text, which is essential given the philological complexity of the corpus. Experts can edit OCR output, resolve variant forms, and insert interpretive notes. This modular and iterative design makes the system a versatile instrument for onomastic research, while also serving broader investigations into linguistic variation, genre conventions, and medieval intertextuality.

The source code is publicly maintained and updated, ensuring transparency and reproducibility. As the project advances, ORNARE is intended not only as a tool for scholarly annotation but also as a model for future digital repertories of medieval literature.

[13] https://cran.r-project.org/web/packages/DT/

[14] https://cran.r-project.org/web/packages/textplot/

[15] At the time of this article's writing, the interfaces accessible to the public do not provide functionalities for direct annotation or modification of the data. This limitation has been deliberately implemented to prevent alterations by non-experts. Future developments will include administrator-level authentication, thereby ensuring a clear separation of functionalities.

Conclusion and Future Works

This paper has presented the first complete implementation of ORNARE, a digital onomastic repertoire conceived as a scalable and interoperable model for the study of names in medieval French romance. Building on earlier methodological investigations and prototype development, we have described a pipeline for the semi-automated extraction, curation, and annotation of names, as well as the development of a dual-access web application: one registry-oriented, for metadata queries across romances, and one onomastic, for querying and annotating character names and variants. The challenges addressed, from OCR processing of paratext-rich editions to the lemmatization and disambiguation of variant spellings, underscore the uniqueness of medieval onomastic data and the necessity for expert-guided, hybrid approaches. ORNARE not only facilitates the recovery and connection of dispersed data across a complex literary tradition, but also enables modelling and visualization of this data in ways previously inaccessible to scholars.

Nonetheless, the project currently faces important limitations that warrant further attention. The annotation of character identities remains incomplete, particularly in cases involving ambiguous or overlapping references. Disambiguation strategies, though methodologically grounded, require refinement to handle the full spectrum of orthographic and contextual variants found in the corpus. Moreover, while initial steps have been taken towards aligning the dataset with CLARIN-compatible ontologies and LOD serialization by means of OntoLex-Lemon,[16] full semantic interoperability has yet to be achieved, limiting the project's integration within broader linguistic infrastructures. Future developments will focus on addressing these issues by completing the annotation process, enhancing disambiguation algorithms with machine-assisted and expert feedback loops, and advancing the semantic alignment of data. Such improvements will ensure not only the interoperability and long-term sustainability of ORNARE but also strengthen its capacity as a research environment that upholds philological rigor while opening new horizons for the digital exploration of names, genres, and imaginaries in medieval narrative.

Future developments may also profit from the ongoing evolution of digital philology infrastructures and models. Recent work has emphasised the need for interoperable frameworks that combine textual scholarship, annotation, and visualization [16], suggesting that onomastic data could be integrated within broader environments of textual literacy and data-driven commentary. Likewise, the move towards ontology-based editorial models [21] opens the possibility of aligning onomastic resources with conceptual architectures designed for digital scholarly editions, thereby reinforcing methodological coherence between textual and onomastic domains. Finally, the emergence of catalogues and repositories of philological tools [22] indicates the value of mapping and connecting existing infrastructures to ensure reusability, transparency, and long-term maintenance. In this perspective, ORNARE could serve as both a use case and a testing ground for extending digital philology practices to the structured management of names—an area where interoperability, annotation, and visualization converge to redefine how we model and interpret medieval cultural networks.

[16] https://www.w3.org/2019/09/lexicog/

Riferimenti

[1] Alberic de Pisançon ed. Foulet 1949. *The Medieval French "Roman d'Alexandre"*. Vol. III, Version of Alexandre de Paris. Text, Variants and Notes to Branch I. Edited by Alfred Foulet. Princeton: Princeton University Press, 1949, pp. 37–60.

[2] Alexandre de Paris ed. Armstrong 1937. *The Medieval French "Roman d'Alexandre"*. Vol. II. Version of Alexandre de Paris. Edited by E. C. Armstrong, D[ouglas] L. Buffum, Bateman Edwards, L[awrence] F. H. Lowe, Princeton – Paris, Princeton University Press – Presses Universitaires de France, 1937.

