A multi-layer markup language for geospatial semantic annotations

Ludovic Moncla, Mauro Gaio

A Multi-Layer Markup Language for Geospatial Semantic Annotations

Ludovic Moncla[*] [†]
Universidad de Zaragoza
C/ María de Luna, 1 - Zaragoza, Spain
moncla.ludovic@gmail.com

Mauro Gaio
LIUPPA - EA 3000
Av. de l'Université - Pau, France
mauro.gaio@univ-pau.fr

ABSTRACT

In this paper we describe a markup language for semantically annotating raw texts. We define a formal representation of text documents written in natural language that can be applied for the task of *Named Entities Recognition* and *Spatial Role Labeling*.

The proposal relies on a multi-layer annotation process based on a core generic layer, which can be freely adapted into more specific layers depending on the intended goal. Our markup language is based on the TEI Guidelines[1] to propose a generic and extensible markup language. This language is particularly dedicated for the text mining task and ready to use to be layered with more semantic relationships between elements of the text.

We show the feasibility of this proposal from a generic annotation of texts describing itineraries toward a geospatial semantic annotation.

CCS Concepts

•Information systems → Data encoding and canonicalization; Language models; *Information extraction;* **•Applied computing → Extensible Markup Language (XML);**

Keywords

geo-semantic tagging; text annotation; expanded named entity

[*]Laboratoire COGIT
73 av. de Paris - 94160 Saint-Mandé, France
[†]LIUPPA - EA 3000
Av. de l'Université - Pau, France

[1]TEI is an international standard for textual markup defined by the TEI Consortium. It provides a guide to best practices for interchange and encoding of textual material in digital form. http://www.tei-c.org/Guidelines/P5/

GIR '15, November 26-27, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2837689.2837700

1 Motivation and Background

Two categories of markup languages can be considered in geospatial tasks, those dedicated to the encoding of spatial data (e.g., GML, KML) and those dedicated to the annotation of spatial or spatio-temporal information in texts, such as in semantic role parsing [4, 10]. SpatialML [6] is more focused on the annotation of static spatial information and does not provide any support to identify spatio-temporal information such as motion. ISO-Space [12] annotates spatio-temporal information and has been also designed for capturing implicit spatial information. Although ISO-Space seems a comprehensive standard for capturing spatio-temporal information, some of its elements remain complex and are not really suited for a fully automatic process. Otherwise TEI is a standard for textual markup and provides a guide to best practices for interchanging and encoding textual information. TEI has been designed to be the more generic and adaptable as possible and provides a generic framework particularly adapted for the customization.

The existing markup languages dedicated to the annotation are more focused on the specification of relations (spatial or spatio-temporal) than in named entities (NE). Standard markup languages consider NE as being only composed of a pure proper name, whereas we consider more complex expressions known as expanded named entities (ENE) [7, 15].

ENE concept is particularly essential in any task that aims to retrieve all or part of evocative context of the NE, such as disambiguation task. For those reasons, we propose the specification of a multi-layer markup language adapted to the annotation of both ENEs and spatial/spatio-temporal relations. The proposed TEI-compliant markup language is based on a core generic layer, dedicated to the annotation of NE, which can be used to share pre-processed corpus of documents. Finally the language is designed to be compliant with an automatic process of annotation.

Although still at an early stage of development, the proposed markup language was applied for a problem of automatic information extraction and toponym resolution described in [8] and for a problem of automatic itinerary reconstruction described in [7].

The remainder of the paper is structured as follows. Section 2 describes the 'core generic layer' of our multi-layer annotation language. Then, Section 3 describes the adaptation of the generic language for a specific semantic role. Finally, Section 4 summarises and concludes this paper.

2 A Generic Markup Language for Expanded Named Entity Representation

The main elements of our proposed generic markup specification are categorized in two groups, the first one refers to the elements contained within expanded named entities, the second one refers to the standard elements describing the text segmentation and a set of global attributes defined by the TEI P5 Guidelines which are available for all TEI elements.

2.1 Expanded Named Entity Representation

2.1.1 Principles

According to [5] there are two categories of proper names: pure and descriptive. Pure proper names can be simple (i.e., composed of a single lexeme) or complex (i.e., composed of several lexemes) and are only composed of proper names. Descriptive proper names refer to a composition of proper names and common names (i.e., expansion). In other words, descriptive proper names overlap pure proper names. Descriptive proper names refer to a NE built with a pure proper name and a descriptive expansion. This expansion can change the type (e.g., location, person, etc) of the initial pure proper name.

An ENE may contain an entity built with both categories of proper names (i.e., pure and descriptive), and that can be composed of one or more concepts, whereas most of works of NER are usually only considering pure proper names. This provides us, in particular, to move beyond reduction of place to a name and a set of coordinates, a model still predominant in Geographic Information Science as specified by [11]. We define several levels of overlapping, (0, 1, 2, etc.) for the representation of ENE. Each level is encapsulated in the previous level.

Level 0. refer to pure proper names. It can be seen as the core component of an ENE. Thus, we consider NE as a special kind of ENE. The following examples (1 - 3) show some examples of entity of level 0:

(1) [Cardiff] → one entity (location)

(2) [Balaruc-le-Vieux] → one entity (location)

(3) [Charles de Gaulle] → one entity (person)

Level 1. refer to descriptive proper names composed of a pure proper name (i.e., an entity of level 0) and a common noun (i.e., expansion). The following examples (4 - 6) show the representation of ENE. We can notice that in these cases, descriptive expansions do not change the intrinsic nature of the object described by the proper name.

