Integration of temporal and semantic components into the Geographic Information. Part II: Methodology

Willington Siabato, M.Sc., Miguel-Angel Manso-Callejo, Ph.D. Phone: +34.91331-1968, Fax: +34.91331-1968, E Mail: { w.siabato; m.manso}@upm.es, Technical University of Madrid, Autovía de Valencia Km. 7.5, 28031 Madrid, Spain

Bruno Martins, Ph.D., Phone: +351.21-4233508, Fax: +351. 21-4233290, E Mail: b.martins@ist.utl.pt, Instituto Superior Técnico, INESC-ID, Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract

The overall objective of this research project is to enrich geographic data with temporal and semantic components in order to significantly improve spatio-temporal analysis of geographic phenomena. To achieve this goal, we intend to establish and incorporate three new layers (structures) into the core of the Geographic Information by using mark-up languages as well as defining a set of methods and tools for enriching the system to make it able to retrieve and exploit such layers (semantic-temporal, geosemantic, and incremental spatio-temporal). Besides these layers, we also propose a set of models (temporal and spatial) and two semantic engines that make the most of the enriched geographic data. The roots of the project and its definition have been previously presented in Siabato & Manso-Callejo 2011. In this new position paper, we extend such work by delineating clearly the methodology and the foundations on which we will base to define the main components of this research: the *spatial model*, the *temporal model*, the *semantic layers*, and the *semantic engines*. By putting together the former paper and this new work we try to present a comprehensive description of the whole process, from pinpointing the basic problem to describing and assessing the solution. In this new article we just mention the methods and the background to describe how we intend to define the components and integrate them into the GI.

Keywords: T-GIS; Temporal GIS; spatio-temporal analysis; Information Retrieval; GIR; TIR; geosemantics.

1 Introduction

In Siabato & Manso-Callejo 2011, we presented the formal definition of the research project entitled *Integration of temporal and semantic components into the Geographic Information through Mark-up Languages.* Outlining this proposal, the project aims to define a set of methods, rules, and restrictions for the adequate integration of *semantic*, *temporal*, and *spatiotemporal* components into the primary elements of the Geographic Information (theme, location, and time) to improve spatio-temporal analysis of geographic phenomena. Although these primary elements were formally defined more than three decades ago (Sinton 1978), they are still used in different systems as they were originally defined, namely single vector formats, database storage methods, spatial database engines, among others. Although these components certainly work quite well in different scenarios and solve an important amount of spatial problems, they alone seem to be insufficient for achieving tough spatio-temporal analyses. We consider that the lack of the semantic and temporal elements in the current storage structures of Geographic Information (GI) is the main reason for the spatio-temporal analyses to be deficient. Due to this, we state that the proposal of a new storage model for incorporating an independent temporal structure and a set of semantic components (temporal and spatial) would optimise such storage and therefore would improve the retrieving, processing, and analysis capabilities of GI into spatio-temporal scenarios.

Into this context, our proposal defines work oriented to the modelling, storage, and retrieval of dynamic GI taking into account the three above mentioned components. In order to integrate such components into de GI, we intend to establish and incorporate three new layers (structures) into the core of data storage process by using mark-up languages as well as defining a set of methods and tools to exploit the new layers. The ultimate objective is the modelling, querying, and retrieval of dynamic geographic features, establishing the necessary mechanisms to store incremental geometries enriched with a temporal structure and a set of semantic descriptors detailing (i) the nature of the represented phenomena, (ii) their temporality, and (iii) their temporal and meaning (semantic) relations. Figure 1 and Figure 2 show and overview of these concepts.

This research project is primarily based on concepts and studies related to space and time; Temporal GIS (T-GIS); semantic, geosemantics and semantic interoperability; annotation of temporal expressions; Geographic Information Retrieval (GIR) and Temporal Information Retrieval (TIR); Spatio-temporal Databases; and other cross-related topics. We do not include in this paper[1] the state of the art since we have already prepared a completed survey of literature about space and time in GIS (Siabato & Manso-Callejo 2012b) and a preliminary study of the T-GIS foundations is also available in (Siabato & Manso-Callejo 2011). In addition to this, we have prepared and made available online an up-to-date dynamic bibliography about the above mentioned topics and others such as Moving Objects, Information Retrieval, Spatial Databases; even some standards quite related to this research have been listed. Moreover, we are also preparing an infographic study of this bibliography. All these elements and updated information about the evolution of this research project are available in (Siabato & Manso-Callejo 2012a).