[3] *ARLIMA – Archives de littérature du Moyen Âge (ARLIMA)*. https://www.arlima.net/

[4] Beltrami, Pietro 1999. "Il Tesoro della lingua italiana delle origini (TLIO) e l'onomastica". R*ivista Italiana di Onomastica* 5 (2): 349-362.

[5] Bianchini, Simona. 2002. "Interpretatio nominis e pronominatio nel Cligés." *Vox Romanica* 61: 181–221.

[6] Caffarelli, Enzo. 2023. "Le funzioni del nome personale." In *Onomastica: un mondo da scoprire*, Treccani Magazine, Sezione Lingua Italiana, 30 giugno 2023. Accessed 20 may 2025. https://www.treccani.it/magazine/lingua_italiana/articoli/parole/funzioni_nome_personale.html.

[7] Chambers, Frank M. 1971. *Proper Names in the Lyrics of the Trobadours*. Chapel Hill: University of North Carolina Press.

[8] Coduras Bruna, M., dir. *Diccionario de nombres del ciclo amadisiano (DINAM)*. Universidad de Zaragoza. Accessed 30 may 2025. https://dinam.unizar.es/

[9] Cordoni, Constanza und Matthias Meyer, eds. 2015. *Barlaam und Josaphat. Neue Perspektiven auf ein europäisches Phänomen. Berlin-Boston: De Gruyter*

[10] DEAFBiblEL. Möhren, Frankwalt. *Complément bibliographique. Frankwalt Möhren - Elektronische Fassung [aktualisiert am 2024-09-04]*. https://alma.hadw-bw.de/deafbibl/

[11] Duval, Frédéric. 2021. "Éditer les noms propres". In Ferlampin-Acher, Christine, Fabienne Pomel et Emese Egedi-Kovács, éd. 2021. *Par le non conuist an l'ome. Études d'onomastique littéraire médiévale*, 61-89. Budapest: Collège Eötvös Jozsef ELTE. Accessed https://mek.oszk.hu/23000/23056/23056.pdf

[12] Ferlampin-Acher, Christine, Fabienne Pomel et Emese Egedi-Kovács, éd. 2021. *Par le non conuist an l'ome. Études d'onomastique littéraire médiévale*. Budapest: Collège Eötvös Jozsef ELTE. Accessed 12 august 2025. https://mek.oszk.hu/23000/23056/23056.pdf

[13] Franklin, Alfred. 1875. *Dictionnaire des noms, surnoms et pseudonymes latins de l'histoire littéraire du Moyen Âge (1100–1530)*. Paris: Firmin Didot.

[14] GRLMA/IV. Frappier, Jean, and Reinhold R. Grimm, eds. 1978–1984. *Grundriss der Romanischen Literaturen des Mittelalters. Vol. IV. Le roman jusqu'à la fin du XIIIe siècle.* Heidelberg: Winter Universitätsverlag.

[15] Hough, Carole, ed., with assistance from Daria Izdebska 2016. *The Oxford Handbook of names and naming*. Oxford: Oxford University press. 10.1093/oxfordhb/9780199656431.001.0001

[16] Italia, P. (2025). Edizione, Annotazione, Visualizzazione. Problemi e prospettive della/per la filologia digitale. Umanistica Digitale, 9(20), 189–219. https://doi.org/10.6092/issn.2532-8816/21183

[17] Klapisch-Zuber, Christiane. 1996. "Quel Moyen Âge pour le nom?" In *L'anthroponymie. Document de l'histoire sociale des mondes méditerranéens médiévaux*. Actes du colloque international organisé par l'École française de Rome, avec le concours du GDR 955 du CNRS 'Genèse médiévale de l'anthroponymie moderne', éd. Monique Bourin, Jean-Marie Martin et François Menant, 473–80. Rome: Publications de l'École française de Rome.

[18] Kuiper, W., H. Hendriks, and S. Koetsier, eds. *Repertorium van Eigennamen in Middelnederlandse Literaire teksten (REMLT) – Répertoire des noms propres dans des textes littéraires en Moyen Néerlandais*. Accessed 12 august 2025 https://bouwstoffen.kantl.be/remlt/Inleiding.pdf/

[19] Langlois, Ernest. 1974. *Table des noms propres compris dans les Chansons de geste*. Genève: Slatkine Reprint. Original Edition Paris 1904.