(4) commune de [Balaruc-le-Vieu]

(5) comunidad autónoma de [Aragón]

(6) général [Charles de Gaulle]

However, when the associated term is not equal to the intrinsic or default type of the pure proper name, it defines a new entity that overlaps the pure proper name. The following examples (7 - 10) illustrate that an entity may contain the name of another entity, and that the new entity may have a different type. For instance, a sentence containing the expanded named entity "sindaco di Venzia" (sindaco = mayor) is referring to the person and not necessarily to the city.

(7) [Cardiff] Downtown→ two entities, **Cardiff** (location) and **Cardiff Downtown** (location)

(8) deltà del [Ebro] → two entities, **Ebro** (location) and **deltà del Ebro** (location)

(9) château de [Versailles] → two entities, **Versailles** (location) and **château de Versailles** (location)

(10) sindaco di [Venezia] → two entities, **Venezia** (location) and **sindaco di Venezia** (person)

Level 2. refer to a descriptive proper name composed of another descriptive proper name. ENE of Level 2 are built with ENE of level 1 and with a descriptive expansion. The following examples (11 - 13) show some ENE of Level 2. The behaviour is the same as for the previous level, the expansion can change the nature of the object described by the ENE of level 1. For instance, in example (13) the person ENE 'général Charles de Gaulle' becomes a reference to a location with the expansion 'avenue' having a geographical sense.

(11) Parque Natural del [Delta del [Ebro]]

(12) [[Cardiff] City] Football Club

(13) avenue du [général [Charles de Gaulle]]

Level 3. ENE at this level are built with ENE of level 2 plus a descriptive expansion. Actually, there is not really a limit to the overlapping. However, it is really rare to find an ENE of level 3 or more. The following examples (14 - 15) show some ENE of Level 3.

(14) la valle glaciale [del lago di [monte [Acuto]]]

(15) the owner of [the restaurant of [[Neuvic] Lake]]

The proposed hierarchy of overlapping of ENE introduces more detailed entities and allows the annotation of more fine-grain NE and less errors of classification.

In our language each level of the ENE can be marked, from the pure proper name to the whole ENE. For instance, considering example (15), using NER method described in [8] produces the following results:

- 'Neuvic' as **proper name** (ENE level 0)
- 'Neuvic Lake' as **geographical name** (ENE level 1)
- 'restaurant of Neuvic Lake' as **place name** (ENE level 2)
- 'owner of the restaurant of Neuvic Lake' as **person** (ENE level 3)

To illustrate the advantage of using the introduced concept of ENE, we have tested four well-known English NER tools (Stanford, Open Calais, Illinois and FreeLing) with example (15). The results are shown below:

- Stanford Named Entity Recognizer[2] annotates one entity:
  - 'Neuvic Lake' as **location**
- Open Calais[3] annotates two NE:
  - 'Neuvic Lake' as **natural feature**
  - 'restaurant of Neuvic Lake' as **facility**
- NER tool of the Cognitive Computation Group[4] [13] of the University of Illinois annotates only one NE:
  - 'Neuvic Lake' as **organization**
- FreeLing[5] annotates one NE:
  - 'Neuvic Lake' as **geographical location**

We argue that for a powerful task, of marking and of categorizing named entities in context, it is essential to consider ENE (15) as a whole entity. Whereas standard NER tools, considering only the entity 'Neuvic Lake', lead to inaccuracies in categorization and this is also true for OpenCalais in despite of taking into account the word 'restaurant'. As it can be observed in the example (15) the bouding ENE refers to a person (the owner of the restaurant) and not a spatial entity while it is yet the case for all constituent entities. Furthermore, the problem of wrong classification due to bad boundaries detection occurs also for smaller named entities such as for example: *the mayor of Paris*. Among the four NER tools previously tested (i.e., Stanford, Open Calais, Illinois and FreeLing), only one (Open Calais) succeeds to detect the ENE as a person. All the other, consider only the named entity 'Paris' as a location.

2.1.2 TEI Based Annotation

According to our proposal of a TEI compliant XML markup language, we will now describe the TEI elements and their attributes used to represent ENE.

The `<name>` element provided by TEI annotates proper names or noun phrases which are equivalent to ENE of level 0. Table 1 shows the current set of `<name>` attributes defined in our customized specification.

In our specification, all the attributes of the `<name>` element are optional and the `type` attribute may refer to the category of the NE (location, person, organisation, etc.) such as the ENAMEX types proposed in the MUC-6 typology or such as those defined by [15, 16]. The `subType` attribute for `<name>` indicates a second level in the classification such as *roadName* if the `type` attribute value is equal to 'location'. Figure 1 illustrates the annotation of pure proper names (i.e., ENE of level 0) using the TEI `<name>` element.

