Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora

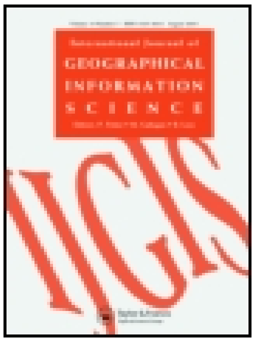
Article *in* International Journal of Geographical Information Science · May 2019

DOI: 10.1080/13658816.2019.1620235

3 authors, including:

Katherine McDonough
Stanford University
13 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Ludovic Moncla
Institut National des Sciences Appliquées de Lyon
34 PUBLICATIONS **195** CITATIONS
SEE PROFILE



International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora

Katherine McDonough, Ludovic Moncla & Matje van de Camp

To cite this article: Katherine McDonough, Ludovic Moncla & Matje van de Camp (2019): Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2019.1620235

To link to this article: https://doi.org/10.1080/13658816.2019.1620235

Published online: 27 May 2019.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2019.1620235

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora

Katherine McDonough[a,b], Ludovic Moncla[c] and Matje van de Camp[d]

[a]The Alan Turing Institute, London, UK; [b]Department of History, Queen Mary University of London, London, UK; [c]INSA Lyon, CNRS, LIRIS UMR 5205, France; [d]De Taalmonsters, Tilburg, The Netherlands

ABSTRACT
Geographic text analysis (GTA) research in the digital humanities has focused on projects analyzing modern English-language corpora. These projects depend on temporally specific lexicons and gazetteers that enable place name identification and georesolution. Scholars working on the early modern period (1400–1800) lack temporally appropriate geoparsers and gazetteers and have been reliant on general purpose linked open data services like Geonames. These anachronistic resources introduce significant information retrieval and ethical challenges for early modernists. Using the geography entries of the canonical eighteenth-century *Encyclopédie*, we evaluate rule-based named entity recognition (NER) systems to pinpoint areas where they would benefit from adjustments for processing historical corpora. As we demonstrate, annotating nested and extended place information is one way to improve early modern GTA. Working with Enlightenment sources also motivates a critique of the landscape of digital geospatial data.

ARTICLE HISTORY
Received 14 May 2018
Accepted 14 May 2019

KEYWORDS
digital history; ethics of GIS; early modern history; geographic information retrieval

1. Introduction

Bruzen de la Martinière's *Le Grand Dictionnaire Geographique et Critique* (1726–1739) and Saugrain's *Dictionnaire universel de la France ancienne et moderne, et de la Nouvelle France* (1726): these are but two of many efforts to connect human knowledge and geography during the eighteenth century. Gathered from scholarship, travel accounts, scientific notes, and vernacular tradition, they are invaluable collations of extant geographic knowledge in Europe. They represent early examples of dictionaries and encyclopedias that chose to organize information geographically. Even when mixed with other classification schemes, geography operates as an anchor in Enlightenment work. Knowing how geographical information was applied should, therefore, be at the heart of encyclopedic analysis. And yet, this aspect of early modern reference works has escaped the attention of scholars working on information culture. This is understandable since most of these works have until recently been difficult to access collectively (because of their dispersion in rare books collections) and read closely and comparatively (because of their structure and length).

CONTACT Katherine McDonough kmcdonough@turing.ac.uk

We use the *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres* (1751–1772) as a model for a critical approach to such early modern geographic knowledge. This project geoparses pre-nineteenth-century French text, new territory for the young field of Geographic Text Analysis (GTA). GTA combines Natural Language Processing (NLP) and Geographic Information Science (GIS) to enable humanities researchers to extract and model spatial information from text sources. More specifically, geoparsing combines Named Entity Recognition (NER), which automatically identifies place names in a text (either using rule-based or algorithmic methods), with georesolution, which uses a gazetteer lookup to match an identified place entity with a location. (A gazetteer is an index of place names with associated metadata, including geospatial coordinates, variant names, etc.)

This experiment evaluates failures in NER rather than successes in georesolution. We intentionally use tools that have not been prepared specifically for this context, a practice that is becoming more common as humanities scholars apply out-of-the-box geoparsers to historical materials. This article functions as a warning light for humanities scholars interested in these methods: there is no such thing as context-agnostic tools for GTA.

Geoparsers used in early GTA research were designed for post-eighteenth-century English corpora and have poor results when used outside of that context. Understanding the sources of geoparser error will generate recommendations for geoparser construction and make GTA research more robust. Such an investigation requires evaluating NER performance on historical corpora in a variety of languages and from different periods. We model this process here for early modern French texts. Based on a comparison with a subset of manually annotated *Encyclopédie* (ENC) articles, our NER evaluation highlights two core areas for improvement. One of these is geoparsers' dependency on anachronistic gazetteers. Another is their limited ability to identify complex or vague spatial information (especially in non-English languages). This article addresses both challenges by briefly unpacking the history of gazetteers and experimenting with a new process of capturing place names alongside useful contextual information. While the latter geographic information retrieval task can be fairly language-dependent, the general idea of accounting for contextual information is not. Likewise, the first issue regarding gazetteers applies across all languages.

Our tests use two relatively unaltered systems akin to what scholars who are not equipped to customize a geoparser would encounter. We compare the Edinburgh Geoparser (EG) (Grover *et al.* 2010, Tobin *et al.* 2010, Grover and Tobin 2014, Alex *et al.* 2015) with Perdido (Gaio and Moncla 2017). EG is a standard geoparser in GTA research, designed for running text in modern English. Perdido was designed for the same style of text in modern French. With this approach, we gained valuable information to guide future adaptations of Perdido by thinking through what worked, what did not work, and why. We stress that the evaluation of EG in tandem with Perdido is not meant to slight EG's performance metrics when used with modern English corpora. We study it alongside Perdido not to criticize its functionality, but to point out to humanities scholars how different geoparsers can be and why customizing them is a core concern.

The ENC[1] is an ideal pivot point to examine geographic information before and after the Enlightenment. With 20.7 million words in 44,632 entries by 134 authors, the ENC represents the culmination of reference publication culture in early modern France.

Geography-classified entries function as palimpsests of place attestations. Edited by Diderot and d'Alembert, the first edition was printed in 17 volumes of text and 11 volumes of plates (Morrissey and Roe Spring 2016 Edition).[2]

What kind of language do ENC authors use to describe places? What is considered to be a useful context for describing a place and what does this tell us about editorial goals for place-name articles? What does the geography of the ENC tell us about the representation of non-European places before and during the Enlightenment? Refining NER will allow us to eventually add references to places from attestations in other early modern geographical reference sources. With this larger corpus, we will study textual representations of geographic knowledge within early modern Europe, in relation to contemporary non-European texts, and, finally, in relation to twenty-first-century digital resources.[3]

Just as we look backward from the eighteenth century through the lens of ENC entries, we must also look forward to consider the ENC as a distant relative of Wikipedia. The questions we ask of Enlightenment texts should be the same questions we ask of Geonames, the Alexandria Digital Library (ADL), or Wikipedia–all commonly used tools in spatial humanities projects–in relation to their production. Who compiled this knowledge? Why? From what sources? What is excluded? What is the relationship between information about the place and that of people, ideas, or events? By adopting this mindset, we prepare a foundation for examining early modern geographical texts and linking these artifacts to their digital descendants. Existing digital resources suffer from some of the same faults as their early print kin – a colonialist perspective of the world foremost – because they usually replicate their predecessors' content. Geoparsers for spatial humanities research on the pre-modern era should account for these biases.

Adapting a geoparser for French will allow us to explore the structure and content of early modern geographical information computationally. It also allows us to analyze techniques for geographic information retrieval challenges common across historical languages: identifying place name variants, associating name variants of the same place, disambiguating different places, and determining what types of relationships exist between gazetteer records and textual attestations of historical places.

We, therefore, address two gaps in NLP research: 1) working with low-resource historical languages like French and 2) working with complex textual structures that combine running text with lists of place names. (Santos *et al.* 2015) demonstrate the challenges that remain in georesolution of historical toponyms. Working primarily with itinerary, or list-style, data where entities do not need to be sorted from the surrounding text, georesolution could be isolated instead as the key research problem. However, we take an NER-centric approach. This allows us to first critique resources like ADL which geoparsers like EG and Perdido depend on to identify place entities.

Anachronistic uses of gazetteers and lexicons in projects identifying places in other centuries constitutes what we define as 'temporal dissonance,' or the mis-application of geographic information explicitly or implicitly documenting a particular time and place to another period and not necessarily the same place. Temporal dissonance occurs in relation to individual place histories as well as overall resource coverage: modern gazetteers contain records that are not historically relevant and are also embedded in digital resources that represent only some parts of the world from the perspective of foreigners. The problem of temporal dissonance is that it ignores historical specificities such as name variation. While ancient Lutèce (the French form of Lutetia, which is the name of the Roman city on the site of modern Paris) and modern Paris may occupy relatively the same space, a gazetteer that only contains the name for one of these (or privileges this during lookups) will fail to identify mentions of Lutèce as a place entity. So while general, non-scholarly gazetteers have been adequate in terms of some georesolution tasks, they are not sufficient for entity recognition.

