Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data

Francesca Frontini, Carmen Brando, Marine Riguet, Clémence Jacquot, Vincent Jolivet

Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data

FRANCESCA FRONTINI
Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa
CARMEN BRANDO
EHESS, CRH UMR 8558 (EHESS-CNRS)
MARINE RIGUET
Labex OBVIL – Université Paris-Sorbonne
CLÉMENCE JACQUOT
Université d'Artois, EA Grammatica
VINCENT JOLIVET
Labex OBVIL – Université Paris-Sorbonne

Abstract
This paper aims to discuss the challenges and benefits of the annotation of place names in literary texts and literary criticism. We shall first highlight the problems of encoding spatial information in digital editions using the TEI format by means of two manual annotation experiments and the discussion of various cases. This will lead to the question of how to use existing semantic web resources to complement and enrich toponym mark-up, in particular to provide mentions with precise geo-referencing. Finally the automatic annotation of a large corpus will show the potential of visualizing places from texts, by illustrating an analysis of the evolution of literary life from the spatial and geographical point of view. **Keywords:** digital literary studies; toponyms; semantic web; geographic databases; maps and visualizations.

Resumo
Este artigo aborda as dificuldades e as vantagens da anotação dos nomes de lugar em textos literários e de crítica literária. Começamos por realçar os problemas de codificação da informação espacial em edições digitais usando o formato TEI, através de duas experiências de anotação manual e da análise de diversos casos. Isto conduzirá à questão de como utilizar os recursos da web semântica para complementar e enriquecer a marcação de topónimos, em particular com georreferenciação rigorosa. Por último, a anotação automática de um grande *corpus* irá mostrar o potencial de visualização de locais a partir de textos, ilustrando a análise da evolução da vida literária segundo um ponto de vista espacial e geográfico. **Palavras-chave:** estudos literários digitais; topónimos; web semântica; bases de dados geográficas; mapas e visualizações.

MATLIT 4.2 (2016): 49-75. ISSN 2182-8830
http://dx.doi.org/10.14195/2182-8830_4-2_3

1. Introduction

We propose here an excursus into the interactions between literary studies and geographical information science. In particular, we shall examine the issue of correctly and efficiently annotating place names in literary texts and (literary) criticism. Clearly, the problems connected to these two tasks are similar but not identical. Literary texts may contain fictional places, while non-fictional texts mostly real ones. Also the goals of such annotations are different; in the first case the purpose of the annotation may be to study the interaction between setting and narrative space[1] within a single text, while in the second case it will often be to enable a diachronic analysis of large corpora in order to find trends and evolutions in the geographical distribution of literary centres and topics. Nevertheless, many similar problems arise, and it is useful to address the common issues of toponym annotation in the literary domain in a comprehensive way, as we shall do in this paper. We shall argue that some reflection is necessary to establish best practices for the appropriate annotation of place names in texts and for their linking to existing geographical databases to be able to retrieve information—typically but not exclusively geospatial information—about them.

The keen interest of researchers in the Humanities for the geographical dimension of information is not a recent phenomenon. The possibility of modelling, storing, analysing (via spatial analysis methods) and visualising geospatial information proposed by Geographical Information Systems (GIS) have been exploited particularly by archaeologists and by historians to study, trace and quantify phenomena taking place on the surface of the Earth. More recently, disciplines within the Digital Humanities (DH) have shown increasing interest in geospatial information. As a proof of this, a Geohumanities special interest group[2] is particularly active within the DH community. The availability of geographical databases such as interoperable gazetteers thanks to Linked Open Data (LOD) initiatives as well as the interactive cartography tools and technologies offered by the Web represents an undeniable opportunity for these communities to explore novel interdisciplinary research ideas.

For what concerns (digital) literary studies, some interesting projects have recently seen the light, inspired in part by the pioneering work of Franco Moretti (Moretti 2007) and Matthew Jockers (Jockers 2013). In general, these are large projects requiring interdisciplinary work among literature researchers, corpus linguists, geographers and cartographers, and others; at the same time, they allow researchers to gain new insights from working on large sets of data and from aggregating them in novel ways, which allows new

[1] For the theoretical complexity of analysing fictional spaces and for a proposal of visualisation of several spatial dimensions in fiction see Hones (2011) and Piatti *et al.* (2013).

[2] "GeoHumanities" http://www.geohumanities.org/. Accessed January 22, 2016.

and interesting spatial relations to become visible by projecting places on maps. We shall mention here just a few of the many existing initiatives of this type. The first, Literaturatlas,[3] based in Zurich, is devoted to the creation of a literary atlas of Europe. Cartography and graphic semiology techniques have traditionally been developed for mapping real spaces; this project instead aims at visibly rendering complex overlays of real and fictional geographies. For instance, they propose solutions to questions such as how to objectively represent the fictional area where a character's dream took place (Piatti *et al.,* 2009; Reuschel *et al.,* 2011; Piatti *et al.,* 2013). The second project is Spatial Humanities[4], based in Lancaster, which develops and applies methodologies for analysing unstructured texts—including large corpora of historical sources (and not exclusively literary ones)—within a GIS environment. As a case study, they constituted a corpus of 1,500,000 words on Lake District literature that was annotated for toponyms to allow researchers to investigate the "literary landscape" of this area. In particular, the use of "a hybrid corpus- and geographic-based methodology" labelled "geographic text analysis" can be used to gain new insight from the texts both by projecting the places on specially designed dot maps, but also by analysing associated concepts for places, by means of corpus linguistics techniques such as collocation extraction (Gregory *et al.* 2011; 2014; 2016). Similar investigations are infrequent for languages (and literatures) other than English; some exceptions are the GIS project at Språkbanken, the Swedish centre for language resources, producing geographic visualizations of large corpora of Swedish Literary texts (Borin *et al.,* 2014) and the SyMoGIH project[5] aiming to add spatial referencing to TEI documents (as well as images, and metadata) by means of an *ad hoc* developed GIS environment, whose resulting data is published as Linked Data (Beretta *et al.,* 2012; 2014). Finally it is important to mention Pelagios, a visual browser for geo-tagged datasets, where datasets can be texts but also archaeological collections, archive records, etc. (Simon *et al.,* 2012; Isaksen *et al.,* 2014). The Pelagios consortium is mostly devoted to investigating mentions of ancient places, but the technical infrastructure and methodology is applicable to any context.

The analysis of such projects shows great advances from a technological point of view, in particular for what concerns the *geoparsing* of texts—namely the technique for the automatic or semi-automatic detection of toponyms (see Leidner and Lieberman, 2011 for an overview)—but common practices

[3] "Ein Literarischer Atlas Europas." http://www.literaturatlas.eu/en/. Accessed January 22, 2016.
[4]"Spatial Humanities | TEXTS, GIS & PLACES." http://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/. Accessed January 22, 2016.
[5] http://www.symogih.org. Accessed January 22, 2016.

for toponym annotation and referencing are still difficult to identify[6], and the problem of how to concretely use existing geographical data sources and how to best enrich textual data in a standardised way that is in line with current practices in DH deserves further investigation from the perspective of identifying best practices for this type of research and allowing for cross-project reuse both of visualization tools and of annotated data.