[2]http://nlp.stanford.edu/software/CRF-NER.shtml
[3]http://new.opencalais.com/
[4]http://cogcomp.cs.illinois.edu/page/demo:view/NER
[5]http://nlp.lsi.upc.edu/freeling/

Table 1: Attributes for `<name>` tag

`type`	category of NE (location, person, organization, etc.)
`subType`	semantic sub-categorization

```
<name xml:id="n1">
  <w type="NPr">Vanoise</w>
</name>
---
<name xml:id="n2" type="location" subtype="roadName">
  <w type="NPr" >GR55</w>
</name>
```

Figure 1: Examples of Named Entities (NE)

In our annotation scheme, we also take advantage of tags provided by TEI for dates and time. These tags are described in the *Core* module of the TEI Guidelines and are available for all TEI documents. They refer to the expressions of date, time or duration in texts. Figure 2 shows some examples of annotation of date and time entities.

```
<date when="2015-07">July 2015</date>
<time when="2015-07-29T20:42:00-05:00">Jul 29 2015
     at 8 pm</time>
<time dur="PT20M">twenty minutes</time>
```

Figure 2: Annotation of date and time

The `<term>` element, defined by the TEI Guidelines, contains a single-word, multi-word, or symbolic designation which is regarded as a technical term. In the current generic layer of the language, we use the `<term>` element for several purposes such as annotating common nouns which refer to the descriptive expansion part of ENE (i.e., common nouns associated with a proper name). Table 2 shows the current set of `<term>` attributes defined in our customized specification.

Table 2: Attributes for `<term>` tag

`type`	*N*, *offset*, *measure*
`subType`	semantic sub-categorization

Table 2 shows the current set of `<term>` attributes defined in our customized specification and Figure 3 shows some examples of annotation of `<term>` elements.

In our proposed customization of TEI, the `type` attribute for the `<term>` element is mandatory and its value must be: *N*, *offset* or *measure*. The *N* value means that the `<term>` element contains a common noun and may refer to the descriptive expansion part of an ENE. Furthermore, in the current version of the language, the *offset* type refers to the expression of spatial or temporal relations. For instance, in the case of an *offset* type, the value of the `subtype` attribute

```
<term xml:id="t1" type="N">
  <w lemma="refuge" type="N">refuge</w>
</term>
---
<term xml:id="t2" type="offset"
     subtype="orientation">
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="nord" type="ADJ">nord</w>
  <w lemma="de" type="PREP">de</w>
</term>
---
<term xml:id="t3" type="offset"
     subtype="direction:final">
  <w lemma="jusque" type="PREP">jusqu</w>
  <w lemma="au" type="PREPDET">au</w>
</term>
---
<term xml:id="t4" xml:lang="en" type="measure">
  <w type="NUM">200</w>
  <w lemma="meter" type="N">meters</w>
</term>
```

Figure 3: <term> elements with different types

may be: *orientation*, *adjacency*, *inclusion*, *direction initial* or *direction final*, depending on the nature of the spatial relations. Finally, the *measure* type annotates distance measures.

As we have seen in the definition of the concept of ENE , there are several levels of expansion. Each level can encapsulate the ENE of lower level. According to the TEI Guidelines, the <rs> element defined in the *Core* module, contains a general purpose name or referring string. In the generic layer specification of our multi-layer markup language, we use the <rs> element for annotating ENE. Furthermore, in our customized specification, a <rs> element is either composed of a <term> element and another <rs> element; or it consists of a <term> element and a <name> element. <rs> elements interpret ENE in a broad manner and can encapsulate all types of <term> elements.

Table 3: Attributes for <rs> tag

type	*expandedName*, *relative*, *sequence*

We specify the <rs> element as having only one attribute: type (Table 3). We define the attribute type as optional and its value must be equal to: *expandedName*, *relative*, or *sequence*. <rs> elements having an *expandedName* type refer to the expression of ENE (see <rs xml:id="rs1"> in Figure 4). They must be composed of a <name> element or another <rs> element and may contain a <term> element with a type value equal to *N*. Furthermore, we use the global attribute n to specify the level of encapsulation (e.g., 0, 1, 2, etc.).

The *relative* type refers to the expression of a proper name (i.e., <name>) or of an ENE of level > 0 (i.e., <rs type="expandedName">), associated with a modifier, i.e. a <term> element having an *offset* value for the type attribute (see <rs xml:id="rs2"> in Figure 4). Finally, the *sequence* value for the type attribute refers to a sequence of several <rs> elements (Figure. 5). A <term> element contained by a <rs type="sequence"> is applied to all the <rs> elements. For instance, the term 'les bourgs' (i.e., small villages) is associated with both *Barioz* and *Bieux* entities (which are the names of small villages).

```
<rs xml:id="rs2" type="relative">
  <term type="offset" subtype="adjacency">
    <w lemma="pres" type="ADV">pres</w>
    <w lemma="du" type="PREPDET">des</w>
  </term>
  <rs xml:id="rs1" n="1" type="expandedName">
    <term type="N">
      <w lemma="chalet" type="N">chalets</w>
    </term>
    <w lemma="de" type="PREP">de</w>
    <w lemma="le" type="DET">la</w>
    <name xml:id="n1">
      <w type="NPr">Gliere</w>
    </name>
  </rs>
</rs>
```

Figure 4: <rs> element: the case of *relative* type

```
<rs type="sequence">
  <term xml:id="t1" type="N">
    <w lemma="le" type="DET">les</w>
    <w lemma="bourg" type="N">bourgs</w>
  </term>
  <w lemma="de" type="PREP">de</w>
  <rs xml:id="rs1" n="0" type="expandedName">
    <name>
      <w type="NPr">Barioz</w>
    </name>
  </rs>
  <w lemma="et" type="CONJC">et</w>
  <w lemma="de" type="PREP">de</w>
  <rs xml:id="rs2" n="0" type="expandedName">
    <name>
      <w type="NPr">Bieux</w>
    </name>
  </rs>
</rs>
```

Figure 5: <rs> element: the case of *sequence* type

2.2 Text segmentation

This section describes the elements that refer to the text segmentation. Structuring of textual information operates on various levels of discourse. They describe the segmentation of a text into traditional linguistic categories such as sentences, words and punctuation marks. These elements are defined by the TEI Guidelines and belong to the *Analysis* module (Simple analytic mechanisms) of TEI.