To model spatial information in historical collections of texts, we need temporally appropriate, attestation-based gazetteers that are often too time-intensive to build manually. NER can be used to extract information for such gazetteer records. However, as we have just seen, in order to have trustworthy rule-based NER systems, we need gazetteers whose content and structure are culturally and historically sensitive, for example, leaving some places 'unlocated,' and instead documenting spatial or social relations. By assigning high value to geospatial coordinates as the key piece of metadata in gazetteers that transforms a place name into digital spatial information, we have ignored other opportunities for annotating location. Confidence in NER results on historical projects would be improved by leveraging gazetteers that handle change over time and spatial relationships. In an attempt to cut through this circular logic linking NER and gazetteers, our research reassesses the process of place entity identification during NER and offers alternative solutions to modeling complex, non-English place references.

2. Review of literature

2.1. Enlightenment spatial histories

We build on recent work on the spaces and places of the Enlightenment as well as digital humanities projects about this period. Studies on the print culture of geography during the Enlightenment focus on the constitution of the discipline of geography (Withers 2008, Mayhew 2010, Withers and Mayhew 2011) rather than he practical origins of geographical information during the early modern period. (Withers 1999) and (Safier 2014) are exceptions to this trend. Furthermore, Enlightenment DH has often focused on the geographies of small groups of elites (political leaders, scientists, philosophes, travelers) (Edelstein 2016, Comsa *et al*. 2016). In contrast, our research is a step towards understanding the mobility of geographic information across local and cosmopolitan, expert and non-expert communities.

Although it is an important example of spatial historical thinking, the heterogeneous ENC makes no pretensions towards exhaustive or exclusive coverage of the entire body of geographic knowledge. In November 1750, Diderot first suggested that the general organization of the ENC could be likened to a *mappemonde* or a map of the world. In June 1751, d'Alembert compared the work to 'a kind of Mappemonde' made up of many more detailed maps: 'These detailed maps will be the different articles of our *Encyclopédie*, and the tree or *système figuré* will be the mappemonde.'[4] Relying on the mappemonde as a source of exhaustive knowledge about a topic is, d'Alembert warned, misguided. 'Anyone who would rely on the Encyclopedic Tree for all knowledge would know no more than he who while having acquired from Mappemondes a general idea of the globe andits principal areas would flatter himself that he knew all the different peoples that inhabit it and the individual states that exist there.'[5] D'Alembert extended the mappemonde metaphor to remind readers that choices made in ordering knowledge and linking general- to particular-scale articles are similar to choosing a map projection: there are as many possibilities for organizing information as there are ways to look at the world from afar.

D'Alembert and Diderot's mappemonde concept allowed for possible geographies existing beyond authors' selective expertise. Acknowledging such limits, the editors conceded that 'maps' would always be re-drawn, new articles written. For example, thanks to the rise of field surveying and improved instrumentation readers of the final volumes would have known that the longitude coordinates provided in many of the articles were incorrect. Like digital resources connecting to Linked Open Data (LOD), the ENC exemplifies the challenge of confronting ever-expanding knowledge from different perspectives. Our new methods for studying the ENC are equally critical of and inspired by early modern technologies for collecting and recording geographic information.

2.2. Geographic text analysis

Unpacking how geoparsers function is a key concern for GTA (Rupp *et al*. 2013, Murrieta-Flores and Gregory 2015, Clifford *et al*. 2016, McDonough and van de Camp 2017). The promise of GTA to semi-automate the process of linking a place name in a text with a point on a map – 2014 an attractive prospect – masks the poor recall and precision that is problematic when humans are replaced with machines.

This paper sits at the intersection of three related problems in GTA. The first is a basic problem of access to relevant, authoritative digital records about early modern places: even if we accept that a substantial number of European and North American cities have suitable matches in existing gazetteers, there is a lack of gazetteers for a local scale of spatial analysis in much of the early modern world. The second is a problem of cross-disciplinary knowledge: there is little understanding about the history of gazetteers as a genre and the provenance of information in frequently used LOD resources. The third is the design problem: NLP tools remain underdeveloped for non-English and historical texts in formats that diverge from running text.

2.2.1. Access

Empty spaces abound in gazetteer-land. Between the documentation of the ancient world (in the Pleiades) and the resources that scholars of the nineteenth and twentieth centuries make do with, there is very little spatial data about places from about 1400–1800. This lacuna is troubling from a historical perspective because the processes of state-formation, revolution, and colonization throughout the early modern period had a significant impact on political geographies. To complicate matters, early modern scholars, scientists, and travelers from around the world were documenting these geographies at a rapidly increasing rate thanks to improving postal and travel infrastructures and printing presses.

The first generation to widely publish information about ancient geography was born in the sixteenth century. Around the same time, attention turned to contemporary geography, and the first gazetteers in the western tradition were published as lists and descriptions of places for readers, travelers, merchants, and administrators. Recovering this knowledge is crucial to addressing spatial questions about early modern history. Without it, we trick ourselves into answering these questions with anachronistic, and potentially culturally suspect spatial data.

Examining the history of gazetteers shines light on geographic information choices of the past and is one step we are taking to decolonize spatial humanities methods. In Europe alone, given the dramatic transformations in borders, administrative jurisdictions, and changing or standardization of place names during the nineteenth and twentieth centuries, it can be a frustrating practice to establish a convincing match between a named place in a historical text and a Geonames record for rural villages, small cities, and even major ones with turbulent geopolitical histories (think Budapest). Outside of Europe, this data challenge takes an explicitly ethical turn as we attempt to find matches to named places in documents with non-European cultural origins.

2.2.2. History

Spatial humanities research depends on rich metadata that connect place information to a point, set of points, or area. The deep maps like those produced from the Corpus of Lake District Writing (Rayson *et al*. 2017, Reinhold *et al*. 2017) to examine relationships between human experiences and spatial features are impossible without a backbone of gazetteers that link a place name to its attributes (location, temporal relevance, elevation, etc.) (Mostern and Johnson 2008, Mostern 2008, Bodenhamer *et al*. 2010, Southall *et al*. 2011, Manning and Mostern 2015, Mostern *et al*. 2016). Gazetteers save scholars time in locating places and provide authoritative information for disambiguating between places with shared names. Their citable records, expressed as LOD, can be used as authorities and referenced by multiple projects. (The Peripleo platform[6] developed by the Pelagios project is a good example of how gazetteer records can help to make links across collections and projects that otherwise would have no common home.) However, in order to leverage LOC, scholars have flocked to common digital gazetteers like Geonames and ADL. This raises problems of temporal dissonance. By depending on these, we remain silent in the face of the historical obliteration of endonyms (place names in the language of the people who live there) and oversimplify ambiguous or contested naming histories. Aware of this problem, many scholars are creating gazetteers for specific times and places that are particularly unrepresented in such common resources. There has been formal work about why historical gazetteer construction is valuable even when ‘global’ geospatial data is freely available online (indeed these discussions led to the World Historical Gazetteer[7]), but there has so far been little discussion of the implications of continuing to integrate those resources (like Geonames) in NER and other GTA processes.

Applying non-period specific (dissonant, anachronistic) information about a place to historical scenarios can wreak havoc with analysis down the line, make it difficult to compare across projects, and introduce error in assumed administrative hierarchies or other metadata. Gazetteer use should attend to these potential consequences of attaching digital place records to historical attestations of places. For example, if I am annotating the Paris article in the ENC, I would typically match gazetteer records to Paris and Lutèce. I could choose the same or different records from one or many gazetteers that are created by a) institutions (Getty Thesaurus of Geographic Names[8]), b) states (Geographic Names Information System[9]), c) a combination of these plus volunteers (Geonames[10]) or d) scholars (Pleiades[11]). The first three are often used interchangeably as place-name authority records or as sources for latitude/longitude coordinates for historical places. With the exception of Pleiades, these resources are ill-suited to early modern applications because they are built primarily on data from nineteenth and twentieth-century national gazetteer projects that did not usually document the provenance of place names. The choice hinges on the research question. In this case, I am interested in the distinction between Roman Lutèce and Enlightenment Paris, and so even though they share topography, I need to distinguish them temporally. Gazetteers that include metadata about the dates associated with a place name are preferential to those that do not (for example, the date of its attestation in a text or on an object). Temporality and place-name provenance go hand in hand.

Sources like Geonames fail to capture the nuance of toponym history and indeed fail to cover large swaths of the inhabited world. Depending on exonyms (for example, place names attributed to non-French places by French people) erases local endonyms from the historical record and perpetuates colonial exonyms. Working with the ENC corpus is an opportunity to investigate and model the ways that Europeans were engaging with place name history and to document instances where local names are debated or omitted in the source material. Rich with references to the classical works of Pliny, Strabo, Polybius, and others, in addition to medieval and early modern travelers and natural philosophers, the ENC data is an excellent foundation for rethinking how to responsibly produce digital records about global places formerly colonized or even simply explored or mapped by Europeans.