[20] Magoun, Francis Peabody, Jr. 1926. "An Index of Abbreviations in Miss Alma Blount's Unpublished Onomasticon Arthurianum." *Speculum* 1 (2): 190–216. https://doi.org/10.2307/2847545

[21] Martignano, C. (2024). *Critical Edition Ontology: a conceptual model for digital critical editions*. Umanistica Digitale, 8(17), 71–94. https://doi.org/10.6092/issn.2532-8816/19469

[22] Martignano, C. (2025). *A catalogue of software tools for digital scholarly editing*. Umanistica Digitale, 9(19), 1–15. https://doi.org/10.6092/issn.2532-8816/21093

[23] Martina, Piero Andrea. 2021. *Il romanzo francese in versi e la sua produzione manoscritta*. Strasbourg: EliPhi.

[24] Meneghetti, Maria Luisa. 2010. *Il romanzo nel medioevo. Francia, Spagna, Italia*. Bologna: il Mulino.

[25] Milazzo, Marta. 2023. "Il nome proprio nella letteratura romanza francese medievale. Per uno stato dell'arte ragionato". *Critica del testo* 16 (3): 129-172. https://doi.org/10.23744/5523

[26] Milazzo, Marta. 2024. "The Onomasticon Arthurianum (et similia). State of the art of a chimera". *Journal of the International Arthurian Society*, 12: 111-133. https://doi.org/10.1515/jias-2024-0005

[27] Milazzo, Marta, and Giorgio Maria Di Nunzio. 2023. "The First Tile for the Digital Onomastic Repertoire of the French Medieval Romance: Problems and Perspectives." In *Linking Theory and Practice of Digital Libraries*, edited by Omar Alonso, Helena Cousijn, Gianmaria Silvello, Mónica Marrero, Carla Teixeira Lopes, and Stefano Marchesin, 317–23. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43849-3_29.

[28] Milazzo, Marta, and Giorgio Maria Di Nunzio. 2024. "The Onomastic Repertoire of the Roman d'Alexandre (ORNARE). Designing an Integrated Digital Onomastic Tool for Medieval French Romance." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 15982–87. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1389/.

[29] Mitterauer, Michael. 2001. *Antenati e santi. L'imposizione del nome nella storia europea*. Trad. Teresa Franzosi. Torino: Einaudi. Original Edition: München 1993.

[30] Moisan, André. 1986. *Répertoire des noms propres de personnes et de lieux cités dans les chansons de geste françaises et les œuvres étrangères dérivées*. 5 vols. Genève: Droz.

[31] *NR.* Colombo Timelli, Maria, Barbara Ferrari, Anne Schoysman, and François Suard. 2014. *Nouveau Répertoire des mises en prose (XIV–XVI siècles)*. Paris: Garnier.

[32] Pfister, Max 1999. "L'importanza della toponomastica per la storia della lingua nella Galloromania e nell'Italoromania". *Rivista Italiana di Onomastica*, 5 (2): 449-464.

[33] Rapisarda, Stefano 2018. *La filologia al servizio delle nazioni. Storia, crisi e prospettive della filologia romanza*. Milan: Mondadori.

[34] Sahle, Patrick 2016. "What Is a Scholarly Digital Edition?". *Digital Scholarly Editing*, édité par Matthew James Driscoll et Elena Pierazzo, Open Book Publishers, https://books.openedition.org/obp/3397.

[35] West, Gerald D. 1969. *An Index of Proper Names in French Arthurian Verse Romances 1150–1300*. Toronto: University of Toronto Press.

[36] West, Gerald D. 1978. *An Index of Proper Names in French Arthurian Prose Romances*. Toronto: University of Toronto Press.

[37] Wiacek, Wilhelmina. 1968. *Lexique des noms géographiques et ethniques dans les poésies des troubadours des XIIe et XIIIe siècles*. Paris: Nizet.

[38] Woledge, Brian. 1954. *Bibliographie des romans et des nouvelles françaises antérieurs à 1500*. Genève-Lille: Droz – Giard.

[39] Woledge, Brian. 1975. *Bibliographie des romans et nouvelles en prose française antérieurs à 1500*. Supplement 1954–1973. Genève: Droz.