(16) Êtes-vous parvenu au refuge du Col de la Vanoise?
Did you reach the refuge of Col de la Vanoise?

The <s> element contains a sentence-like division of a text. It may be used to annotate sentences, or any other non-overlapping segments such as complete and non-nesting segmentation of a text. Figure 6 shows the result of the annotation of sentence (16).

```
<s>Êtes-vous parvenu au refuge du Col de la Vanoise?
    </s>
```

Figure 6: Annotation of a sentence

The <w> element represents a grammatical word. The word segmentation depends on which characters are defined as words dividers. Whereas for most of Indo-European languages the word separator is typically a blank space, the

East Asian languages such as Chinese and Japanese differ, i.e., words are not explicitly delimited. In our work we are mainly focused on Indo-European languages.

Table 4: Attributes for <w> tag

lemma	Canonical form of the word
type	Part-of-speech
subType	Semantic sub-categorization

Table 4 shows the current set of <w> attributes defined in our specification and Figure 7 shows the result of the annotation of sentence (16).

```
<w lemma="être" type="V">Êtes</w>
<w lemma="vous" type="PRO">-vous</w>
<phr type="verbal">
  <w lemma="parvenir" type="V"
      subtype="motion:final">parvenu</w>
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="refuge" type="N">refuge</w>
  <w lemma="du" type="PREPDET">du</w>
  <w lemma="col" type="N">Col</w>
  <w lemma="de" type="PREP">de</w>
  <w lemma="le" type="DET">la</w>
  <w lemma="Vanoise" type="NPr">Vanoise</w>
  <pc force="strong" type="interrogative">?</pc>
</phr>
```

Figure 7: Annotation of words, grammatical phrases and punctuation

The lemma attribute for <w> is optional and may contain the canonical form of the word. The type attribute for <w> is mandatory and contains the lexical categories of word (part-of speech). Although there are significant variations depending on the language, common parts of speech are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, numeral, article or determiner. The label for each category is usually defined by abbreviations such as N, V, ADJ, ADV, etc. Different POS tagsets have been defined and used in well-known projects such as the Brown Corpus[6], the Penn Treebank[7] [14] and the French Tree Bank[8] [1]. The subType attribute for <w> is optional and may contain semantic information. For instance, in the current version of the language, subType is used to classify verbs. The possible values are: *motion initial*, *motion median*, *motion final*, *perception* and *topographic*.

The <phr> element defined by the TEI Guidelines in the *analysis* module represents a grammatical phrase. The type attribute may be used to indicate the type of phrase such as noun phrases, prepositional phrases, verbal phrases, etc. Figure 7 shows such annotation on sentence (16).

The <pc> element contains a character or string of characters regarded as constituting a single punctuation mark. Table 5 shows the current set of <pc> attributes.

All the attributes of the <pc> element are optional. The force attribute for <pc> indicates if the considered punctuation mark is a separator for words or phrases. The type attribute for <pc> indicates the kind of punctuation. Figure 7 illustrates the annotation of punctuation.

Table 5: Attributes for <pc> tag

force	*strong*, *weak*, *inter*
type	*declarative*, *imperative*, *interrogative*, *exclamatory*

All these elements used for the text segmentation and the representation of ENE define the generic layer of the multi-layer markup language. They have been applied for the text mining and NER tasks.

3 Towards a Geospatial Semantic ML

3.1 Overview

This section describe the adaptation to transform the generic core layer towards a geospatial semantic markup language: the second layer.

We propose some guidelines for a TEI compliant markup language for encoding spatial information. Some elements belonging to the generic layer of our multi-layer markup language (<term> and <rs>) are turned into more specific elements embedding geospatial semantics.

According to the *Namesdates* module of the TEI Guidelines, the content of <geogFeat> elements is defined as a common noun identifying some geographical feature (e.g., valley, mount, etc.) contained within a spatial NE. This is the equivalent of the definition of a descriptive expansion part of an ENE (i.e., <term type="N"> element) having a geographical denotation and associated with a spatial NE. Thus, in our customized specification, the <term type="N"> elements having a geographical sense (i.e., city, lake, river, etc.) which are used in conjunction with <name> elements are turned into <geogFeat> elements. Figure 8 shows two examples of <geogFeat> elements.

```
<geogFeat>
  <w lemma="lac" type="N">lac</w>
</geogFeat>
<name>
  <w type="NPr">Grattaleu</w>
</name>
---
<geogFeat>
  <w lemma="torrent" type="N">torrent</w>
</geogFeat>
<w lemma="de" type="PREP">de</w>
<w lemma="la" type="DET">la</w>
<name>
  <w type="NPr">Leisse</w>
</name>
```

Figure 8: Annotation of geographical feature names

According to the *Namesdates* module, the <geogName> element identifies a name associated with some geographical feature such as 'River Thames' or 'col de la Vanoise'. Thus, <rs type="expandedName"> elements which refer to geographical names are turned into <geogName> elements. Figure 9 shows an example of annotation of a <geogName> element.