To take a common example, *Barbarie* occurs about 200 times in the Geography subcorpus in reference to the north coast of Africa, such as in the *Retel* entry:

> RETEL ou Arratame, (Géog. mod.) province d'Afrique en Barbarie; son étendue est d'environ 20 lieues, le long de la riviere le Ris; elle confine Ã la province de Sulgumesse, & Ã celle de Métagara. (D. J.)[12]

Geonames has no 'Barbary' or 'Barbary Coast' record, and so there was no adequate way to relate the ENC occurrences to a Geonames match. Wikipedia is also indecisive about the Barbary Coast toponym. In December 2017, an editor suggested that the English article 'Barbary Coast' be merged with the 'Tamazgha' article.[13] Tamazgha is a twentieth-century neologism created to describe the Greater Maghreb. Cases like this highlight the political sensitivity of gazetteer use and construction.

Does it matter if one chooses the same Geonames record to match with both Paris and Lutèce or matches 'Barbarie' with Algeria instead of Tamazgha? Yes. But the argument for 'no' has weight: it depends on the scale of the places you are locating, what you plan to do with your located places, and, unavoidably, how much time you have to invest in documenting and selecting gazetteer records for your places. The argument for 'yes' is, however, an issue of ethical responsibility for humanities scholars becoming participants in the digital marketplace of geographic information. For not only do commonly used gazetteers lack temporal metadata linked to historical place name forms, they simply do not contain content about large portions of the inhabited earth. Just why this is, of course, is related to the history of gazetteers that this project explores. In such cases where digital data simply does not exist or exists in overtly imperialist settings, our 'aim should be not to introduce further bias' when we document early modern places in digital humanities projects (Acheson *et al*. 2017). We have to decide the degree to which our research contributes to defining what a place is.

Geographers concerned with the 'black holes of informational capitalism' (Graham *et al*. 2015) that leave many places invisible in digital geospatial infrastructure argue that 'digital code and content thus do not just reflect the world but also produce it' (Ballatore *et al*. 2017). We would extend these concerns about the ways that gazetteer coverage of twenty-first-century places impacts the production of knowledge about a place in the past. Just as we must produce gazetteers that are culturally sensitive to living communities, we must also create and use gazetteers that are chronologically and culturally sensitive to the places where people lived in the past.

We aim to create a model for future research that accounts for historical endonyms (and potentially recent ones like Tamazgha) in tandem with documentation of the ways that Europeans assigned names to places around the world. This history of naming practices and the ways that reference works like the ENC perpetuate these conventions must be transparent. Like our work, the Cultures of Knowledge Project's EMPlaces[14] and the World Historical Gazetteer provide resources for scholars to establish temporal harmony between places we encounter in early modern research, and the anachronistic gazetteers we have depended on to disambiguate and locate them.

2.2.3. Design & extended named entities

In order to process the ENC text, we faced a core challenge in GTA: NLP tools designed for alternate kinds of corpora (newspapers, social media, lists) and certainly not for early modern French. Initial humanities research with NLP has been very encouraging for identifying and disambiguating spatial information within historical print materials (Alex *et al*. 2015, Simon *et al*. 2015, Santos *et al*. 2015, Cooper *et al*. 2016). (Won *et al*. 2018) evaluates NER tools for the Reassembling the Republic of Letters EU COST Action project and emphasizes two key problem areas: identifying place names and disambiguating results of geocoding.

Disambiguation problems occurring during geocoding often stem from misselection of place entity candidates in Geonames, Wikipedia, etc. In order to extricate ourselves from the circular logic of disambiguating based on gazetteers with limited coverage, we need better identification methods. Irregular spelling and authorial frankness about the confusion between two places that may be the same place highlights the need for a process to capture complex information about place entities. This includes those that are embedded within other types of entities and those that occur in proximity to other entities. This will facilitate disambiguation down the line. Therefore, while geocoding remains an unsolved problem (Wing 2015, Fernando and Bruno 2016, Gritta *et al*. 2018a, 2018b), this paper focuses on the first challenge of identifying historical entities.

Spelling variation remains a significant issue in texts that are not as well edited as the ENC and in corpora made up of many texts as (Butler *et al*. 2017) have shown. Where variation does occur in the ENC, it is because authors intentionally take up thorny onomastic issues. The *Zimara* article, for example, contains multiple spellings of this place combined with contested locations (Zimyra or Zimira at the source of the Euphrates, Simyra in Syria by the sea, etc.).

> ZIMARA, (Géog. anc.) ville de la grande Arménie, selon Solin, qui la place au pié du mont Capotes, où l'Euphrate prend sa source. On lisoit ci-devant dans les exemplaires imprimés de Pline, l. V. c. xxiv. Zimyra, ou Zimira; mais comme l'a remarqué le P. Hardouin, c'étoit une faute insigne: car Simyra est une ville de Syrie au bord de la mer Méditerranée. La correction que ce savant religieux a faite, est appuyée sur les meilleurs manuscrits qui lisent Zimara. C'est ainsi qu'écrit Ptolomée, l. V. c. vij. qui marque Zimara dans la petite Armenie au bord de l'Euphrate, mais assez loin de la source de ce fleuve. Tout cela s'accorde avec les itinéraires. (D. J.)[15]

Variants and alternative names occurring in an entry can reflect separate places, one place that has many alternate names or variations on one name. During our annotation phase, we aim to interpret what each ENC author's intended message was about the most probable attestation. We also plan in the future research to connect these competing attestations to resources of endonyms for non-French places. The inadequacy of equating every article headword with just one Geonames or DBpedia record is a significant factor in building a gazetteer of the early modern world based on textual attestations. One way of dealing with this is to annotate contextual clues, beginning with nested entities and those with spatial offsets and other kinds of extensions.

Another concern is the appearance of religious, supernatural, literary, or mythical geographies. Often they have attested relationships to real places (for example, in the Greek tradition). Rather than ignoring these as non-locatable, in future outputs (a gazetteer, visualizations, or maps) we envision documenting non-locatable places similarly to real places that are difficult to locate because of lost documentation and the absence of physical remains. Places not only have relations to other places, real and imaginary, they also have connections to other entities, in particular, people. *Stilo*, for example, is one of a subset of Geography articles that exist as vehicles for biographies since the editors forbade biographical entries in the ENC.[16] Thomas Campanella, a son of Stilo, is described at length. The *Stilo* entry also features two imaginary, utopian destinations where Campanella's work, *The City of the Sun* (1602) is compared negatively to Thomas More's *Utopia* (1516).

Our efforts to apply extended named entities (ENE) (Gaio and Moncla 2017) to this corpus reflects the significance of the ways that place is embedded within social and spatial relations. This will enable us to augment, if not replace, the quest for geospatial coordinates with other types of information that leverage contextual relations. For example, the City of the Sun has no latitude and longitude, but it has links to Campanella and to Stilo. 'El Dorado' has a less tenuous connection to actual places (Central and South America) than 'purgatory,' but the latter nonetheless has links to other imagined places and people. This approach is generalizable: early modern sources are not alone in containing associations – cultural, social, or geographical – between places and other entities. However, it is particularly important to document ties between people, institutions, events, and land in an era of extensive patrimonial power.

Our contribution to NER design focuses on improving identification of these complex place entities. Leveraging the immediate context of words linked to ENC place names was originally conceived as a disambiguation method for generating a likely location or area for places with no clear match to current gazetteers (McDonough and van de Camp 2017). This grows out of information retrieval (Adams and Janowicz 2012, Adams and McKenzie 2013) and spatial relation extraction research (Mani *et al*. 2010, 2011a, Kordjamshidi *et al.* 2011b, Snoussi *et al.* 2012) as well as research on imaginary places and literary mapping (Murrieta-Flores and Howell 2017). ENE offers an opportunity to document spatial relations and motion verbs, spatial offsets (south, near), and features (*la ville de Paris*) (Moncla *et al.* 2017). Whereas most NER systems will not identify a place name as a place entity if it is found within another kind of entity, we wish to capture this nested information. Perdido therefore also logs place information embedded within a person's title, an institution, an event, or other entity types (see example 1). By linking place names to context, Perdido identifies places that are difficult to disambiguate (like Zimara) and those that cannot be geolocated because they are imaginary, mythical, or extraterrestrial. Perdido's initial design to accommodate ENE is adapted here for the early modern historical context. (EG does not support this functionality because of different design priorities and we only evaluate results from Perdido's ENE identification.) ENE are useful during disambiguation but remain tricky to interpret. For instance, example 2a refers to the ancient city of Troy whereas example 2b is used to describe one Egyptian city named earlier in the text.

In this experiment, we test Perdido's ENE annotation process. The concept of ENE is similar to the concept of nested named entities (Byrne 2007) or tree-structured named entities (Dinarelli and Rosset 2012). An ENE has a proper name (pure or descriptive) at its core and includes one or more concepts (see example 1). A pure proper name is built out of one or more lexemes but includes only proper nouns. A descriptive proper name contains a pure proper name plus a 'descriptive expansion' that can change the type of entity of the embedded pure proper name (Moncla *et al.* 2017). The concept(s) may change the type of the entity, as in example 1 where a place name is embedded in a person's title.