In what follows, we are going to introduce the problem of place name annotation and detection within the framework of DH and computational linguistics. First, we shall introduce the problem of toponym annotation with external linked data sources and the Text Encoding Initiative (TEI) standard for the encoding of such information in texts. Then, the specific issues concerning the identification of toponyms in literature and in critical essays will be exemplified by two annotation experiments, one on fiction and one on criticism. These experiments will help us to identify some of the major issues (concerning mainly temporal and spatial vagueness), and to define the best way to tackle them in TEI. In view of these experiments, we will also be able to analyse in detail the best way to link place names in texts with external geographical databases published as Linked Data (resulting in the Web of Data) that can provide both referencing and additional information (notably geographic coordinates for geo-visualisation purposes); existing databases will be compared based on their advantages and disadvantages in terms of completeness and homogeneity. The use of uniform resource identifiers (URIs) will be then recommended and TEI annotation cases outlined. Finally, we shall briefly illustrate how automated algorithms for toponym recognition, powered by external resources, can be used for the geographical analysis of large quantities of texts. More specifically a further case study on a literary criticism corpus will show how the annotation of place names can help studying the evolution of literary life over space and time.

2. Named Entities and toponyms in text

Named Entities (NE) are linguistic expressions that stand like rigid designators (Kripke, 1980) for individuals; named entities normally include proper names of Persons, Geographical Places, Organizations, but also temporal references such as dates. So for instance "William Shakespeare", "Paris", "Sorbonne" are examples of NEs.

The manual annotation of NEs in texts is important for the production of richer digital editions, but also for the training of automatic Named Entity Recognition (NER) systems (see the extensive survey by Nadeau and Sekine,

[6] Indeed, project pages and papers often tend to focus on the visualizations and on the analyses that have been derived, rather than on presenting and discussing the annotated texts; when an annotation schema is present, it follows the TEI conventions as they will be presented in the next paragraphs.

2007). The most important problems in the annotation of NEs are represented by the detection of the actual boundaries of each mention in the text, by the attribution of each mention to a class, and by the disambiguation, namely the identification of the referent of the mention. Let us take the following example (in French):

> "Voilà ! J'avais eu affaire, rue de la Pépinière, près de la place Saint-Augustin, et je revenais par le boulevard Malesherbes en l'intention de prendre l'omnibus à la Madeleine. Tout à coup, au coin de la rue des Mathurins, un homme se dressa devant moi en criant : "Madame ou mademoiselle, [...]." (*Le passant de Prague*, Guillaume Apollinaire)]

Here we find five mentions of NEs, more specifically toponyms: three mentions of streets, one mention of a square and one referring to a building. Notice that the latter, Madeleine, is an ambiguous term as the same superficial form may refer in different contexts to the Church of la Madeleine in Paris, the square in which the church is located and a river in Belfort[7]. At the same time the same entity, such as the church in Paris in our case, may be referred to by using different superficial forms, such as "la Madeleine" and "l'église de la Madeleine". Such mentions pose problems for search and information retrieval in large collections of texts for research purposes, as a plain text search may produce very unclean results.

Enriching mentions with a link to a referent by means of a unique identifier is crucial for the semantic annotation of texts. This is done by pointing to an external resource, such as a URI in the LOD cloud. For instance, in natural language processing, the automatic annotation of NEs is generally accompanied by the linking of such entity mentions to a DBpedia link added to clarify which external entity is the referent of a given mention in the text. Such is the behaviour of the popular tool DBpedia Spotlight (Mendes *et al.,* 2011; for an overview of NEL systems see also Hachey *et al.,* 2013).

The purpose of NE annotation in DH is to enrich digital editions with such information that allows users to retrieve different mentions of the same entity in many texts (e.g. "M. Hugo", "Victor Hugo"), but also to link it to external sources of structured information (e.g. DBpedia[8] and Bibliothèque Nationale de France – BnF[9] entries for Victor Hugo) for disambiguation purposes. This information can later be used for text mining and querying (e.g. "find all mentions of authors born after 1750"), but also for aggregation

[7] Ambiguity can be high even when limiting the scope to the same class of entities, here to toponyms. Inter categorical ambiguity is even higher (Madeleine can also refer to a person or to the famous cookie); that is why the classification of entities is also an important step.
[8] http://fr.dbpedia.org/page/Victor_Hugo. Accessed January 22, 2016.
[9] http://data.bnf.fr/11907966/victor_hugo/. Accessed January 22, 2016.

and visualisation. A RDF query language, *SPARQL*,[10] can be used to retrieve all the information available for a given entity in the LOD cloud, by using the entity's identifier. This keeps the annotation in the text to a minimum, and enriches documents with an always-growing set of knowledge (see also Van Hooland, 2015).

Different typologies of texts contain different classes of NEs. Literary essays and fictional works typically contain classes such as places, authors, edition titles, organisations such as universities or publishing houses, fictional characters and places, places which existed in another epoch, etc.

Toponyms constitute a special case of NE, since they are referring directly to objects associated with portions of physical space. Their correct annotation and linking is crucial for their aggregation and cartographic visualisation, since they naturally allow for a spatial representation of the text as the projection on a map of all the locations mentioned in it. Annotation of places in texts further allows for the geographic search of texts, enabling users to search for texts that mention places located within a specific area (i.e. spatial queries). In order to do this, it is important to disambiguate each entity by linking it to an appropriate repository, containing or linking to as much geospatial information as possible. This is particularly important as there exists a special version of SPARQL, named *GeoSPARQL*,[11] specifically designed to handle LOD datasets containing spatial information (spatial operators are typically *intersection*, *within*, *touch*).

In this context, we are mostly interested in the specific problems identifying place names in texts, disambiguating them by providing an external referent using LOD sources, and by annotating them in a way that is compatible with current TEI standards.[12]

The TEI defines and maintains a widespread standard for the representation of texts in digital form. As to NEs annotation, it is possible to define organisations and persons using the XML tags, *OrgName* and *PersName*, respectively. The specifications related to place name annotations[13] propose the use of two XML tags *geogName* and *placeName*. The latter is used to annotate relative or absolute place names. Besides, it provides the possibility of adding information concerning the different levels of detail, for instance districts, areas, countries, settlements and blocs. The *placeName* tag may contain an *offset* tag that can be used to isolate text containing vague information related to a toponym; such is the case of the segment "north of" in the sentence "north of France" which provides directional information (north, south, west, east). The definition of vague places is used to indicate places for which

[10] http://www.w3.org/TR/rdf-sparql-query/. Accessed January 22, 2016.
[11] GeoSPARQL was published as standard by the Open Geospatial Consortium (OGC).
[12] http://www.tei-c.org/index.xml. Accessed January 22, 2016.
[13] http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ND.html. Accessed January 22, 2016.

no fixed boundary can be given (Jones *et al.*, 2008). The *geogName* tag is an alternative to the *placeName* tag; it allows one to distinguish the generic part of a toponym such as "Mont" in "Mont Saint-Michel".

As for referencing and disambiguating named entities in general and toponyms in particular, two strategies can be adopted within TEI. The first is that of use the attribute *key*, which can contain a textual identifier for the annotated entity. So for instance, if two places have the same name but refer to two different places their TEI annotation will contain different keys. Keys are internally defined and have just a disambiguating function.