[6]http://www.comp.leeds.ac.uk/ccalas/tagsets/brown.html
[7]http://www.cis.upenn.edu/~treebank/
[8]http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

```
<geogName type="T" subtype="PASS">
  <geogFeat>
    <w lemma="col" type="N">col</w>
  </geogFeat>
  <w lemma="de" type="PREP">de</w>
  <w lemma="le" type="DET">la</w>
  <name>
    <w lemma="Vanoise" type="NPr">Vanoise</w>
  </name>
</geogName>
```

Figure 9: Annotation of geographical names

As we have seen in the definition of the concept of ENE and in the description of the `<rs>` element, we defined several levels of encapsulation. The global `n` attribute may be also used to indicate the level of encapsulation of `<geogName>` elements (in the same way that it was done for the `<rs>` element). Figure 10 shows an example of encapsulation of two `<geogName>` elements.

```
<geogName type="S" subtype="RHSE" n="2">
  <geogFeat>
    <w lemma="refuge" type="N">refuge</w>
  </geogFeat>
  <w lemma="du" type="PREPDET">du</w>
  <geogName type="T" subtype="PASS" n="1">
    <geogFeat>
      <w lemma="col" type="N">Col</w>
    </geogFeat>
    <w lemma="de" type="PREP">de</w>
    <w lemma="le" type="DET">la</w>
    <name>
      <w lemma="Vanoise" type="NPr">Vanoise</w>
    </name>
  </geogName>
</geogName>
```

Figure 10: Encapsulation of `<geogName>` elements

According to our customized specification, the attributes `type` and `subtype` are optional and their values refer to the nature of the geographical feature such as lake, mountain, valley, city, etc. In the current version of the language, we propose to follow the classification introduced for the GeoNames Ontology[9]. The nine categories of feature classes of GeoNames are shown in Table 6. The *code* column lists the nine possible values for the `type` attribute and the *feature code* column shows some examples of feature classes which are used for the `subtype` attribute. And according to our customized specification, these two attributes can be also used for the `<name>` element when it is not included in a `<geogName>` element.

According to the TEI Guidelines, the `<offset>` element marks that part of a relative temporal or spatial expression which indicates the direction of the offset. Thus, the generic `<term type="offset">` elements referring to spatial relations are turned into `<offset>` elements in the geospatial semantic layer of our proposal.

According to our specification of the geospatial semantic layer, the `<offset>` element annotates spatial relations expressed in texts. At the current level of development, we distinguish three categories of spatial relations: topological relations [2], directional relations [3] and distances. Currently we consider two main types of topological relations:

[9]http://www.geonames.org/ontology/

Table 6: GeoNames feature classes

Name	Code	Feature code
Administrative boundaries	A	first-order administrative division (ADM1)...
Area	L	locality (LCTY), park (PRK)...
Hydrographic	H	stream (STM), lake (LK)...
Hypsographic	T	valley (VAL), pass (PASS)...
Populated place	P	populated place (PPL), farm village (PPLF)...
Road / Railroad	R	trail (TRL), street (ST), road (RD)...
Spot	S	school (SCM), resthouse (RHSE)...
Undersea	U	canyon (CNYU), reef (RFU)...
Vegetation	V	forest (FRST), cultivated area (CULT)...

adjacency and inclusion. Furthermore, distance relations are described with the `<measure>` element in the paragraph below.

Table 7: Attributes for `<offset>` tag

`type`	*orientation*, *direction initial*, *direction final*, *meet*, *inside*, ...
`subtype`	sub-categorization

Table 7 shows the current set of `<offset>` attributes defined by our customized specification and Figure 11 shows three examples of annotation of `<offset>` element.

```
<offset type="meet" subtype="near">
  <w lemma="pres" type="ADV">pres</w>
  <w lemma="du" type="PREPDET">des</w>
</offset>
---
<offset type="orientation" subtype="north">
  <w lemma="au" type="PREPDET">au</w>
  <w lemma="nord" type="ADJ">nord</w>
  <w lemma="de" type="PREP">de</w>
</offset>
---
<offset type="direction:final">
  <w lemma="jusque" type="PREP">jusqu</w>
  <w lemma="au" type="PREPDET">au</w>
</offset>
```

Figure 11: Annotation of `<offset>` elements

The `type` attribute is mandatory and its value shall be: *orientation*, *direction initial* or *direction final*, *meet*, *inside*... The value of the optional `subtype` attribute depends on the value of the `type` attribute. For the *orientation* `type`, the value of the `subtype` attribute shall be: *south*, *east*, *northwest*, *above*, *behind*, etc. For the *adjacency* `type`, the value of the `subtype` attribute shall be: *next*, *near*, etc. And for the *inclusion* `type`, the value of the `subtype` attribute shall be: *in*, *inside*, etc.

According to the TEI Guidelines, the `<measure>` element contains a word or phrase referring to some quantity, usually comprising a number, a unit, and a commodity name. Thus, according to our specification of the geospatial semantic layer, the `<term type="measure">` elements referring to distance relations are turned into `<measure>` elements in the

geospatial semantic layer of the language.

type	*distance*
unit	unit identifier
quantity	numeric value

Table 8: Attributes for <measure> tag

Table 8 shows the current set of <measure> attributes defined in our customized specification. All attributes are optional. The unit attribute indicates the units used for the measurement. The value shall be expressed in the International System Units (SI)[10] such as meter and second. The value of the quantity attribute shall be a numeric value. Figure 12 shows an example of annotation of <measure> element.

```
<measure xml:lang="en" type="distance" unit="m"
    quantity="200">
  <w type="NUM">two</w>
  <w type="NUM">hundred</w>
  <w lemma="meter" type="N">meters</w>
</measure>
```

Figure 12: Annotation of <measure> element

The <placeName> element is defined by the TEI Guidelines as containing an absolute or relative place name. Thus, according to our specification of the geospatial semantic layer, the <rs> elements of the generic layer of our proposal referring to geographical places (i.e., containing a <geogName> or <name type="place">) are turned into <placeName> elements. Furthermore, we specify that <geogName> elements must be included into <placeName> elements.