(1) a. *le roi de* *France*
'the king of France'

b. *le chevalier de* *Jaucourt*
'knight of Jaucourt'

(2) a. *ville de* *Troie*
'city of Troy'

b. *ville d'* *Egypte*
'city in Egypt'

Making the distinction between those two cases is very important. In the first case, the feature type embedded in the ENE can be used to disambiguate the embedded entity, whereas in the second case the ENE can be used to resolve ambiguities concerning another entity (such as an article headword). Incorporating ENE will eventually improve the feature classification process (therefore refining the type of gazetteer records that might apply to a particular entity during geoparsing). Fundamentally, it picks up on complex spatially driven entities common in historical texts.

3. Methodology

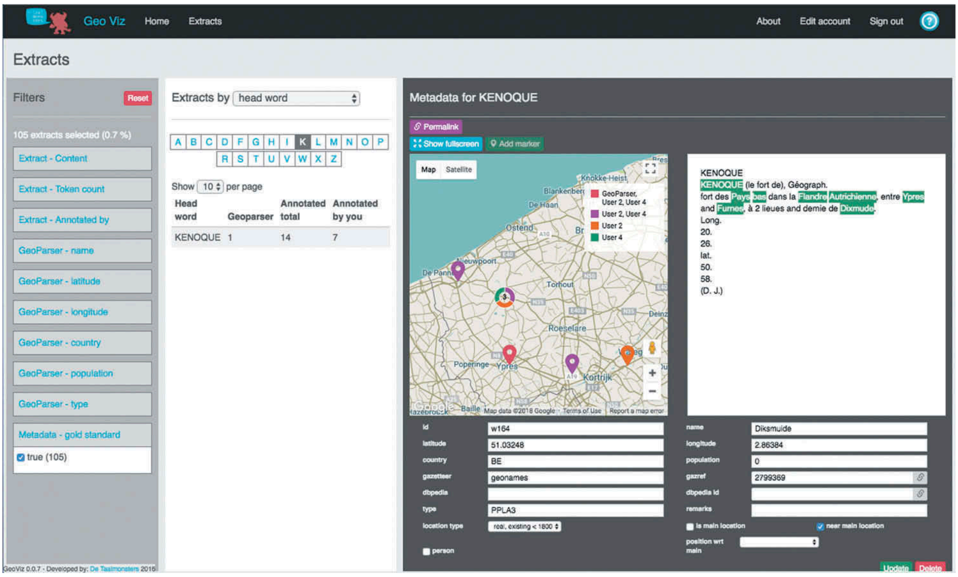
3.1. Data

Our corpus contains 14,445 entries classified as *Geographie* in the ENC. These were provided as one XML file. We removed the XML header and formatting and split each article into a separate text file. We did not alter or modernize the language given earlier research on early modern English texts that found modernization superfluous for NER processing (Won *et al*. 2018). The articles were sub-classified as geography (which includes feature types [Montagnes] and the names of peoples [les Caribes]), modern geography, ancient geography, holy/religious geography, and historical geography (sub-category use varies by volume). Some articles have dual classifications with Mythology, Astronomy, etc. The chevalier de Jaucourt wrote over 8,000 and Diderot about 1,200 of these articles while the remainder were signed by contributors such as d'Alembert, d'Holbach, the abbé Mallet, and Robert de Vaugondy or left unsigned (about 4,800). The articles range from one sentence to 49,000 words. Each article begins with a 'headword' and classification followed by running text about the place. Neither EG nor Perdido was designed to handle the stand-alone headwords. Given the desire to apply NER to dictionary-style texts, improving headword recognition is one of our goals.

3.2. Data model

To evaluate NER output, we developed a gold standard from a set of ENC articles. In the future work, this will be the basis for new domain-specific training data and gazetteer records.[17] We designed the GeoViz interface (Figure 1) for the annotation process.[18] One can consult or edit the full text, validate or add metadata, and map any located places identified in the text (McDonough and van de Camp 2017). The data model accounts for multiple types of places relative to the eighteenth century: existing before 1800 (in existence during the publication of the ENC), historical [no longer extant], biblical, extra-terrestrial, mythical, and, unknown (Table 1). We also annotated spatial context in relation to the main location of each article (is location, near, contained by, south of, etc.) to assist with disambiguation in the future.

For this exercise, EG output was GeoViz input. EG output included the ENC article title (Headword), identified place names, and one Geonames match. Because EG did not identify many named places in the text, GeoViz also allows editors to select new place names. (The average number of EG annotations per extract was 1.27 compared to 10.44 for KM.) The annotation process was broken down into the following steps: 1) validate the Geonames match as suitable; 2) as needed, correct Geonames match (gazref); 3) use GeoViz to automatically query and add (where possible) DBpedia record id; 4) add metadata about location type, relation to headword, place name embedded within person name, other notes. We privileged Geonames feature type PPL for inhabited communities. ADM was used for regions (with additional research to distinguish between ADM1-2-3 and the infrequent ADM1H [historical]). Other types included AREA (for cultural/historical regions vs. state or institutional regions), CONT (continent), and natural feature codes for bodies of water, mountains, etc.[19]



105 extracts selected (0.7 %)

☑ true (105)

Extracts by head word

Show 10 per page

Head word	Geoparser	Annotated total	Annotated by you
KENOQUE	1	14	7

Metadata for KENOQUE

KENOQUE
KENOQUE (le fort de), Géograph.
fort des Pays bas dans la Flandre Autrichienne, entre Ypres
and Furnes, à 2 lieues and demie de Dixmude.
Long.
20.
26.
lat.
50.
58.
(D. J.)

id	w164	name	Diksmuide
latitude	51.03248	longitude	2.86384
country	BE	population	0
gazetteer	geonames	gazref	2799369
dbpedia		dbpedia id	
type	PPLA3	remarks	
location type	real, existing < 1800	□ is main location	☑ near main location
□ person		position wrt main	

Figure 1. GeoViz annotation interface.

Table 1. Location types for a gold standard set.

Location type	Place identified as type
Real (existing before 1800)	1710
Real, historical (no longer existing in 1800)	420
Biblical	12
Unknown	4
Mythical	4
Real, extraterrestrial	1

In determining whether or not the Geonames (and DBpedia) match was appropriate, annotators used expert domain knowledge. Real places existing at the time in which the author was writing (ca. 1760) posed fewer problems than other location types. However, even these places (especially smaller communities) could have different names and jurisdictions today. Historical (pre-1400), biblical, and mythical location types were a more formidable challenge. First, we examined the underlying data sources for Geonames and DBpedia. We then considered the extent to which place names (with the same or distinct character strings) could represent two different places or the same place name in different languages or naming traditions. Finally, we considered the geographical similarity of the eighteenth-century attestation with the digital gazetteer record: places for which evidence could be found that established shared coordinates could also be considered to be the same place. It is a gross understatement to say that these choices were difficult.

For the gold standard annotations, 77% of the named places could be manually matched with a DBpedia record while only 22% were matched with a Geonames record. Regardless of how good or bad these results might seem quantitatively, these statistics gloss over the qualitative intentions of the texts. For example, in the *Lissus* entry, there are clear Geonames and DBpedia records that potentially match the headword. But the entire point of this article is to suggest multiple locations for ancient places referred to by the word *Lissus*: one refers to a city in Dalmatia cited by Pliny, the second to a place in Crete, and the third to a river in Thrace. In seeking to break the pattern of replicating data structures that simplify historical geography rather than contextualize it (and to honor the intentions of the ENC contributor), we would ideally not select the *same* record for each named Lissus (1,2,3) or at least we would establish different relationships between these attestations and a digital record. In the next phase of GeoViz development, we will improve the ability to delineate these relations. For the time being, we linked the headword instance of *Lissus* with a gazetteer record but left the next two entities unlinked.

In GTA research, georesolution (matching each named place to one set of coordinates defined in a pre-existing resource) has been the preferred result. Instead, by leaving space for historical attestations that have no relevant digital record matches, future GTA work should offer contextual location methods. As the first step in this process, this paper presents initial results for applying ENE to the ENC text. Annotations using this model will be available soon as training data for several tasks of geographic information retrieval including NER and toponym resolution. For instance, this can be used for training a new model of annotation for machine learning approaches specifically for French historical texts. These data may also be used for evaluation of rule- or learning-based methods.

Our gold standard set of 100 articles was randomly selected. Inter-Annotator Agreement (IAA) for identifying words as places (Table 2) and for selecting the same gazetteer records (Table 3) was sufficiently high to be applicable for future machine learning uses. Word ID agreement measures whether all words are included in the entity or not. The high Kappa score for human-human annotators (0.98 between KM and LP for Word ID agreement and 0.97 for Gazetteer reference agreement) reflects the similarity in human choices. Given these IAA measures, we determined that annotator 3 (KM) would be sufficient for the gold standard (McDonough and van de Camp 2017).