Another strategy is to use the attribute *ref*, which contains an identifier within a source of reference, providing more information on the identified location. The source of information may be internal to the document, in the form of a list of places (encoded using the <listPlace> tag[14]), containing descriptions of the places mentioned in the text and including alternate names (with temporal information as to the time of use of each variant), location (for instance the country when the toponym is a city) and of course the geographic coordinates (<geo> tag[15]). Alternatively, the source is an already existing and publicly available one, in the form of a Linked Data set whose URIs can be directly used as links. Clearly the second strategy is the privileged one, as referents exist independently from any corpus and they are used and reused by larger communities and thus benefit from corrections and updates. Having instead a data silo describing places for each corpus would represent an important amount of redundant work and would not comply with Linked Data principles. The two strategies are exemplified in the following, for London UK and London, Ohio using TEI.

(1) Internal reference strategy:

```
<placeName ref="#London,_Ohio">London</placeName>
<placeName ref="#London">London</placeName>
<listPlace type="cities">
    <place xml:id="London,_Ohio">
    <placeName>London Ohio</placeName>
            <location>
             <country>USA</country>
            <geo>39.8875 - 83.4450</geo>
            </location>
    </place>
    <place  xml:id="London">
    <placeName>London</placeName>
```

[14] http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-listPlace.html. Accessed January 22, 2016.
[15] http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-geo.html. Accessed January 22, 2016.

```
            <location>
                <country>UK</country>
                <geo>51.969604 -2.893146</geo>
            </location>
        </place>
    </listPlace>
```

(2) DBpedia as an external reference:

```
<placeName ref="http://dbpedia.org/page/London,_Ohio">
            London</placeName>
<placeName ref="http://dbpedia.org/page/London">
            London</placeName>
```

You can easily see how much more efficient the second option is. URIs, just as normal web links, can be accessed online. But, unlike URLs, they contain structured data that is meant to be readable by machines as well. So in this case a machine can access the two links for London and London Ohio from DBpedia and automatically tell us that one is located in UK and the other in USA, as well as retrieve the geo-coordinates for both; in (1) this information has to be encoded in the document. If a whole text is annotated as in (2), aggregated counts such as the overall number of mentions for USA and UK cities can be produced without further manual intervention. We shall see later what kind of information is available in different data sets and why that is important.

In order to better investigate the types of problems related to the annotation of NE according to the TEI standard, we decided to perform an experiment of manual annotation of some of the texts already available in TEI format in the digital library of Labex OBVIL.[16] The next section describes this experiment and our findings.

3. Identification and classification issues with toponyms: two experiments on fiction and essay samples

The study of the history of literature, and more generally of ideas, requires the analysis of both primary and secondary sources. It is clearly interesting to study place names in literature, in order to see what real and imaginary places are mentioned and how this is important in creating the fictional space of the work. Moreover by analysing large quantities of texts, researchers can be able to identify what toponyms are mostly mentioned in which epochs. This is

[16] http://obvil.paris-sorbonne.fr/. Accessed January 22, 2016.

also true for essays, especially of past texts, which are an important source for the study of evolution of literary and cultural life over time.

We thus chose to perform two experiments on two texts that are part of the OBVIL digital library, and that were the object of current studies (Riguet 2015, in press); such texts were chosen based on the high frequency of toponyms mentioned and on the importance of the spatial aspects in both texts. The first is *L'Hérésiarque et cie*,[17] a collection of 23 short stories by Guillaume Apollinaire, published in 1910. Many of these short stories describe a character's travels and wanderings around the world. One of the most famous tales ("Le passant de Prague") uses the typical fictional figure of the *Wandering Jew* who describes here his journey through time and space. Many tales take place in Paris and allow us to precisely locate and follow the character's itinerary. Here, it would be interesting to study how the character interacts with the geographical features by means of spatial relations (e.g. goes along the river, enters the castle, crosses the bridge, etc.) in order to trace his journey. The annotated sample contains around 54,166 words. 395 place names were manually annotated.

The second is the famous essay by Ernest Renan, *Qu'est-ce qu'une nation*?[18] a historical lecture published in 1882 dealing with the basis of French national identity, and more specifically the influence of as well as the attitudes towards foreign nations in that epoch. The text contains around 8,500 words. 174 place names were manually annotated.

The results of the manual annotation presented some common issues for both texts. In these texts a place may be vague, it may be that political boundaries are imprecise (e.g. "Europe occidentale"), constantly changing (e.g. Europe in 1850 vs. Europe in 1950), or they are perceived differently by people ("several big cities of America", "the main cities of Europe", "the five corners of the World" (with capital W). Places may also be referenced relatively to another place, for instance "Southampton's suburbs", "a principality of northern Germany", "a small state of the Balkans". A place may also have alternative, vernacular names, (e.g. "Old World" for Europe). A place may have existed in the past such as "Babylon", "Gaule", "Russia". Moreover, a place may be symbolic or abstract such as "Hell" or "Heaven". Places can be referred to by using descriptions such as "the country of Italians", "the river of Paris", "the capital of Germany", "the country ruled by Philip VI". Moreover, some of these problems can combine, for instance, the "historical kingdom of Bohemia" is both an old place and a vague one because of its unstable boundaries.

Typically geo-political entities are polysemous, as they can be both places and organisations such as the "Roman Empire" or "Charlemagne's empire".

[17] http://obvil.paris-sorbonne.fr/corpus/apollinaire/apollinaire_heresiarque-et-cie.xml Accessed Januaryn22, 2016.
[18] http://obvil.paris-sorbonne.fr/corpus/critique/renan_nation.xml. Accessed January 22, 2016.

Similar examples are "the Vatican" and the "House of Habsburg". In other cases there is no ambiguity at the level of identification (it is clearly a place name) but the referent may be ambiguous, as in "Hraschin" that may refer to the castle of Prague but also to the specific district of the town according to context.

Current TEI specifications for place name annotation (as described above) allow us to deal with many of these cases, for instance, with vague places and composite places. First of all, given the problem in distinguishing between toponyms and generic place names, we propose the extended use of *placeName* to tag both proper toponyms and more general descriptions. Then we can make use of other TEI tags to solve specific problems. Here follow some issues and examples of annotation.

- Cases such as "le royaume de Juda" (the kingdom of Juda), "la Bohême" (Bohemia), whose borders are quite unstable and today include several different countries can be annotated using the *bloc* tag, that is normally recommended for a geo-political unit composed at least by two states or countries:

```
le <placeName>
        <bloc type="Nation">royaume de Juda</bloc>
</placeName>
```

- The same case is true for empires, such as "l'empire des Habsbourg":

```
l'<placeName>
        <bloc type="Nation">empire des Habsbourg</bloc>
</placeName>
```

- For other cases of vagueness that do not imply places with sub-parts ("la banlieue de Southampton", "un petit État des Balkans", "une principauté d'Allemagne du Nord", "une petite localité du Queensland") the *offset* tag seems a better solution:

```
la <placeName>
        <offset>banlieue de</offset>
        <settlement type="city">Southampton</settlement>
</placeName>

un <placeName>
        <offset>petit État des</offset>
        <bloc type="Union">Balkans</bloc>
</placeName>
```

Clearly, these examples pertain identification only, and do not remove the necessity of adding a *ref* attribute to provide for linking. In other words, TEI annotation can provide the user with information about the fact that the place is either vague or has internal subdivisions, but no straightforward way can be found to encode temporal information. In particular it is difficult to signal with TEI tags or attributes the fact that a place does not exist at present or that existed with different borders or a different name when the annotated text was written. The same is true of other types of information, for instance the fact that a given place is fictional or abstract. In fact, the complexity of providing temporal information for places makes it difficult to see how this kind of information could be actually provided within TEI as a textual mark-up. We strongly believe that such type of annotation is better stored in external databases and accessed there via linking, as is the case of geographical coordinates.