Table 9: Attributes for <placeName> tag

type	*absolute* or *relative*

```
<placeName type="absolute">
  <geogName type="S" subtype="RHSE" n="2">
    <geogFeat>
      <w lemma="refuge" type="N">refuge</w>
    </geogFeat>
    <w lemma="du" type="PREPDET">du</w>
    <geogName type="T" subtype="PASS" n="1">
      <geogFeat>
        <w lemma="col" type="N">Col</w>
      </geogFeat>
      <w lemma="de" type="PREP">de</w>
      <w lemma="le" type="DET">la</w>
      <name>
        <w lemma="Vanoise" type="NPr">Vanoise</w>
      </name>
    </geogName>
  </geogName>
</placeName>
```

Figure 13: Annotation of an absolute <placeName>

We distinguish between two types of <placeName>: absolute and relative and we define the type attribute as optional (Table 9). Absolute <placeName> elements refer to standard spatial ENE and relative <placeName> elements refer to spatial ENE associated with spatial relations (i.e., <offset> and <measure> elements). In other words, the <rs type="expandedName"> elements defined in the generic layer are turned into <placeName type ="absolute"> and <rs type ="relative"> elements are turned into <placeName type="relative">. Figure 13 and Figure 14 show an example of annotation of an absolute and a relative <placeName> element respectively.

```
<placeName type="relative">
  <measure type="distance" unit="m" quantity="200">
    <w type="NUM">200</w>
    <w lemma="mètre" type="N">mètres</w>
  </measure>
  <offset type="orientation" subtype="north">
    <w lemma="au" type="PREPDET">au</w>
    <w lemma="nord" type="ADJ">nord</w>
    <w lemma="de" type="PREP">de</w>
  </offset>
  <name type="place">
    <w type="NPr">Pau</w>
  </name>
</placeName>
```

Figure 14: Annotation of a relative <placeName>

The <place> element is defined by the TEI Guidelines as a generic element containing data about a geographic location. In the current version of our specification of the geospatial semantic layer, we use this element to replace the generic <rs type="sequence"> element. Thus, we consider that the <place> element refers to the definition of a spatial area from the association of several locations (Figure 15). According to our specification, <place> elements can contain the <type> and <subtype> attributes described in the <geogName> element and referring to the feature class of the spatial object.

```
<place type="P" subtype="PPLS">
  <geogFeat>
    <w lemma="le" type="DET">les</w>
    <w lemma="bourg" type="N">bourgs</w>
  </geogFeat>
  <w lemma="de" type="PREP">de</w>
  <placeName xml:id="pn2">
    <name>
      <w type="NPr">Barioz</w>
    </name>
  </placeName>
  <w lemma="et" type="CONJC">et</w>
  <w lemma="de" type="PREP">de</w>
  <placeName xml:id="pn1">
    <name>
      <w type="NPr">Bieux</w>
    </name>
  </placeName>
</place>
```

Figure 15: Annotation of a <place> element

We customize the <phr> element described in the generic layer of our multi-layer markup language for annotating motion events and perception expressions.

According to our specification of the geospatial semantic layer, the *subtype* attribute (optional) indicates the semantic of the verbal phrase, its value shall be: *motion* or *perception*.

[10] http://www.bipm.org/en/publications/si-brochure/

Table 10: Attributes for <phr> tag

type	type of the phrase
subtype	semantic sub-categorization (e.g., *motion* or *perception*)
function	a second level of semantic sub-categorization

The *function* attribute (optional) indicates the motion class.

In the current version of our proposal we consider a set of six motion classes based on the classifications of verbs of motion proposed by [9]: *leave*, *hit*, *reach*, *external*, *internal*, and *cross*. Figure 16 shows the result of the annotation of sentence (16) using the elements available in our geospatial semantic layer.

```
<s>
<w lemma="être" type="V">Êtes</w>
  <w lemma="vous" type="PRO">-vous</w>
  <phr type="verbal" subtype="motion"
       function="reach">
    <w lemma="parvenir" type="V"
         subtype="motion:final">parvenu</w>
    <w lemma="au" type="PREPDET">au</w>
    <placeName type="absolute">
    <geogName type="S" subtype="RHSE" n="2">
      <geogFeat>
        <w lemma="refuge" type="N">refuge</w>
      </geogFeat>
      <w lemma="du" type="PREPDET">du</w>
      <geogName type="T" subtype="PASS" n="1">
        <geogFeat>
          <w lemma="col" type="N">Col</w>
        </geogFeat>
        <w lemma="de" type="PREP">de</w>
        <w lemma="le" type="DET">la</w>
        <name>
          <w lemma="Vanoise" type="NPr">Vanoise</w>
        </name>
      </geogName>
    </geogName>
  </placeName>
  </phr>
  <pc force="strong" type="interrogative">?</pc>
</s>
```

Figure 16: Annotation of a <phr> element

3.2 Encoding Geometric Properties of Spatial Features

The TEI Guidelines describe also some elements for encoding geometric properties of spatial features. According to the *Namesdates* module, the <location> element defines the location of a place as a set of geographical coordinates and the <geo> element contains any expression of a set of geographic coordinates. However, geographic coordinates such as latitude and longitude values are not often available directly in the textual description and must be retrieved from external geographic resources.