Table 2. Word ID agreement between annotators (A1 & A2) (McDonough and van de Camp 2017).

A1	A2	Extracts	Words	Annotated words A1	Annotated words A2	Kappa
KM	LP	91	28,516	1107	1107	0.9849
KM	SB	10	1160	92	80	0.9836
LP	SB	10	1160	84	80	0.9775

Table 3. Gazetteer reference agreement (McDonough and van de Camp 2017).

A1	A2	Extracts	Words	Annotated words A1	Annotated words A2	Kappa
KM	LP	91	28,516	1107	1107	0.9714
KM	SB	10	1160	92	80	0.9603
LP	SB	10	1160	84	80	0.9620

3.3. Geoparsing

EG and Perdido are rule-based geoparsers (Figure 2). Like (Won *et al*. 2018) we received a version of the default EG distribution with a French Part-of-Speech (POS) Tagger swapped in. Aside from that, no other adaptations to process French text were made (e.g. lexicon lookup or rule-related changes). EG, therefore, accounts for French parts of speech, but it implements NER based on rules developed for English. This is a recognized problem for scholars working in non-English languages and even in historical (usually pre-twentieth century) English. Perdido is a fully French pipeline.

EG is built as a pipeline of components (tokenizer, sentence-splitter, POS tagger, lemmatizer, chunker, NER) (Alex *et al*. 2015). Both EG and Perdido make use of similar POS preprocessing and rules (lexical lookups, linguistic context). The EG nertag component is a rule-based named entity recognizer similar to the MUC7 named entity evaluation: numex (money and percentages), timex (dates and times) and enamex (persons, organizations and locations) (Chinchor 1998).[20] The ADL name list is the place-name lexicon.[21] EG uses two rules to contend with single-word and multi-word place names. Multi-word places cannot be re-identified as other types of entities while single-word places can. Single- and multi-word places can be contained within other types of entities (as part of the enamex pipeline), but once they are embedded within another entity the internal markup of the place is not preserved as place-related entity data.

Perdido was designed to annotate ENE and motion events (Moncla *et al*. 2014, Gaio and Moncla 2017) and is based on the combination of a preprocessing step, where POS tagging and lemmatization are performed, and a main process, which implements a finite state transducers cascade. This cascade uses transducers (i.e. graphs) to apply rules for annotating named entities and spatial context. During this task, entities are classified as place, person, date, or other based on the local context available in the text (using a lexical lookup, see Table 4). If there is no specific context (e.g. *Paris* vs *ville de Paris*), entities are classified as unknown and a second classification step is performed during the georesolution task based on a gazetteer lookup method.

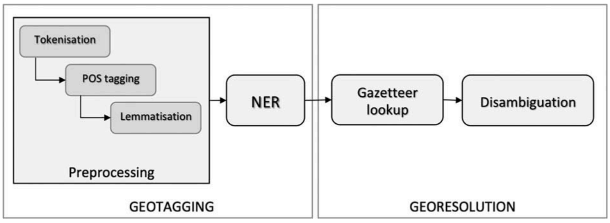


Figure 2. Overview of EG and perdido pipeline.

Table 4. Lexicons used in Perdido NER.

Type	Examples
First names	*Jules, Thomas, Louis, Achille, Hercules, …*
Titles, professions	*roi, duc, abbé, chevalier, écrivain …*
Demonyms	*romains, grecs, gaulois, …*
Geographical terms	*ville, fleuve, vallée, province, …*

Both depend on place names in the ADL lexicon, which lacks most place names in French. The ADL began with data from the US government (the Geographic Names Information System for domestic places[22] and the Geographic Names Processing System for foreign places[23]) because they 'offer the best available worldwide coverage of geographic place names' (Hill *et al*. 1999). However, it should be standard practice to question the coverage of the ADL (or any off-the-shelf lexicon/gazetteer) for historical research uses. ADL's lack of foreign-language name variants severely limits what is identified as a place during NER.

For georesolution, geoparsers usually use gazetteer lookup methods to associate geo-coordinates to place names. During this process, many ambiguities may arise (Smith and Mann 2003). Most significantly for this research, one name may refer to several locations (referent ambiguity) and that one place can have several names (reference ambiguity). Thus, toponym disambiguation (Leidner 2007) is a very important task in identifying the location corresponding to a place name found in the text.

In both EG and Perdido, we used only Geonames as a gazetteer during georesolution. Given our focus on NER evaluation rather than georesolution, for EG, we did not limit the query by feature type and we saved only one match (acknowledging that the Geonames result is unranked). Not enough place names were identified for the ranking algorithm's preference for places clustered near each other to have a visible effect.[24] Perdido uses the gazetteer lookup method with two objectives. The first is to determine if the entities classified as 'unknown' during the NER task refer to a place. The second objective is to find the coordinates of entities classified as places. We configured Perdido to obtain multiple results (5) in order to experiment with ranking methods at a later date. When dealing with ENE, Perdido queries Geonames for both extended and embedded entities. Information captured by ENE (i.e. feature types) can eventually be used for disambiguation (Moncla *et al*. 2014).

4. Results

4.1. NER analysis

The results measure the recall, precision, and F1-score for EG and Perdido against the gold standard (Table 5). Recall measures NER entities (true positives) against human-annotated entities (both true positives and false negatives). Precision is the number of correctly identified entities (true positives) over the total number of identified entities (true and false positives). The F1-score is the weighted average of these two measures.

Recall from EG was low given the lack of language- and text structure-specific adaptations (9.2%). But Perdido was not as high as expected (only 55.58%) based on other tests with nineteenth-century French novels (which had recall of 99.7% for Parisian street names) (Moncla *et al*. 2017). Precision of EG results was high, 94.64% compared to only 75.71% for Perdido.

Table 5. NER evaluation.

	Recall	Precision	F1-score
EG	9.20%	94.64	16.78%
PERDIDO	55.58%	75.71%	64.10%

EG's low recall should not be interpreted as poor performance since this is a tool designed for other kinds of corpora. KM found an average of 10.44 places per extract compared to the 1.27 from EG (and 9.79 from Perdido – including false positives). As a point of comparison, using EG on a corpus of early modern English letters resulted in 53.8% precision, 52.4% recall, and a 53.1% F1-score (Won *et al*. 2018). Like those letters, place names occur in the ENC in multiple languages (Greek, Spanish, English, German, and more), which disrupts the NER process.

With a 64.10% F1-score, Perdido underperformed compared to its other implementations (hiking descriptions, nineteenth-century novels). Why might this be? In addition to the multiple language issue (common in early modern corpora), there are issues with detecting proper names that stand apart from other text. Headwords, for example, were only identified twice in the entire exercise as place entities. Using a different POS tagger may resolve this problem. Previous Perdido experiments involved quite different tasks and/or types of text such as analyzing hiking descriptions containing mainly spatial entities (Gaio and Moncla 2017) or urban road names in novels (Moncla *et al*. 2017). Extracting entities from the articles of the ENC is more challenging. Not all entities only refer to places, some are expressed in foreign languages, and the use of a dictionary of proper names to classify place or person names is less efficient given the historical context.

For example, false positives in identification occur more often in Perdido than in EG. Perdido has a habit of picking up person names as place names: Copernicus, Pliny, Cato, Strabo, and Bayle all registered as places. These names were not embedded in ENE with contextual information about the type of entity. Furthermore, the names have referents existing in Geonames thus making them erroneously locatable (Bayle is a city in France, and Pliny is a town in West Virginia). ENE information related to a person (e.g. title, profession) helps with manual classification, but person names that are only one word are more difficult to handle. For example, *roi de Chypre, chevalier de Jaucourt, empereur Louis* are appropriately classified as people's names, whereas the figures above or any person without a title are not. The NER component of the Perdido processing chain needs to be improved to reduce the number of false positives. However, the solution of using dictionaries of person names is not sufficient (Perdido already uses a dictionary of common first names): in the case of the ENC, many early modern names are no longer common. Thus, one solution would be to query LOD resources, each adapted for a type of entity (historical names), and not just geographical gazetteers. (For example, access to authority records of persons from the *Bibliothèque nationale de France* would improve these results.) In addition to individuals, demonyms (e.g. *Romains, Gaulois*) also were captured as places (despite an explicit demonym lexicon).

Name variants and names that simply do not register with EG were identified as places by Perdido, despite using the same underlying place lexicon (ADL). For example, the *Zimara* article names 14 places. Perdido recognized 11, two of which were false positives (persons). Perdido did not identify variants on Zimara (Zimira, etc.). EG recognized no places. Similarly, the *Siam* article contained 64 named places. EG recognized no place names while Perdido identified 44 including 12 false positives (Siamois and other demonyms). *Siam* was identified as a place by Perdido. In the gold standard annotations, *Siam* was matched to the *Kingdom of Thailand* Geonames record (1,605,651). *Siam* is an exonym applied to the Thai state by outsiders including early modern Europeans. False negatives can derive from gaps in the ADL data, but also from the rule-based approach.