In the following paragraph we shall analyse in more detail the types of information that are available in LOD datasets that could be potentially used to link and enrich annotation of place names in texts. This analysis will help us to at least partly solve some of the problems left out by the present paragraph.

4. LOD for toponym linking

As we have seen, the usually inherent vagueness associated to toponyms makes difficult to systematically assign a unique identifier. Most existing geographic databases provide coverage only for Real World, currently existing places, such as geopolitical entities, geographic features, monuments, which are represented using various types of geometries (as points polylines, polygons) thanks to a Geodesic system (usually WGS84).

Some attempts have been made to include the temporal dimension in geographic databases; for instance Pleiades[19] is a gazetteer containing Mediterranean place names for Antiquities[20]. No existing resource offers the same amount of coverage for the Modern world but some information can be derived from existing resources; for instance in DBpedia Gare d'Orsay has a property specifying that it was in service between 1900 and 1979.

Given the aforementioned issues, the ideal database for place linking and annotation in texts would provide at least geo-coordinates for places, but at best more complex geometries such as areas, the period of existence of plac-

[19] http://pleiades.stoa.org/. Accessed January 22, 2016.

[20] Other similar LOD databases contain temporal information for a specific geographic area only, such as for instance the New York City Chronology of Place, a Linked Open Data Gazetteer ("NEH Grant Details: NYC Chronology of Place, a Linked Open Data Gazetteer." 2015. Accessed January 22, 2016. https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=HD-51618-12.

es when relevant, the possibility of setting the degree of fuzziness (yes or no), a satisfactory coverage for the targeted areas.

Geographic data is made available in the form of Linked Data mostly by government agencies and research communities. Several LOD datasets are available, but they all have pros and cons, and no optimal solution available for annotating literary texts. Each dataset is conceived based on its own perception of place. In general, the data is built from a geographic database point-of-view but sources such as Pleiades and Getty have a scope more compatible with the Humanities, in other words, they were created for the Humanities and by humanists. Table 1 summarises the different solutions available, considering criteria that may be relevant to the DH.

At first sight, it seems that the LinkedGeoData source (Stadler *et al.,* 2012), derived from the OpenStreetMap project[21], a collaborative project which aims at creating an open geographic database of the World, is the most convenient LOD data repository for our needs. Because of the possibility of assigning complex geometries to places, this is particularly useful to build maps. The external linking to other LOD sources, especially DBpedia, is quite interesting for enriching a digital edition. DBpedia provides direct access to additional, non-geographic information (for instance that the architect of the ancient Orsay train station is Victor Laloux). The multilingual support is also very important because we focus on French texts and in many cases we find most of place names (and their alternatives) only in English. However, the missing support to historical places and the missing temporal information about places are important drawbacks of this data set. We recently tested French DBpedia and Geonames for automatic linking of place names in digital humanities (Brando *et al* 2015); preliminary results showed that the former outperforms the latter for recall, in part due to the coverage of historical places. In the future, more experiments will have to be performed on some of the data sets listed above, considering different comparison criteria.

Apart from the correct treatment of temporal information, other research questions remain open that make it difficult to achieve an optimal annotation and linking. The first set of problems concern vagueness, in that some places are defined in a fuzzy way, which makes it difficult to associate them to a point or area. The second set of problems concerns the time dimension, which is particularly important for geopolitical entities (Gaule) or artificial landmarks (Gare d'Orsay), which may come into existence, change or cease to exist over time.

[21] http://openstreetmap.org. Accessed January 22, 2016.

LOD source/ criteria	DBpedia	Geonames	Linked Geo Data, derived from OSM	Getty Thesaurus of Geographic Names	Pleiades
Latitude/ Longitude	Yes	Yes	Yes	Yes	Yes
Geometries (location and form)	No	No	Yes	No	No
External linking	No	Yes	Yes	Yes	Yes
Vernacular knowledge	Yes	No	Yes	No	No
Multilingual support	Yes	Yes	Yes	Limited	No
Homogeneous World coverage / complete attribute data	No	No	No	No	No
Coverage of historical places	Few	Few	No	Mostly in English	Only Mediterranean and ancient places
Time information	Implicit	No	No	No	Yes

Table 1. Comparative table showing pros and cons of several available geographic DBs.

Finally, a set of problems is linked to fictional or symbolic places (Heaven or Hell). These entities have a clear spatial dimension, and have an important role in fictional narratives. They may be very complex and articulated (Dante's Map of Hell) but they too need to be retrievable when querying for the places in which the action takes place. In difficult cases, when places are not present in any existing LOD source, and others are too vague to be assigned an external URI, encoding space and time information directly within the TEI annotation could be a fallback strategy, though the best option in most cases would be to eventually create a dedicated resource to be published as Linked Data.

Overall, the presented experiments have shown that the annotation of real places is generally possible within the current TEI specifications, but an adequate linking target is crucial as not all existing resources contain the required information. The one exception to this is represented by abstract or fictional places as well as old ones, which pose severe problems, and might require deeper investigation (see Joliveau, 2009 for an interesting discussion).

Finally, access to temporal information is vital for DH texts, and more effort is needed by the DH and the GIS communities in order to create more appropriate geo-historical LOD sources.

The degree of connection of the chosen link is also important when different options are available; so for instance Geonames entries of places provide a link to the corresponding DBpedia entry, when this exists. Thus in such a case using a Geonames link to identify a toponym in the text provides immediate access to richer information. From the analysis of such issues sets of open questions emerge, that make it difficult to achieve an optimal annotation and linking[22].

Despite these problems, the use of external links in annotation can be crucial, since it allows for the retrieval of additional information and the cartographic representation of places. For instance, if we point to an external resource, we are able to access all the information about that place that resource can provide, such as the country in which it is contained, whether some geographical features are present in its vicinity (rivers, mountains, ...), the number of inhabitants, when it was founded and, if relevant, when it ceased to exist. This in turn means that a corpus annotated with such external identifiers can answer more complex queries and selected visualisations can be produced. So for instance, a corpus of literary essays might be used to separately retrieve mentions to literary centres from different geographical areas; a corpus of novels might be searched for locations that contain cities along a river; small centres can be contrasted to large centres, metropolises and capitals.

Clearly such aggregated and filtered queries can only make sense when dealing with large corpora, that can hardly be pre-processed using manual annotators. To this purpose, natural language processing techniques can be useful in automatically annotating named entities in large texts. In what follows, we present an experiment of automatic annotation, providing an overall analysis of the aggregated results with a focus on literary centres and nations over time. This experiment is meant to show how, despite all aforementioned problems, the detection of the spatial dimension can help to highlight interesting phenomena relating to the history of literature.

[22] A preliminary version the two texts discussed in the previous section with added toponym annotation and linking to Geonames can be found at https://github.com/cvbrandoe/REDEN/blob/master/input/<apollinaire_heresiarque-et-cie-gold.xml> <renan_nation_only_placeNameTag-gold.xml>. Accessed January 22, 2016.