Figure 17 shows an example of annotation using the <location> and <geo> elements. The <bloc> and <country> elements (optional) indicates the continent and the country respectively to which the location belongs.

As we have defined the concept of ENE as an encapsulation of several levels of expansion (see Section 2.1), according to our specification the <location> and <geo> elements can be

```
<placeName type="absolute">
  <name>
    <w type="NPr">Pau</w>
    <location>
      <country key="FR" />
      <bloc type="continent" key="EU" />
      <geo>43.301667 -0.368611</geo>
    </location>
  </name>
</placeName>
```

Figure 17: Annotation of the <location> element

nested in various elements depending on the ENE to which it refers. For instance, in Figure 17 the <location> element is nested in the <name> element and refer to the location of the spatial NE 'Pau', whereas for relative ENE (e.g., 'sud de Pau') the <location> element is nested in <placeName> element.

Furthermore, a location may be specified by using a non-TEI XML vocabulary such as GML and KML. Then, we also propose a mapping via gazetteer unique identifiers, i.e. the use of RDF identifiers to interlink with resources on the Web of Linked Data such as Geonames or DBpedia.

3.3 Indication of Uncertainty

The <certainty> element indicates the degree of certainty associated with some aspects of the text markup. This element is described in the *Certainty* module of the TEI Guidelines.

Table 11: Attributes for <certainty> tag

target	URI data pointer
locus	*name*, *start*, *end*, *location*, *value*
assertedValue	alternative value
degree	degree of confidence

Table 11 shows the current set of <certainty> attributes defined in our specification. The target attribute indicates the element to which the certainty is applied using the URI syntax. The target attribute is optional and if it is not expressed, the certainty relies to its parent element. The locus attribute indicates the aspect concerning which certainty is being expressed. The locus attribute is mandatory and its value shall be one of the following: *name*, *start*, *location*, *value*. The *name* value indicates that the uncertainty relies on the name of the element to which the <certainty> element refers. The *start*, *end* and *location* values indicate respectively whether the start, the end or both the start and the end of the element are correctly identified. The *value* value indicates that the uncertainty concerns the content of the element. The assertedValue attribute provides an alternate value for the aspect of the considered markup. For instance, when the value of the locus attribute is equal to *name*, the value of the <assertedValue> refers to the alternate value of the name of the element in question. And finally, the degree attribute indicates the degree of confidence expressed by the <certainty> element.

With respect to the problem of NE classification, the <certainty> element can be used to indicate the degree of certainty of the type assigned to a NE. Figure 18 shows an example of a <certainty> element applied to a <placeName>

element. The *name* value of the `locus` attribute and the *rs* value of the `assertedValue` attribute indicate that the `<placeName>` element may be a `<rs>` element.

```
<placeName xml:id="pl1">
  <certainty target="#pl1" locus="name"
      assertedValue="rs" degree="0.6" />
  <name>
    <w type="NPr">Paris</w>
  </name>
</placeName>
```

Figure 18: Annotation using the `<certainty>` element

Furthermore, the `<certainty>` element can be also used to indicate the certainty degree of the toponym disambiguation task. In this case, the `<certainty>` element must be associated with the `<geo>` element. Figure 19 shows an example in which the uncertainty relies on the geographical location of the spatial entity.

```
<placeName xml:id="pn1" type="absolute">
  <name>
    <w type="NPr">Pau</w>
    <location>
      <geo xml:id="geo1">43.301667 -0.368611</geo>
    </location>
  </name>
</placeName>
<certainty target="#geo1" locus="value" degree="0.8"
    />
```

Figure 19: Annotation using the `<certainty>` element

4 Discussion

The main idea of this paper is to propose a general framework for people to create their own specific markup language based on a core generic layer particularly dedicated for the text mining tasks such as information extraction, data mining, but also for deeper linguistic processing such as semantic parsing and co-reference resolution. The core generic layer is also ready to use to be layered with more semantic relationships between elements of the text.

To our knowledge NER is an important pre-processing step for most of these tasks and automatic NER process for Indo-European languages might be more or less challenging (e.g. for German it is especially challenging). This is the reason why our contribution is based on examples from this task and our proposal relies on the TEI standard which is widely used in digital humanities and linguistics for Indo-European languages. Thus, the objective is to define several specific languages, each one adapted to a specific need and all based on the same generic core layer. The proposed generic core layer may be used to create and share pre-processed corpus.

Table 12 shows a summary of the different elements defined by the generic and the geospatial layers of our multi-layer markup language.

Unlike a great deal of current research, dealing with NER, that are considering only pure proper names or very few entities that we defined as ENE of level 1 such as Eiffel Tower and River Thames, we consider in our proposal both categories of proper names (i.e., pure and descriptive).