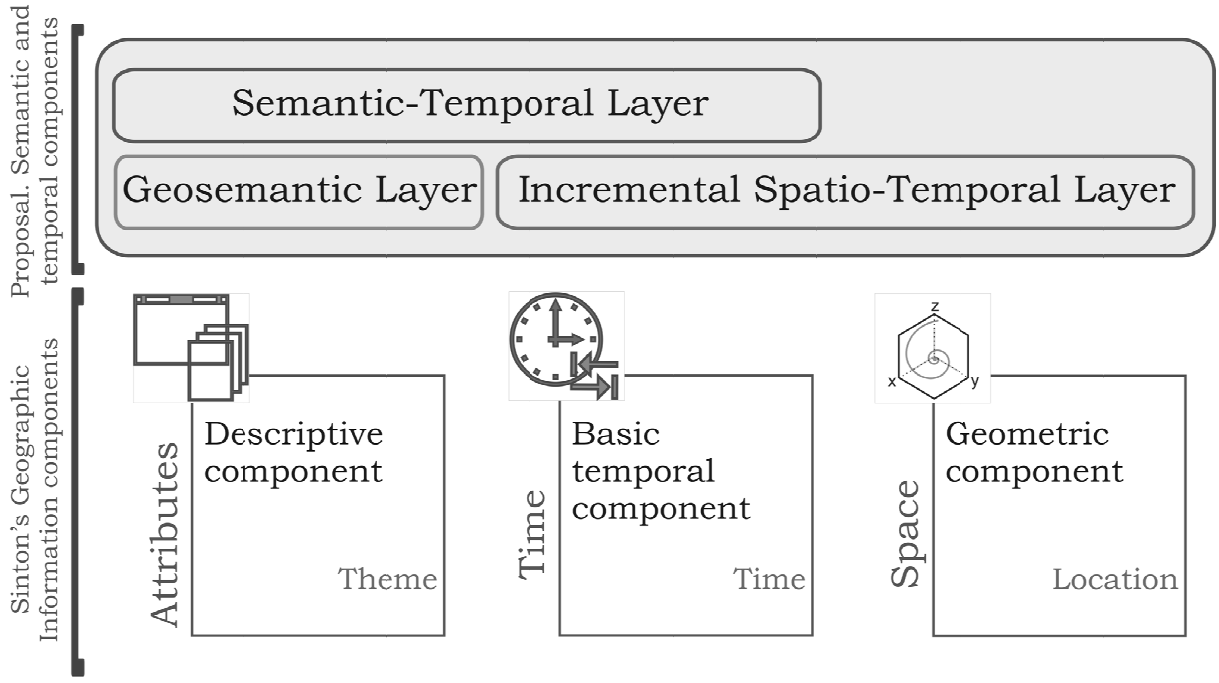

Figure 1. Basic components of the GI and new proposed layers. The three layers relates the three primary components, the point of convergence is the time. A specific geosemantic layer describes nature of features.

[1] The research plan includes the publication of six incremental papers, each one of which will describe a specific part of the project. This is the second of the series. Part III, IV, V, and VI will describe the models, the semantic layers, the semantic engines, and the integration of the proposed model.