5. French literature and the world: a preliminary experiment of automatic annotation and analysis of place names in the Corpus Critique

In this experiment, automatic Named Entity annotation and data aggregation is used to analyse the geographical spaces emerging from large quantities of texts. Although automated natural language processing tools cannot reach the levels of accuracy of a manual annotator, they are pretty accurate when a proper domain adaptation is performed, and when run on large corpora they can be used to extract valuable information from texts and to detect hidden facts and trends.

Here we want to use this approach to identify place names, notably those cities and nations that are most represented in the French literary discourse, in order to study the progressive increase of cosmopolitanism, the opening to foreign literature and the possible convergences with historical events. More specifically, the study of toponyms in a diachronic perspectives aims to identify how foreign nations slowly emerge in French literary discourse, with political and ideological implications. For instance, when exactly does Russian literature make its appearance in the French literary landscape? Which part does Germany have in these texts when the French-Prussian war breaks out in 1870? Or more generally, how does the literary discourse categorise the relationships between France and the rest of the world?

In order to answer these research questions, we have investigated the *Corpus Critique*, a diachronic corpus of French literary essays originally published between 1824 and 1932 by authors such as Bergson, Zola, Sainte-Beuve, Bourget, Faguet, Taine, Brunetière, Lamartine and the Goncourts. The corpus contains texts that are crucial for the understanding of the French culture over time, some dealing with literary criticism others with history, politics, science and philosophy and it is used to carry out diachronic studies in the history of literature and ideas. This collection is part of the Labex OBVIL digital library, and is made available in open access in TEI format.[23] Each text contains indications as to its publication date, and this allowed us to analyse the geographical references in a diachronic perspective, to observe the evolution of literary life and literary discourse.

For the automatic extraction of place names from the *Corpus Critique* we exploited a natural language processing pipeline that has been particularly adapted to work with French literary essays. It is composed of UNERD (Mosallam *et al.,* 2015[24]), a Named Entity Recognition detector, and REDEN (Frontini *et al.*, 2015; and Brando *et al.,* 2015), a Named Entity Linking tool. They automatically recognise mentions of places, persons and organisations in a text by using linguistic information and pre-loaded dictionaries, and

[23] http://obvil.paris-sorbonne.fr/corpus/critique/. Accessed January 22, 2016.
[24] For the UNERD version that was domain adapted for Corpus Critique see http://obvil-dev.paris-sorbonne.fr/unerd/unerd-tei/. Accessed January 22, 2016.

connect them to existing linked data sources. In the present case only place names were taken into account. As for the external geographical sources, DBpedia and Geonames were used.

By ranking results by number of occurrences of place mentions, we first derive a list of seven influential nations of the epoch, namely, France, Italy, England, Spain, Russia, Germany, and Greece. Subsequently we treat these occurrences as a quantitative variable for building a map and use it to project a circle onto the corresponding nation with a diameter proportional to the frequency of its mentions; by sliding the temporal bar (where the minimal time unit is the year), the data displayed on the map dynamically changes as mentions are filtered by the year of publication of the texts where they appear. Besides analysing the results in these maps, we also consulted the frequencies of mentions of other toponyms such as important cities of these nations. Not surprisingly, France is the most cited toponym throughout the whole century, a fact that highlights the nationalism of French literary discourse (see Figure 1). Other frequently cited countries relate to collective representations of that age: Greece (Grece, Athene) as a cradle of culture and creativity, Italy (Italie, Rome) as an artistic model since the Renaissance. But the diachronic perspective allows us also and most crucially to analyse the evolutions that took place during the 19th century in the collective representation of foreign nations. Russia for instance is only cited ten times before 1880, but imposes itself in the critical discourse between 1880 and 1900 (see Figure 3), only to become invisible again: this peak clearly corresponds to the discovery of Russian literature (and of the works of Dostoyevsky and Gogol in particular) in France. As for Germany, it progressively emerges with a stronger and stronger presence between 1870 and 1920, to become the third most cited toponym after France and Paris. More generally, we can observe an increasing interest in exoticism and in the broadening of the French cultural horizon, with the introduction in 1890 of places that were altogether absent from the literary discourse before, such as Africa and Japan.

These results offer a first view on the geographic dimension of the *Corpus Critique*, helping the researcher to validate prior assumptions or guiding in further research. The analysis of the results shows that such a basic NLP approach can only offer partial solutions. Indeed, the automatic extraction of toponyms deserves to be enriched by other forms of text mining, such as adding names of human groups (ethnonyms), which are normally annotated in TEI as organization names (*orgName*), and adjectives of nationality (as in "la littérature russe", Russian literature). On the other hand, the detection and disambiguation of references to places thanks to geographical databases in the form of linked data makes it possible to represent a global cartography that shows, in a dynamic and interactive way, the relationships between France and foreign nations according to their representation in the literary discourse of the 19th century. This, we believe, is sufficient to give the reader an idea of what can be achieved by a more thorough annotation.

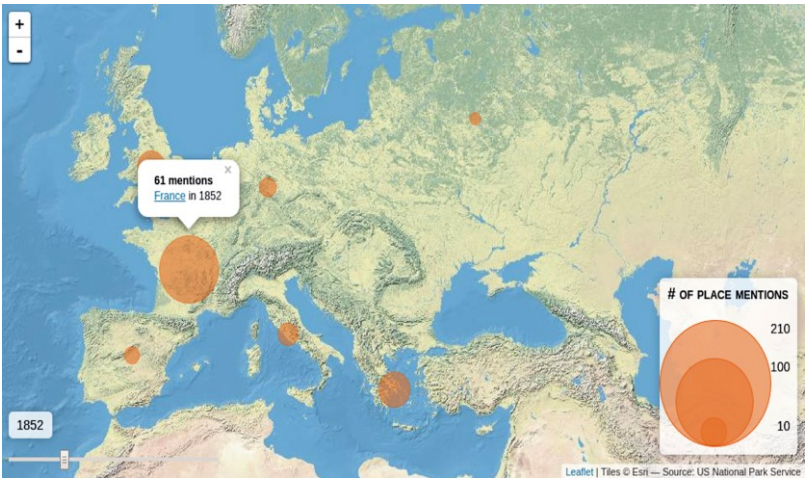

Figure 1. Mentions per nation in the *Corpus Critique* in 1852 (France: 61; Italy: 7; England: 13; Spain: 4; Russia: 2; Germany: 5; Greece: 17).

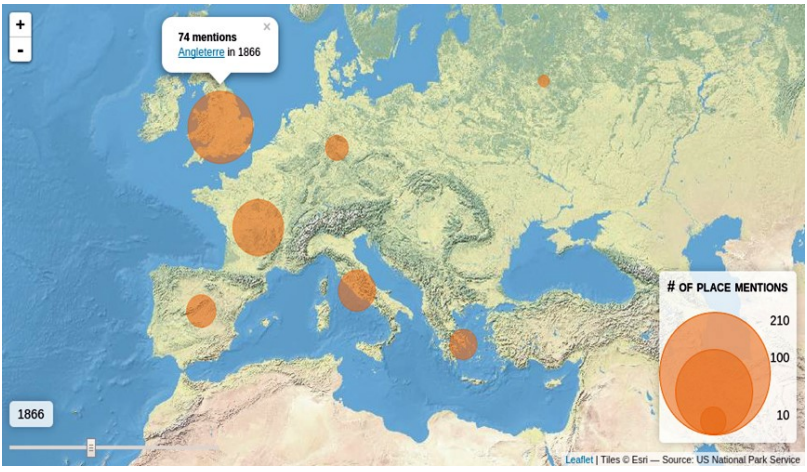

Figure 2. Mentions per nation in the *Corpus critique* in 1866 (France: 45; Italy: 24; England: 74; Spain: 15; Russia: 2; Germany: 9; Greece: 13).