Table 12: Summary of the tagset defined for the Generic and the Geospatial layer of our multi-layer markup language

Textual elements	Tagset of the Generic layer	Tagset of the Geospatial layer
word	`<w>`	
sentence	`<s>`	
punctuation	`<pc>`	

spatial and temporal relations
measure expressions
expansion of ENE
	`<term>`	`<offset>`
		`<measure>`
		`<geogFeat>`

NE *(level 0)*	`<name>`	

ENE *(level > 0)*	`<rs>`	`<geogName>`
		`<placeName>`
		`<place>`

verbal phrase	`<phr>`	

Figure 20 shows the result of the annotation of sentence (16) using all the elements and attributes defined in our customized specification of a geospatial semantic language.

```
<s>
  <w lemma="être" type="V">Êtes</w>
  <w lemma="vous" type="PRO">-vous</w>
  <phr xml:id="phr1" type="verb" subtype="motion"
      function="reach">
    <w lemma="parvenir" type="V"
        subtype="motion:final">parvenu</w>
    <w lemma="au" type="PREPDET">au</w>
    <placeName xml:id="pn1" type="absolute">
      <certainty target="#pn1" locus="name"
          assertedValue="rs" degree="1.0"/>
      <location>
        <geo xml:id="geo1">51.969604 -2.893146</geo>
        <country key="FR" />
        <bloc type="continent" key="EU" />
        <certainty target="#geo1" locus="value"
            degree="1.0"/>
      </location>
      <geogName type="S" subtype="RHSE" n="2">
        <geogFeat>
          <w lemma="refuge" type="N">refuge</w>
        </geogFeat>
        <w lemma="du" type="PREPDET">du</w>
        <geogName type="T" subtype="PASS" n="1">
          <geogFeat>
            <w lemma="col" type="N">Col</w>
          </geogFeat>
          <w lemma="de" type="PREP">de</w>
          <w lemma="le" type="DET">la</w>
          <name>
            <w lemma="Vanoise" type="NPr">Vanoise</w>
          </name>
        </geogName>
      </geogName>
    </placeName>
  </phr>
  <pc force="strong" type="interrogative">?</pc>
</s>
```

Figure 20: Example of annotation

We applied our proposal of a multi-layer markup language for the annotation of geospatial information from textual descriptions of itineraries. Then, we use this encoding of geospatial information in order to automatically reconstruct the described itinerary. The automatic reconstruction of itineraries task is based on a calculation (multi-criteria ap-

proach) using all the elements annotated from the textual description. Given the work in progress of our modeling work, the information concerning the reconstructed itinerary is not yet introduced in the annotation. Despite this we have shown how it is possible to define new layers according to the needs and aims. For instance, users can define a third layer derived from the second layer of the multi-layer markup language using non-TEI elements and also non-consuming tags which may interlink already tagged elements such as NE or ENE with verbs or any other kind of element (e.g., spatial or temporal relations, etc.). On our side we are considering in the short term to specify a third layer introducing more semantic and dedicated to the representation of motion using some ISO-Space tags such as the spatial link elements (`<QSLINK>`, `<OLINK>`, `<MOVELINK>`, `<MLINK>`).

5 Acknowledgments

This work has been partially supported by : the Communauté d'Agglomération Pau Pyrénées (CDAPP) and the Institut National de l'Information Géographique et Forestière (IGN) through the PERDIDO project ; the Spanish Government (project TIN2012-37826-C02-01); and the Aragon and Aquitaine Regional Governments through the transborder Aragon-Aquitaine cooperation programme 2014.

References

[1] A. Abeillé, L. Clément, and F. Toussenel. Building a treebank for french. In A. Abeillé, editor, *Treebanks*, number 20 in Text, Speech and Language Technology, pages 165–187. Springer Netherlands, Jan. 2003.

[2] M. Egenhofer and R. Franzosa. Point-set topological spatial relations. *International journal for Geographical Information Systems*, 5(2):161–174, 1991.

[3] A. U. Frank. Qualitative spatial reasoning with cardinal directions. In *Proc. of the Seventh Austrian Conference on Artificial Intelligence*, pages 157–167. Springer, 1991.

[4] D. Gildea and D. Jurafsky. Automatic Labeling of Semantic Roles. *Comput. Linguist.*, 28(3):245–288, Sept. 2002.

[5] K. Jonasson. *Le nom propre*. Duculot, Belgique, Louvain-la-Neuve, 1994.

[6] I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. Spatialml: Annotation scheme, corpora, and tools. In *Proc. of LREC'08, European Language Resources Association (ELRA)*, Marrakech, Morocco, may 2008.

[7] L. Moncla, M. Gaio, and S. Mustière. Automatic itinerary reconstruction from texts. In *Proc. of GI-Science 2014*, Geographic Information Science, pages 253–267, Vienna, Austria, Sept. 2014.

[8] L. Moncla, W. Renteria-Agualimpia, J. Nogueras-Iso, and M. Gaio. Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. In *Proc. of the SIGSPATIAL '14*, pages 183–192, New York, NY, USA, 2014. ACM.

[9] P. Muller. A qualitative theory of motion based on spatio-temporal primitives. In *Proc. of the Sixth International Conference on Knowledge Representation and Reasoning (KR98)*, pages 131–141, 1998.

[10] S. Pradhan, K. Hacioglu, W. Ward, J. H. Martin, and D. Jurafsky. Semantic Role Parsing: Adding Semantic Structure to Unstructured Text. In *Proc. of ICDM '03*, pages 629–632, Washington, DC, USA, 2003. IEEE Computer Society.

[11] R. S. Purves and C. Derungs. From space to place: place-based explorations of texts. *International Journal of Humanities and Arts Computing*, 1(9):74–94, 2015.

[12] J. Pustejovsky, J. L. Moszkowicz, and M. Verhagen. A linguistically grounded annotation language for spatial information. *TAL*, 53(2):87–113, 2012.

[13] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL*, 6 2009.

[14] B. Santorini. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical report, Department of Computer and Information Science, University of Pennsylvania, 1990.

[15] S. Sekine, K. Sudo, and C. Nobata. Extended named entity hierarchy. In *Proc. of LREC'02, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.

[16] M. Tran and D. Maurel. Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *Traitement Automatique des Langues*, 47(3):115–139, 2006.