If *Zimara* raises questions about the problems of picking one attestation to define a place, *Siam* offers another case study in determining how to model endonyms and exonyms that occur in the historical record.

Finally, another category of alternate names – historical names – posed a challenge for lexicons and rules. EG usually picked up *Grece, Troyes, Cyclades*, and *Palestine*, but not the historical places of *Gallia Belgica, la Perse, Béni-Arax*, or *Salapia*. Perdido recognized *la Perse* and *Salapia*, but not the others. As we move forward to modeling records for a gazetteer drawing on the information in similar early modern works, being able to attach temporal metadata to occurrences of endonyms, exonyms, and historical names may smooth the path towards understanding their origins and application over time. The Trismegistos gazetteer record for *Salapia (Salpi)*, for example, includes chronological data for attestations.[25]

Using multiple gazetteers is advised in georesolution (Alex *et al*. 2015, Gregory *et al*. 2015), and we should do the same in NER to improve these gaps and errors in identification. Won *et al*. (2018) use the new EM Places gazetteer from Early Modern Letters Online (EMLO) to complement the lexicons used in the EG, NER-Tagger, Stanford NER, Polyglot, and spaCy NER systems. Using multiple attestation-based gazetteers (like Pleiades and Trismegistos) and other resources as lexicons for persons and places increases the chances of finding entities that are temporally and culturally appropriate for a corpus.

4.2. Preliminary results: ENE

It is frustrating to lose many place names during the NER process when non-place entities are separated from place entities. Omitting complex entities actually distorts the places that are captured (for example, identifying only Gallia instead of Gallia Belgica). Furthermore, place names embedded within person names are typically lost, even when these places could be significant spatial referents in the text.

Perdido allows for a multi-step classification process that accounts for complex, embedded, or simple ENE. Among the 1721 entities annotated by Perdido, 396 (i.e. 23%) are embedded in ENE. Additionally, 51 ENE refer to a person's name and 33 of these extend a place name (i.e. 65%). Saint Francis of Assisi in the *Montefalco* article adds a new spatial reference, and, if Montefalco's location were in doubt, *Assisi* could be a useful point for narrowing the frame of reference relevant to this article.

> MONTE-FALCO, (Géogr.) petite ville d'Italie dans l'état de l'Eglise, au duché de Spolete, sur une montagne, près du Clitunno. Long. 30. 25. lat. 42. 58. Elle se vante d'avoir donné la naissance à sainte Claire en 1193. Cette pieuse amie de saint François d'Assise établit un couvent dont elle fut abbêsse, sonda l'ordre des religieuses qui portent son nom, mourut en 1253, & fut canonisée peu de tems après par le pape Alexandre IV.[26]

Institutions also derive meaning from their location, and the location entity would be misrepresented without this attached feature (*la commanderie de Malte, le parlement de Metz*). Ideally, we would like to make it possible to include specific types of results in separate queries so as not to mix ENE and non-ENE entities when this would be undesirable (e.g. one would not always want to locate the *duc de Berri* or *la reine de Navarre*).

ENE appeared for more than just personal titles and institutions. The ENE *La carte de France* (the Cassini maps) reminds us that objects related to places pop up frequently in these articles (including currencies like *trente sols de Hollande* or *l'argent de Siam*). Collective peoples and areas were expressed frequently as complex entities, for example: *la royaume de Fez* and *la vaste région de l'Amérique*. Wars and treaties reference places and occur on a regular basis *la guerre de Candie, le traité de Munster*. Creatures and mythical things crop up (*satyres de Perse* and *la chimére de Lycie*) as do events (*la fameuse ambassade de Siam* or *la celebre journée d'Actium*). ENE are extremely important for attestations of bodies of water and other natural features (*embouchure du Rhône*), and indeed the offset clarifies what is being referenced.

5. Discussion

Our results serve as a reminder that geoparsing remains a highly problematic process for non-English, non-modern corpora. Technical questions about entity identification feed into broader concerns about the political assumptions underlying the design of geoparser components. Anachronistic use of lexicons or gazetteers alongside historical place evidence raises ethical questions about the reproduction of imperial scientific and settler place-naming practices and the erasure of other naming cultures around the world. Ignoring ENE or spatial offsets has consequences for early modern corpora (and undoubtedly also for other corpora). Privileging georesolution as the goal of geographic text analysis ignores rich contextual information that defines a place even if it does not locate it geospatially.

ENC entries are echoes of the oral and text traditions of sharing geographic knowledge. This makes it a fitting corpus for posing questions about naming histories and the structural origins of geographic information. The ENC holds many texts within it. It can be seen as a deep, if intentionally incomplete, gazetteer taking stock of European-gathered geographic information around the 1750s. The depth of knowledge about place in these volumes defies a simple 1:1 or even 1:3 place-name to location identification: how could a place that Jaucourt says is in ruins possibly have a Geonames match? Such a conundrum inspires a new understanding of the links between early resources like the ENC and the digital gazetteers we use today. Questioning digital data provenance and coverage reminds us to do the same with the ENC. This raises the issue of what counts as an attestation or authoritative statement about a place's existence, the ideal seed of a gazetteer record in the past and today.

Studying ENC authors' critical approaches to place names in the eighteenth century can help us to think through disambiguation methods today. Jaucourt's article on Zimara models the contained chaos at the heart of many entries.[27] First, he justifies the selection of the article headword (though often titles are immediately followed by the phrase 'or [some other spelling variation]'). In the case of Zimara, he then not only documents contested location reports (near the source of the Euphrates [Solon] or not [Ptolemy]), he also remarks on a typo in printed editions of Pliny's *Natural History* that confuses Zimara (a city in Armenia minor) with Simyra (a city in Syria). The article avoids committing to where the city is in relation to the Euphrates except to say that 'all of the itineraries agree' with Ptolemy's rendition. We are left with a hint that Ptolemy is correct, but the article confusingly begins with the Solon description. Zimara, and many other places, escape easy definition. The structure of this article suggests that it is the information at the end of the article that has been vetted by the author: it is this information rather than the leading description that is most reliable. And yet, there is no explicit rejection of Solon's location. Zimara is, in the ENC, as much a construction of the classical authors as it is a physically located place. The contentious history of places is recorded in physical footprints as well as media. The ENC is instructive in documenting variant spellings, recording disputes, and leaving the 'where' open for discussion.

These days, the Pleiades[28] and DARE[29] gazetteer records for Zimara preserve attestations from (the corrected) Pliny[30] as well as the Peutinger Table and the Itinerarium Antonini. Jaucourt's research has held up well, but in these new records, we do lose sight of the times when scholars got things wrong. In the quest to locate, digital gazetteer records may lose the 'incorrect', but historically significant evidence. It can be useful to preserve documentation of these mistakes, of past confusion. Being able to locate Zimara today is certainly useful, but it is also important to understand that people living in the medieval or early modern periods would not have had access to this certainty.

One way to address locational ambiguity is to further develop Perdido's ENE. This would refine methods for identifying different types of ENE and documenting the hierarchies or other relations in which they appear. Additionally, using different algorithms to find near matches for variants like *Zimira* and to link these to spatial offsets or other nearby entities will improve disambiguation efforts. ENE are an asset for refining geographic information retrieval, both individually (since they retain nested meanings) and collectively (since a set of neighboring ENE can be examined to help provide approximate locations for challenging historical places). ENE has the potential to capture the scope of geographic information as it was shaped in a period when accurate locations were: 1) difficult to measure and 2) not necessarily useful.

For example, Népi is a city that no longer has any inhabitants by the eighteenth century. It is found within the Papal States. We should, however, be able to determine an approximate location from contextual details like the listed distances from Rome and Magliano, the latitude, and the proximity of the tributary Triglia.

> NEPI, (Géog.) ancienne petite ville dépeuplée d'Italie, au patrimoine de S. Pierre, sur la riviere de Triglia, qui se jette dans le Tibre, avec un évêché suffragant du Pape, Ã 8 lieues N. de Rome, 4 S. O. de Magliano. Long. 30. 2. lat. 42. 12.[31]

Because articles are often vehicles for biographical material about the ancients and the moderns, we can also document this relationship between people and place as a substitute for location (à la Assisi). The article *Mégalopolis* is an éloge of Philopaemen and his student, the Greek historian Polybius, who both hailed from this city. Mégalopolis, attested in eight separate sources (with different spellings or different names altogether [Leontari]), is likely included in the ENC for this sole purpose.[32] It would seem logical, therefore, to treat the relationship between Megalopolis and Philopaemen and Polybius as a social location. It was surely a piece of knowledge that held more certainty and significance than location in the eighteenth century. ENC articles documenting the homeland (*la patrie*) of influential people, therefore, deserve further attention.

Whereas the feature type embedded in the ENE can often specify the nature of the embedded entities (e.g. *riviere de Triglia*), many ENE are used to describe another entity. For instance, in the previous article, the ENE *ville de Péloponnese* describes Mégalopolis.