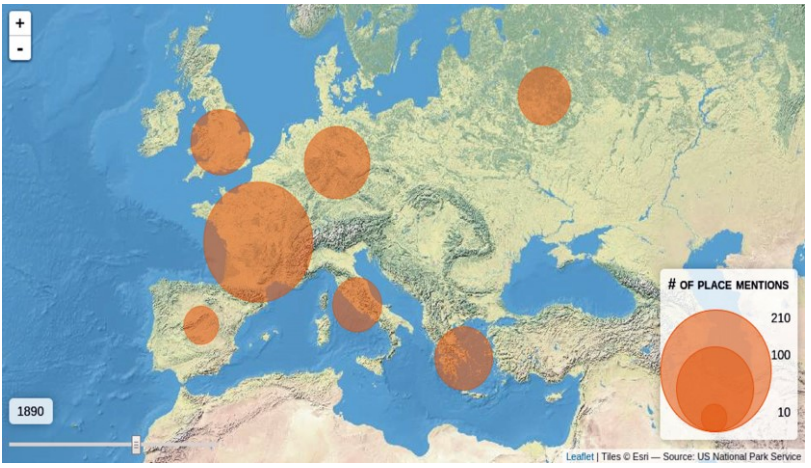

Figure 3. Mentions per nation in the *Corpus Critique* in 1890 (France: 207; Italy: 42; England: 61; Spain: 21; Russia: 48; Germany: 75; Greece: 58).

6. Future developments and conclusion

In this paper, we have discussed the problem of how to annotate place names in texts that have relevance for literary analysis and literary criticism (both primary and secondary sources), as well as for history of ideas in general; we have shown how to annotate them using TEI standards, and how to add references to external data sources in the LOD cloud in order to enrich the texts with additional information. We have also seen that not all LOD sources are the same, and that the ideal referencing of places in texts would require the perfectioning of existing resources with further knowledge, in particular about time spans for real places, as well as the creation of new resources, in particular for fictional places. Finally, we have proposed an example of what types of analysis are enabled by combining annotation of place names in texts and external geographical information.

Currently, Labex OBVIL is continuing both the manual annotation and the automatic analysis of place names in the digitised texts of its online library. More specifically, researchers are currently analysing the Apollinaire corpus, extending the work presented in this paper − on *L'Hérésiarque et cie* − to the *Calligrammes*, an anthology of poems. In the case of fiction and in particular poetry, the identification and spatial representation of toponyms has not only the function of identifying diachronic trends, but also of investigating how different spaces contribute to and enrich the fictional and poetic description.

Further developments from these premises could take different directions. First of all, we are working towards the creation of an easy-to-use open-source web-based instrument for the TEI-compliant annotation and linking of place names in texts to existing gazetteers, using the aforemen-

tioned NLP pipeline and allowing for manual correction. In the case of missing information, a TEI-compliant local index of places (<listPlace>) could also be generated and used to integrate information. *Ad hoc* visualization could then be automatically generated, deriving the geographic coordinates of places both from the local index and from the LOD sources. It is well known that the adoption of standards is promoted by providing freely available tools that support such standards. In this case the tool would allow researchers to generate cartographic projections of their corpora by using TEI. As a related issue, it would be interesting to find ways to connect local indexes of places found in digital TEI editions to the main geo-data sets such as DBpedia or Geonames so that the additions made by individual researchers can benefit the whole community.

Secondly, the semantics of place mentions could be made more complex, especially for fiction, differentiating between places that are just mentioned, and places where the action actually takes place. A complex taxonomy is proposed in Piatti *et al.* (2013), allowing for the annotation of dreamt, longed for or remembered places. In such cases as these, an extension of TEI would be required, as this information is clearly mention-specific, and belongs in the text. An interesting move in this direction is found in Ciotti *et al.* (2014), who propose an Open Annotation Data Model (OA) that can be used to make more complex annotations of entities in TEI texts, including toponyms.[25]

Thirdly, it should be possible to annotate and collect indirect geographical information also from textual elements that, though not toponyms, bear a relationship to places. So for instance, mentions of names of nationality, such as "Italian", "German", "American" could enrich and better substantiate the analysis carried out on the *Corpus Critique* on the relationship between French literary discourse and the rest of the world. At the same time, here too a careful reflection on the annotation of such elements in TEI is required, as they cannot be treated as place names.

Finally, the relationship between mentions of places and other parts of the text is very important to extract ideas, sentiments and opinions associated to different places. Collocations could be automatically extracted for place names in texts to retrieve names or adjectives recurrently associated to certain toponyms.[26] So for instance, it could be possible to see if certain places or nations are associated with specific literary movements, or have a positive or negative connotation. This too could be used to enrich analysis aiming to identify changes in the attitudes towards certain nations over time.

[25] To cite the authors themselves, "we can define OA as an RDF vocabulary (formally expressed in OWL 2), which allows the expression of the relationship between an annotation and its object".

[26] A similar approach is proposed in Murrieta-Flores *et al.* 2015 for historical texts and in Gregory *et al* 2016 for literary texts.

Acknowledgments

This work has been done within the LABEX OBVIL project, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme "Investissements d'avenir" under the reference ANR-11-IDEX-0004-02; it was also partly supported by a "Frenand Braudel" incoming scholarship from the Fondation Maison Sciences de l'Homme, Paris.

References

BERETTA, Francesco, and Pierre Vernus (2012). "Le projet SyMoGIH et la modélisation de l'information: Une opération scientifique au service de L'histoire." *Les Carnets Du LARHRA* 1: 81–107.

BERETTA, Francesco, Djamel Ferhod, Séverine Gedzelman, and Pierre Vernus (2014). "The SyMoGIH Project : Publishing and Sharing Historical Data on the Semantic Web." *Digital Humanities 2014. Conference Abstracts*. EPFL, Lausanne / UNIL, Lausanne. 469–470. https://halshs.archives-ouvertes.fr/halshs-01097399

BORIN, Lars, Dana Dannélls, and Leif-Jöran Olsson (2014). "Geographic Visualization of Place Names in Swedish Literary Texts." *Literary and Linguistic Computing* 29.3: 400–404. doi:10.1093/llc/fqu021.

BRANDO, Carmen, Francesca Frontini, and Jean-Gabriel Ganascia (2015a). "Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets." *New Trends in Databases and Information Systems*. Communications in Computer and Information Science, Springer: 505–14.

———— (2015b). "Linked data for toponym linking in French literary texts." *Proceedings of the 9th Workshop on Geographic Information Retrieval (GIR '15)*. Eds. Ross S. Purves and Christopher B. Jones. ACM, New York, NY, USA, Article 3, 2 pages. doi:10.1145/2837689.2837699.