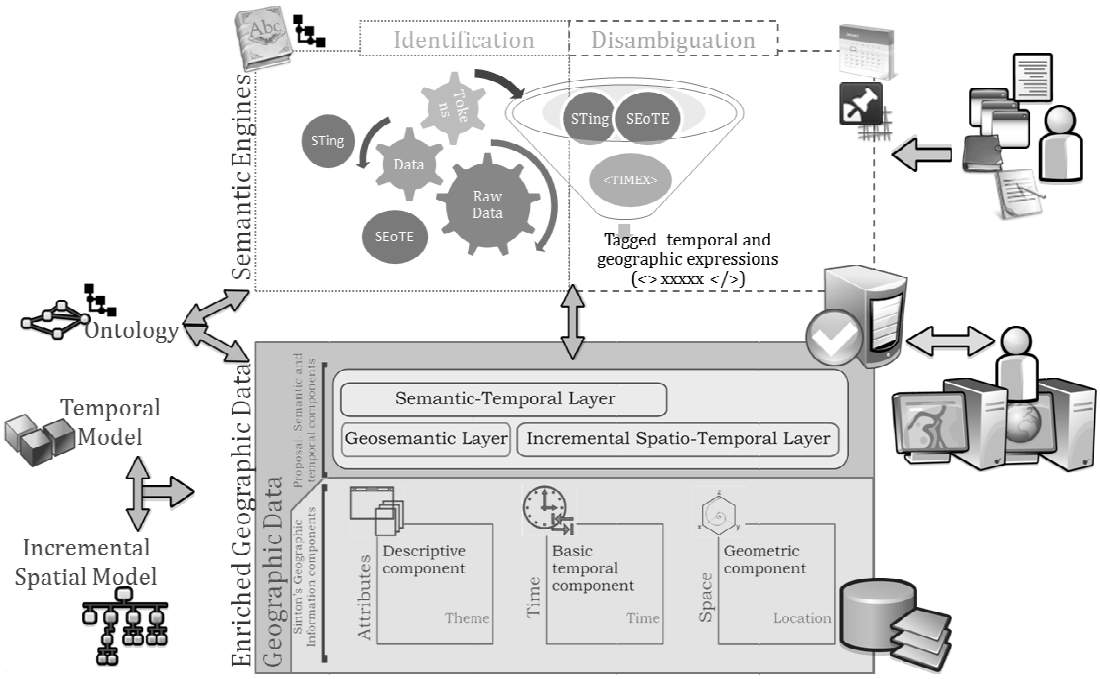

Figure 2. Schematic overview of the proposal as a whole system. In this scheme, we represent the main parts of the proposal (models, layers, and engines) and the flow of the process. The basic geographic data will be enriched through the three layers, which are in turn based on the temporal and spatial models. Once the enriched geographic data is ready, it is necessary to deploy a set of methods for exploiting them, part of these methods are the semantic engines. These engines can interact directly with both the data and the user to solve and identify temporal and geographic expressions. In this sense, user can also provide any kind of textual entrance (e.g. documents, queries, tables) for processing it; this characteristic allows the handling of textual documents through GIS. By putting this together, user can interact with new data structures and make the most of the enriched data through new methods of analysis that will be based mainly on temporal interval algebra.

Thus, in this new article we present part of the foundations and the methodology that we will follow to define and to implement the components and methods that are expected to enrich and to improve geographic information. The following sections describe in detail the methodology for the design and implementation of three components. As we just describe the methodology, we do not go into detail about the design and characteristics of each component. In section 2, we present the *Semantic Engine of Temporal Expressions* (SEoTE) and the *Spatio-Temporal semantic engine* (STing). Then, we briefly introduce the temporal model on which this research will be based, in this section we introduce how the temporal behaviour of geographic features can be modelled through *The Cube of Time* and the *Timing Points*. We mention the spatial incremental storage model. Finally, in section 3, we present some preliminary conclusions and the next steps of this research.

2 Methodology and background

In this section, we present the foundations of the methods that are used for the definition of the semantic engines as well as the core concepts for their integration into our proposal. Although the formal definition of each component is out of the scope of this paper, we intend to show clearly the background and its justification.

Figure 3 shows how, since a theoretical viewpoint, different features of different nature will be able to interact directly with others and establish relations by themselves. These relations will enable data to identify features of same type and

nature (circles) in addition to follow links or relations (intersections) that can be established e.g. by some heuristics or simple rules. This could make the system more flexible since it is not dependent on relational rigid models. Given this, the main component are the semantic taggers, through of which data will know *who they are*, *what they are*, and *their temporal references*, in this sense, task of semantic annotation plays an important role in this research.