Thus, in this case, the feature type embedded in the ENE (i.e. *ville*) refers to the nature of the entity *Mégalopolis* and not to the entity *Pèloponnese*. This leads us to consider several types of ENE and to use spatial entities expressed in an extract as knowledge used to describe a place or its spatial context. On the one hand, this adds a difficulty in the disambiguation process (i.e. disambiguation of embedded entities and proximate info), but, on the other, this allows us to find information using relations between entities (e.g. hierarchical, topological, social).

Annotating the gold standard (and eventually making a gazetteer) in a way that addresses the ethical issues above has translated, in this phase of research, to sticking as closely as possible to the authorial and editorial intentions of the ENC. Any gazetteer based on attestations from European sources will inevitably be host to a range of names and place histories that could be contested (like *Barbarie*). The only objective possible given the source material is to work forward from the evidence, use opportunities like LOD to enrich place name histories that have complex lives beyond the source text, and resist the temptation to geospatially locate **every** named place. This text-based, bottom-up approach means that we must depend less on the top-down strategy of finding places in existing lexicons. Becoming comfortable with maps that will be explicitly incomplete, as they would have been for Diderot himself, is a healthy exercise in spatial history (as Harley would surely agree). What will be striking are the ways we generate to analyze what cannot be mapped. Creatively using ENE annotations is one opportunity to re-imagine spatial history that produces not just maps, but other kinds of visualizations that can cope with non-geospatial or temporally diverse information.

6. Conclusion

NER analysis is the first step in unpacking GTA processes and assessing their function for non-English, non-modern, formally complex texts. For EG, the consequences of working with non-English language materials meant that recall was poor. For Perdido, the challenge of working with place names in multiple languages (not only French), differently structured texts, and different types of ENE produced lower than expected results. The gold standard exercise with EG helped us to pinpoint specific areas to adapt in Perdido.

Our overarching project aims to create a better NER system for early modern French language that will lead to attestation-based lexicons and gazetteers feeding back into the geoparsing process. In combination with other scholarly and locally created gazetteers, it will become possible to resolve certain gaps in historical gazetteer coverage around the world. However, coverage of past geographies will never be complete. This is one reason why we should reassess geolocation as a key GTA goal: not every place that humanities scholars want to locate has enough extant documentation to permit this, nor is this necessarily the most useful task for all research questions. By thinking about the structure of, and qualitative relations between, spatial information rather than the location of only geospatial places, we might move beyond plotting points on a map as the primary interpretative process in spatial history. This will open up the spatial humanities to new opportunities in analyzing the language that people used to describe their experiences among, and ideas about, real and imaginary, ruined and extant, earthly and lunar places.

Perdido's ENE are one method of capturing qualitative spatial information and links to other types of entities. ENE assist in three ways: they 1) document place names embedded within non-place entities, 2) connect place names to spatial offsets that change the nature of the place in question, and 3) provide contextual details that can establish relationships between and improve disambiguation of adjacent place names. ENE are a good fit for corpora where the distinctions between entity types are not precisely the same as they are today (e.g. early modern people referred to commonly by titles, which are expressions of power related to territorial claims). They can account for the inter-relations between places, people, events, objects, and institutions native to early modern cultural contexts.

6.1. Future research

Geoparsing early modern texts brings home how much spatial humanities has left to discuss about the implications of using lexicons and gazetteers like ADL and Geonames. How should we digitally express the complex relationships between pieces of evidence? First, we can move beyond the georesolution goal to instead focus on leveraging context. This means that we need to capture other kinds of spatial information to augment simple place name extraction. As in the *Zimara* entry, the context may not necessarily represent 'correct' information, but it does offer a glimpse of the status of knowledge about a place during a specific historical moment. These details reflect historical spatial knowledge and our methods should be flexible enough to include them. Our experiments demonstrate the rigidness of uniquely rule-based NER. The main problems are due to the specificities of the language used in early modern French texts. Indeed, both Perdido and EG make use of a part-of-speech analysis (during preprocessing) based on models not adapted to this language. This is another way that GTA has implemented NLP tools that are anachronistic. Thus, a first improvement would be to build new models, either for part-of-speech tagging or for NER using the ENC and other early modern dictionary-style texts like the Saugrain and La Martinière mentioned at the beginning of this article. Using our growing annotated corpus as training data, we wish to combine machine learning approaches with rule-based NER, using both to improve results for entity recognition in early modern texts. Using the concept of ENE, we can imagine a hybrid method implementing a machine learning approach for NER improved by a rule-based method for retrieving more information related to the named entities already annotated. A rule-based solution may also help to build up the training data and iteratively improve results of the machine learning approach, lowering the high barrier to entry to machine learning for projects like ours that have to annotate their own data.

Notes

1. https://encyclopedie.uchicago.edu/.
2. The ENC corpus was provided by the ARTFL Project.
3. The GÉODISCO project (Approche GéOmatique et linguistique du DISCours encyclOpédique des Lumières à Wikipédia) is working on these larger questions (Denis Vigier, Thierry Joliveau, Ludovic Moncla, and myself) https://www.msh-lse.fr/projet19/geodisco.
4. https://encyclopedie.uchicago.edu/node/174.
5. https://encyclopedie.uchicago.edu/node/88.
6. http://peripleo.pelagios.org/.

7. http://whgazetteer.org/.
8. http://www.getty.edu/research/tools/vocabularies/tgn/index.html.
9. https://geonames.usgs.gov/.
10. http://www.geonames.org/.
11. https://pleiades.stoa.org/.
12. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/14/1149/.
13. https://en.wikipedia.org/wiki/Tamazgha.
14. http://www.culturesofknowledge.org/?p=8455.
15. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/17/2929/.
16. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/15/2555/.
17. Gold standard data for this article is available on our Github repository: https://github.com/kmcdono2/hgeo. Improvements to this dataset are underway.
18. GeoViz code is available at https://github.com/Taalmonsters/GeoViz.
19. https://github.com/kmcdono2/hgeo/wiki/GeoViz-Annotation-Methodology.
20. http://groups.inf.ed.ac.uk/geoparser/documentation/v1.1/html/geotag.html#pipelines.
21. http://legacy.alexandria.ucsb.edu/gazetteer/.
22. https://geonames.usgs.gov/domestic/.
23. http://geonames.nga.mil/gns/html/index.html.
24. http://groups.inf.ed.ac.uk/geoparser/documentation/v1.1/html/georesolve.html#ranking.
25. https://www.trismegistos.org/place/14287.
26. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/10/2932/.
27. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/17/2929/.
28. https://pleiades.stoa.org/places/629106.
29. http://dare.ht.lu.se/places/22405.html.
30. http://latin.packhum.org/loc/978/1/351/432-439#351.
31. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/11/535/.
32. https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/10/1356/.

Acknowledgments

We thank the ARTFL Project; Claire Grover and Beatrice Alex; and Laure Philip and Simon Burrows. This paper benefited greatly from early discussions with Jason Ensor.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Katherine McDonough is a Senior Research Associate on the Living with Machines digital history project at The Alan Turing Institute and a Research Fellow in the Department of History, Queen Mary University of London. She works on the history of infrastructure and information.

Ludovic Moncla is an Associate Professor in Computer Science at INSA Lyon (Department of Industrial Engineering) and LIRIS laboratory (UMR 5205 CNRS). His research interests are oriented towards pluridisciplinary aspects of Natural Language Processing (NLP), information retrieval, data mining, digital humanities and geographical information science (GIS).

Matje van de Camp works as an independent researcher and scientific programmer with her company De Taalmonsters. Her research interests lie in the field of computational linguistics, specifically information extraction, sentiment analysis, and social network extraction. She develops interfaces and software to assist others with the annotation, curation, and dissemination of research materials.

ORCID

Katherine McDonough http://orcid.org/0000-0001-7506-1025
Ludovic Moncla http://orcid.org/0000-0002-1590-9546

References

Acheson, E., Sabbata, S.D., and Purves, R.S., 2017. A quantitative analysis of global gazetteers: patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64, 309–320. 8. doi:10.1016/j.compenvurbsys.2017.03.007

Adams, B. and Janowicz, K., 2012. On the geo-indicativeness of non-georeferenced text. *In: ICWSM*, 375–378. 9. doi:10.1094/PDIS-11-11-0999-PDN

Adams, B. and McKenzie, G., 2013. Inferring thematic places from spatially referenced natural language descriptions. *In*: Sui, Daniel Z., Elwood, Sarah, Goodchild, Michael F. (Eds.), *Crowdsourcing geographic knowledge*. Netherlands: Springer, Vol. 10, 201–221.

Alex, B., *et al*., 2015. Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, 9 (1), 15–35. 2, 8, 14, 17. doi:10.3366/ijhac.2015.0136

Ballatore, A., Graham, M., and Sen, S., 2017. Digital hegemonies: the localness of search engine results. *Annals of the American Association of Geographers*, 107 (5), 1194–1215. 8. doi:10.1080/24694452.2017.1308240

Bodenhamer, D.J., Corrigan, J., and Harris, T.M., 2010. *The spatial humanities: gis and the future of humanities scholarship*. Bloomington, IN: Indiana University Press. 6.