CIOTTI, Fabio, Maurizio Lana, and Francesca Tomasi (2014). "TEI, Ontologies, Linked Open Data: Geolat and Beyond." *Journal of the Text Encoding Initiative* 8 (December). doi:10.4000/jtei.1365.

FRONTINI, Francesca, Carmen Brando, and Jean-Gabriel Ganascia (2015). "Semantic Web based Named Entity Linking for Digital Humanities and Heritage Texts." *Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference*: 77-88.

GREGORY, Ian N., Andrew Hardie (2011). "Visual GISting: Bringing Together Corpus Linguistics and Geographical Information Systems." *Literary and Linguistic Computing* 26.3: 297–314. doi:10.1093/llc/fqr022.

GREGORY, Ian, Alistair Baron, David Cooper, Andrew Hardie, Patricia Murrieta-Flores, and Paul Rayson (2014). "Crossing Boundaries: Using GIS in Literary Studies, History and Beyond." *Collections électroniques de l'INHA. Actes de Colloques et Livres En Ligne de l'Institut National D'histoire de L'art*. INHA. https://inha.revues.org/4931.

GREGORY, Ian, and Christopher Donaldson (2016). "Geographical Text Analysis: Digital Cartographies of Lake District Literature." *Literary Mapping in the Digital Age*. Eds. David Cooper, Christopher Donaldson, and Patricia Murrieta-Flores. London: Routledge. 67–87.

GROSSNER, Karl, Krzysztof Janowicz, and Carsten Keßler (2016, forthcoming). "Place, Period, and Setting for Linked Data Gazetteers." *Placing Names: Enriching and Integrating Gazetteers*. Eds. Merrick Lex Berman, Ruth Mostern, and Humphrey Southall. Bloomington, IN: Indiana University Press.

HACKEY, Ben, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran (2013) "Evaluating Entity Linking with Wikipedia." *Artificial Intelligence* 194: 130–50. doi:10.1016/j.artint.2012.04.005.

HONES, Sheila (2011). "Literary Geography: Setting and Narrative Space." *Social & Cultural Geography* 12.7: 685–699.

KRIPKE, Saul (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

JANOWICZ, Krzysztof (2009). "The Role of Place for the Spatial Referencing of Heritage Data." *Proceedings of the Cultural Heritage of Historic European Cities and Public Participatory GIS Workshop*: 17–18.

ISAKSEN, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares (2014). "Pelagios and the Emerging Graph of Ancient World Data." *Proceedings of the 2014 ACM Conference on Web Science*. WebSci '14. New York, NY: ACM. 197–201. doi:10.1145/2615569.2615693.

JOCKERS, Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. Chicago, IL: University of Illinois Press.

JOLIVEAU, Thierry (2009). "Connecting Real and Imaginary Places through Geospatial Technologies: Examples from Set-Jetting and Art-Oriented Tourism." *The Cartographic Journal* 46.1: 36–45.

JONES, Christopher B., Ross S. Purves, Paul D. Clough, and Hideo Joho (2008). "Modelling Vague Places with Knowledge from the Web." *International Journal of Geographical Information Science* 22.10: 1045–1065.

LEIDNER, Jochen L., and Michael D. Lieberman (2011). "Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language." *SIGSPATIAL Special* 3.2: 5–11. doi:10.1145/2047296.2047298.

MENDES, Pablo N., Max Jakob, Andrés García-Silva, and Christian Bizer (2011). "DBpedia Spotlight: Shedding Light on the Web of Documents." *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11. New York, NY, USA. ACM: 1–8. doi:10.1145/2063518.2063519.

MORETTI, Franco (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London, New York: Verso.

MOSALLAM, Yusra, Alaa Abi-Haidar, and Jean-Gabriel Ganascia (2014). "Unsupervised Named Entity Recognition and Disambiguation: An Ap-

plication to Old French Journals." *Advances in Data Mining. Applications and Theoretical Aspects*. Springer: 12–23.

MURRIETA-FLORES, Patricia, and Ian Gregory (2015). "Further Frontiers in GIS: Extending Spatial Analysis to Textual Sources in Archaeology." *Open Archaeology* 1.1: 166-175. doi:10.1515/opar-2015-0010.

NADEAU, David, and Satoshi Sekine (2007). "A survey of Named Entity recognition and classification." *Lingvisticae Investigationes* 30.1: 3–26. doi:10.1075/li.30.1.03nad.

PIATTI, Barbara, Anne-Kathrin Reuschel, and Lorenz Hurni (2013). "Dreams, Longings, Memories–Visualising the Dimension of Projected Spaces in Fiction." *Proceedings of the 26th International Cartographic Conference,* Dresden. http://www.literaturatlas.eu/files/2014/01/Piatti_ICC2013_final.pdf

PIATTI, Barbara, Hans Rudolf Bär, Anne-Kathrin Reuschel, Lorenz Hurni, and William Cartwright (2009). "Mapping Literature: Towards a Geography of Fiction." *Cartography and Art*. Amsterdam: Springer. 1–16.

REUSCHEL, Anne-Kathrin, and Lorenz Hurni (2011). "Mapping Literature: Visualisation of Spatial Uncertainty in Fiction." *The Cartographic Journal* 48.4: 293–308.

RIGUET, Marine (in press). "L'impact de la physiologie dans la critique littéraire de la fin du XIXe siècle: l'exemple de Claude Bernard." *Actes du colloque Littérature et Science au XIX siècle*. Eds. Elsa Courant et Romain Enriquez. ENS Ulm. *Épistémocritique*.

———— (2015). "Les éditions numériques de textes littéraires par le Labex OBVIL: la critique littéraire de 1850 à 1914." Presented at *Journée d'études HumaN'Doc*, Bibliothèque nationale de France, November 2015. 26. Jan. 2016. https://www.youtube.com/watch?v=gbzIMgngo1g.

SIMON, Rainer, Elton Barker, and Leif Isaksen (2012). "Exploring Pelagios: A Visual Browser for Geo-Tagged Datasets." *International Workshop on Supporting Users' Exploration of Digital Libraries*. Paphos, Cyprus: 23-27.

STADLER, Claus, Jens Lehmann, Konrad Höffner, and Sören Auer (2012). "LinkedGeoData: A Core for a Web of Spatial Open Data." *Semantic Web* 3.4: 333–354.

VAN HOOLAND, Seth, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle (2015). "Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections." *Digital Scholarship in the Humanities* 30.2: 262-279. doi:10.1093/llc/fqt067.

Annex

TEI file[27]	# TEI documents	Publication year
chamfort_ebauches.xml	1	1824
sainte-beuve_derniers-portraits.xml	1	1852
murger_propos-ville.xml	1	1853
nisard_histoire-01.xml; pontmartin_causeries-litteraires.xml	2	1854
lamartine_cours-familier-01.xml; lamartine_cours-familier-02.xml; taine_saint-simon.xml	3	1856
lamartine_cours-familier-03.xml; lamartine_cours-familier-04.xml; pontmartin_causeries-samedi.xml	3	1857
lamartine_cours-familier-05.xml; lamartine_cours-familier-06.xml	2	1858
sainte-beuve_portraits-01.xml; sainte-beuve_portraits-02.xml	2	1862
renan_vie-de-jesus.xml	1	1863
deschanel_physiologie.xml; sainte-beuve_portraits-03.xml; taine_positivisme-anglais.xml	3	1864
barbey-aurevilly_romanciers.xml; janet_crise-philo.xml	2	1865
taine_litterature-anglaise1.xml	1	1866
baudelaire_curiosites-esthetiques.xml	1	1868
taine_philosophie-art-grece.xml	1	1869
vacherot_science-conscience.xml	1	1870
janet_problemes-xix.xml	1	1872
gautier_portraits-contemporains.xml	1	1874
taine_france-t1.xml	1	1875

[27] TEI files can be accessed online one by one preceding the file names by the following URL http://www.obvil.paris-sorbonne.fr/corpus/critique/ Accessed January 22, 2016.

rod_assommoir.xml	1	1879
barbey-aurevilly_goethe-diderot.xml; barbey-aurevilly_poesie.xml; brunetiere_etudes-critiques-01.xml; charpentier-paul_mal-du-siecle.xml; st-victor_masques1.xml	5	1880
egger_parole.xml; stapfer_etude-litterature-moderne.xml; zola_naturalisme.xml; zo-la_roman-exp?rimental.xml	4	1881
brunetiere_?tudes-critiques-02.xml; re-nan_nation.xml; st-victor_hommes-dieux.xml	3	1882
brunetiere_roman-naturaliste.xml; re-nan_reforme.xml	2	1883
becq-de-fouquieres_art-mise-en-scene.xml; guyau_problemes-esthetique.xml	2	1884
deschanel_romantisme.xml; savine_etapes-naturaliste.xml	2	1885
lemaitre_contemporains1.xml; le-maitre_contemporains2.xml; pardo-bazan_naturalisme.xml	3	1886
brunetiere_banqueroute-du-naturalisme.xml; brunetiere_etudes-critiques-03.xml; ca-ro_sand.xml; goncourt-edmond-et-jules_journal-01.xml; goncourt-edmond-et-jules_journal-02.xml; le-maitre_contemporains3.xml; nisard_essais-ecole-romantique.xml; renan_discours-et-conferences.xml	8	1887
france_vie-litteraire-01.xml; goncourt-edmond-et-jules_journal-03.xml; goncourt-edmond-et-jules_prefaces-et-manifestes.xml; henne-quin_critique-scientifique.xml; morice_demain-questions-esthetique.xml	5	1888
bergson_conscience.xml; guyau_art.xml; hen-nequin_ecrivains-francises.xml; lemai-tre_impressions-03.xml; nisard_aegri.xml; nisard_histoire-02.xml; nisard_histoire-litterature-03.xml; st-victor_theatre.xml	8	1889
barbey-aurevilly_litterature-etrangere.xml; brunetiere_nouvelles-questions-critique.xml; france_vie-litteraire-02.xml; goncourt-edmond_journal-04.xml; lanson_conseils.xml; le-goffic_romanciers-d-aujourd-hui.xml; le-maitre_impressions-04.xml; renan_avenir-	9	1890

science.xml; renard_princes-critique.xml		
faguet_politiques-moralistes-01.xml; france_vie-litteraire-03.xml; goncourt-edmond_journal-05.xml; huret_enquete-litteraire.xml	4	1891
france_vie-litteraire-04.xml; goncourt-edmond_journal-06.xml; lemaitre_impressions-06.xml; rod_idees-morales.xml	4	1892
barine_musset.xml; lemaitre_impressions-07.xml	2	1893
doumic_ecrivains.xml; goncourt-edmond_journal-07.xml; jarry_divers.xml; lemaitre_impressions-05.xml; monod_maitres-histoire.xml; renard_critique.xml	6	1894
albalat_mal-decrire-roman-contemporain.xml; boutroux_lois-naturelles.xml; brune-tiere_science-et-religion.xml; durkheim_regles-methode-sociologique.xml; goncourt-edmond_journal-08.xml; lemaitre_impressions-08.xml	6	1895
bergson_matiere.xml; doumic_jeunes.xml; goncourt-edmond_journal-09.xml; gour-mont_masques1.xml; gour-mont_masques2.xml; le-maitre_contemporains6.xml; le-maitre_impressions-09.xml	7	1896
durkheim_empirisme-rationaliste-de-taine.xml	1	1897
bazalgette_esprit.xml; lemaitre_impressions-10.xml; rod_essai-sur-goethe.xml	3	1898
gourmont_langue.xml; le-maitre_contemporains4.xml; le-maitre_contemporains5.xml; le-maitre_contemporains7.xml	4	1899
barres_taine.xml; bergson_rire.xml; gour-mont_culture-des-idees-1.xml	3	1900
souriau_imagination-artiste.xml	1	1901
albalat_formation.xml; beaunier_poesie.xml; brunetiere_metaphysique-positiviste.xml; faguet_politique-comparee.xml; gour-mont_chemin.xml; gourmont_style.xml; sega-len_observation-medicale.xml	7	1902

bazalgette_latin.xml; taine_derniers-essais.xml	2	1903
faguet_en-lisant-nietzsche.xml	1	1904
albalat_ennemis.xml; bourget_etudes1.xml; bourget_etudes3.xml; gourmont_promenades-philosophiques-1.xml	4	1905
bougle_idees_egalitaires.xml; faguet_anticlericalisme.xml; souriau_reverie-esthetique.xml	3	1906
lasserre_romantisme-francais.xml	1	1907
gourmont_promenades-philosophiques-2.xml	1	1908
flat_femmes.xml; ghil_poesie-scientifique.xml	2	1909
faguet_etudes-litteraires-18e.xml; faguet_rousseau-contre-moliere.xml	2	1910
durkheim_jugements-de-valeur.xml; gheon_directions.xml	2	1911
bourget_pages-de-critique.xml	1	1912
dupuy_poetes-et-critiques.xml; equilbecq_litterature-merveilleuse-des-noirs.xml; faguet_la-fontaine.xml; gaultier_bovarysme.xml	4	1913
brunetiere_evolution-des-genres.xml	1	1914
bergson_france.xml	1	1915
barres_familles.xml	1	1917
bourget_essais-psychologie-01.xml; bourget_essais-psychologie-02.xml; lemaitre_impressions-11.xml	3	1920
bergson_duree.xml; daudet-leon_stupide19e.xml	2	1922
faguet_art-de-lire.xml; ghil_dates-et-oeuvres.xml	2	1923
albalat_souvenirs.xml	1	1924
albalat_comment.xml; gourmont_promenades-philosophiques-3.xml	2	1925
lasserre_romantiques.xml; souday_gide.xml	2	1927

bourget_temoignages-2.xml; bourget_temoignages.xml	2	1928
souday_livres-du-temps-02.xml	1	1929
souday_livres-du-temps-03.xml	1	1930
bergson_sources.xml	1	1932
Total: 171		

Table 2. List of TEI documents of the *Corpus Critique* used in the experiments.

Mention	**# occurrences by period**							
	1824-1858	**1862-1869**	**1870-1879**	**1880-1889**	**1890-1899**	**1900-1909**	**1910-1917**	**1920-1932**
France	458	300	199	821	845	581	207	258
Italy	250	141	29	206	186	107	33	115
England	76	144	35	246	160	127	23	42
Spain	43	41	13	276	75	49	10	24
Russia	23	49	10	80	120	40	12	45
Germany	52	41	28	198	178	99	43	62
Greece	119	144	29	249	114	55	36	81

Table 3. Number of mentions for seven nations classed by period.