Butler, J.O., *et al*., 2017. Alts, abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, 13 (1), 58–81. 8. doi:10.1080/15420353.2017.1307304

Byrne, K., 2007. Nested named entity recognition in historical archive text. *In: International Conference on Semantic Computing (ICSC 2007*), September, 589–596. 10. doi:10.1094/PDIS-91-4-0467B

Chinchor, N.A., 1998. *Overview of muc-7/met-2*. San Diego, CA: Science Applications International Corp. 14.

Clifford, J., *et al*., 2016. Geoparsing history: locating commodities in ten million pages of nineteenth-century sources. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49 (3), 115–131. 5. doi:10.1080/01615440.2015.1116419

Comsa, M.T., *et al*., 2016. The French enlightenment network. *The Journal of Modern History*, 88 (3), 495–534. 4. doi:10.1086/687927

Cooper, D., *et al*., 2016. *Literary mapping in the digital age*. London: Routledge. 8.

Dinarelli, M. and Rosset, S., 2012. Tree-structured named entity recognition on OCR data: analysis, processing and results. *In*: *Language Resources Evaluation Conference (LREC*), May. Istanbul, Turkey, 10. doi:10.1094/PDIS-11-11-0999-PDN

Edelstein, D., 2016. Intellectual history and digital humanities. *Modern Intellectual History*, 13 (1), 237246. 4. doi:10.1017/S1479244314000833

Fernando, M. and Bruno, M., 2016. Automated geocoding of textual documents: a survey of current approaches. *Transactions in GIS*, 21 (1), 3–38. 8.

Gaio, M. and Moncla, L., 2017. Extended named entity recognition using finite-state transducers: an application to place names. *In*: *9th International Conference on Advanced Geographic Information Systems, Applications, and Services*, Nice, France. 2, 9, 14, 16.

Graham, M., De Sabbata, S., and Zook, M.A., 2015. Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, 2 (1), 88–105. 8.

Gregory, I., *et al*., 2015. Geoparsing, GIS, and textual analysis: current developments in spatial humanities research. *International Journal of Humanities and Arts Computing*, 9 (1), 1–14. 17. doi:10.3366/ijhac.2015.0135

Gritta, M., Pilehvar, M.T., and Collier, N., 2018a. Which Melbourne? Augmenting geocoding with maps. *In*: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 1285–1296. 8

Gritta, M., *et al.*, 2018b. Whats missing in geographical parsing? *Language Resources and Evaluation*, 52 (2), 603–623. 8. doi:10.1007/s10579-017-9385-8

Grover, C. and Tobin, R., 2014. A gazetteer and georeferencing for historical English documents. *In*: *Proceedings of LaTeCH 2014 at EACL 2014*. *Gothenburg, Sweden*: Association for Computational Linguistics, 119–127. 2.

Grover, C., *et al.*, 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368 (1925), 3875–3889. 2. doi:10.1098/rsta.2010.0149

Hill, L., Frew, J., and Zheng, Q., 1999. Geographic names: the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*, 5 (1), 15. doi:10.1045/dlib.magazine

Kordjamshidi, P., *et al.*, 2011a. Relational learning for spatial relation extraction from natural language. *In*: *International Conference on Inductive Logic Programming*. Heidelberg: Springer, 204–220. 10.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.F., 2011b. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8 (3), 4. 10.

Leidner, J.L., 2007. Toponym resolution in text: annotation, evaluation and applications of spatial grounding. *SIGIR Forum*, 41 (2), 124–126. 15. doi:10.1145/1328964

Mani, I., *et al.*, 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44 (3), 263–280. 10. doi:10.1007/s10579-010-9121-0

Manning, P. and Mostern, R., 2015. World-Historical Gazetteer. 6.

Mayhew, R.J., 2010. Geography as the eye of enlightenment historiography. *Modern Intellectual History*, 7 (3), 611–627. 4. doi:10.1017/S1479244310000259

McDonough, K. and van de Camp, M., 2017. Mapping the encyclopédie: working towards an early modern digital gazetteer. *In*: *1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, Redondo Beach, CA, USA: ACM, 16–22. 5, 9, 12, 13, 14.

Moncla, L., *et al.*, 2014. Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus. *In*: *22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14. Dallas, TX, USA: ACM, 183–192. 14, 15.

Moncla, L., *et al.*, 2017. Automated geoparsing of Paris street names in 19th-century novels. *In*: *1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, November. Redondo Beach, CA, USA, 10, 16.

Morrissey, R., Roe, G., and Spring, 2016. *Edition. Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc., eds. Denis Diderot and Jean le Rond d'Alembert*. University of Chicago. ARTFL Encyclopédie Project. 3.

Mostern, R., 2008. Historical gazetteers: an experiential perspective, with examples from Chinese history. *Historical methods. A Journal of Quantitative and Interdisciplinary History*, 41 (1), 39–46. 6. doi:10.3200/HMTS.41.1.39-64

Mostern, R. and Johnson, I., 2008. From named place to naming event: creating gazetteers for history. *International Journal of Geographical Information Science*, 22 (10), 1091–1108. 6. doi:10.1080/13658810701851438

Mostern, R., Southall, H., and Berman, M.L., 2016. *Placing names: enriching and integrating gazetteers*. Bloomington: Indiana University Press. 6.

Murrieta-Flores, P. and Gregory, I., 2015. Further frontiers in GIS: extending spatial analysis to textual sources in archaeology. *Open Archaeology*, 1 (1), 5. doi:10.1515/opar-2015-0010

Murrieta-Flores, P. and Howell, N., 2017. Towards the spatial analysis of vague and imaginary place and space: evolving the spatial humanities through medieval romance. *Journal of Map & Geography Libraries*, 13 (1), 29–57. 10. doi:10.1080/15420353.2017.1307302

Rayson, P., *et al.*, 2017. A deeply annotated testbed for geographical text analysis: the Corpus of lake district writing. *In*: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'17. New York, NY, USA: ACM, 9–15. 6.

Reinhold, A., *et al*., 2017. Exploring deep mapping concepts: crosthwaites map and wests picturesque stations. *In*: *Proceedings of Workshops and Posters at the 13th International Conference on Spatial Information Theory (COSIT 2017)*, Lecture Notes in Geoinformation and Cartography. Cham: Springer, 265–273. 6.

Rupp, C., *et al*., 2013. Customising geoparsing and georeferencing for historical texts. *IEEE International Conference on Big Data*, Silicon Valley, CA, 59–62. 5.

Safier, N., 2014. The Tenacious Travels of the Torrid Zone and the global dimensions of geographical knowledge in the eighteenth century. *Journal of Early Modern History*, 18 (1–2), 141–172. 4. doi:10.1163/15700658-12342388

Santos, J., Anastácio, I., and Martins, B., 2015. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80 (3), 375–392. 3, 8. doi:10.1007/s10708-014-9553-y

Simon, R., *et al*., 2015. Linking early geospatial documents, one place at a time: annotation of geographic documents with recogito. *e-Perimetron*, 10 (2), 49–59. 8.

Smith, D.A. and Mann, G.S., 2003. Bootstrapping toponym classifiers. *In*: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*. Stroudsburg, PA, USA: ACL, 45–49. 15.

Snoussi, M., Gensel, J., and Davoine, P.A., 2012. Extending TimeML and SpatialML languages to handle imperfect spatio-temporal information in the context of natural hazards studies. *In*: *Proceedings of AGILE2012 conference*. Avignon (France): Springer, 117–122. 10. doi:10.1094/PDIS-11-11-0999-PDN

Southall, H., Mostern, R., and Berman, M.L., 2011. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5 (2), 127–145. 6. doi:10.3366/ijhac.2011.0028

Tobin, R., *et al*., 2010. Evaluation of georeferencing. *In*: *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10. New York, NY, USA: ACM, 7:1–7: 8.2.

Wing, B.P., 2015. Text-based document geolocation and its application to the digital humanities. Available from: https://repositories.lib.utexas.edu/handle/2152/40313. 8.

Withers, C.W.J., 1999. Reporting, mapping, trusting: making geographical knowledge in the late seventeenth century. *Isis*, 90 (3), 497–521. 4. doi:10.1086/384413

Withers, C.W.J., 2008. *Placing the enlightenment: thinking geographically about the age of reason*. Chicago, IL: University of Chicago Press. 4.

Withers, C.W.J. and Mayhew, R.J., 2011. Geography: space, place and intellectual history in the eighteenth century. *Journal for Eighteenth-Century Studies*, 34 (4), 445–452. 4. doi:10.1111/j.1754-0208.2011.00441.x

Won, M., Murrieta-Flores, P., and Martins, B. 2018. Ensemble Named Entity Recognition (NER): evaluating NER tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5. 5, 2. 8, 11, 13, 16, 17. doi:10.3389/fdigh.2018.00002