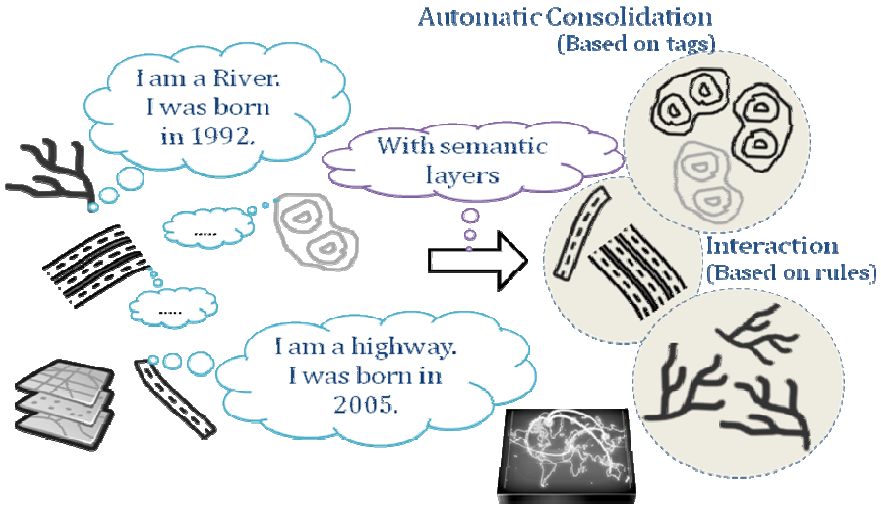

Figure 3. Schematic representation of the integration of data due to the semantic taggers. With the semantic layers, data will be aware of their nature and will be able to create relations by themselves. So, our vision in a very utopian world is being able to see how data can interact in a system just by incorporating them into the system. We want to provide data with the capability of interact without having to be immersed in a static model.

2.1 Semantic components

Semantic is a very general term used in different fields and which can refer to different significances. It is also often too vague. In this project, we refer to the term semantics in its purest definition, i.e. the meaning of the things. In this sense, the semantic components look for defining what the geographic elements and the temporal elements stand for, since a natural language viewpoint. In fact, this approach is quite related to the linguistic semantics, which refers to the study of meaning used to understand human expression through language. The idea of the semantic components and the layers of the system (see Figure 2) is to translate that meaning sense from language to data. Due to this meaning, data could be able to interact as shown in Figure 3.

In this research project, the definition of the semantic components comprises in essence two single tasks: *tagging* and *processing*. The *tagging task* includes the annotation of the data with specific tags, either temporal or spatio-semantic, as well as grant data the possibility of identifying and recognizing semantic expressions onto a text e.g. through a corpora (Corporation 2004, Ferro et al. 2010, Group 2008, Group 2010, Pustejovsky et al. 2003) or by using semantic dictionaries or ontologies (Lutz & Klien 2006). The *processing task* includes the identification and disambiguation of the expressions. In the strict sense, first task is completed when data are created or modified and the second one when data are queried or analysed. In conclusion, we must firstly define the semantic layers to be able to use the semantic engines for searching and analyzing data collections or user inputs (see Figure 2).

For the *tagging task* we will use a set of mark-up languages (e.g. TIMEX2, TimeML, and SpatialML). We will also define some general heuristics for the resolution of specific cases; these heuristics will also improve the accuracy in the identification and disambiguation of the expressions (F-measure).

For the definition of the *processing task* we will use the Natural Language Processing techniques, specifically the Named Entity Recognition (NER) (Nadeau & Sekine 2007). The methods on which we will base come from the Information Retrieval area; we include some specific techniques applied into the Geographic and Temporal Information Retrieval fields. Over the last few years the temporal reference resolution problem has been addressed by many different researchers (Ahn et al. 2007, Mani & Wilson 2000, Miller 2003) as well as the geographic references (Goldberg et al. 2009, Kornai & Sundheim 2003, Markowetz et al. 2005, Martins et al. 2010). The processing task will be conditioned by the semantic engines and their training subtasks which depend on annotated corpora, gazetteers, rules, and ontologies. In this project, we will define and create a specific ontology for temporal expressions. For the spatial entity recognition, there exist huge and very reliable geographic collections such as the GeoNames[2] gazetteer among others, this research project will be supported on this comprehensive projects and collections. This task is initially just related to identification of geographic entities in the natural language queries defined by users, although it can be also extended to general use purposes. In fact, these engines will support the tagging of geographic data as well as permit the handling of textual documents through GIS. Keeping all these basic concepts in mind, in the next sections we describe the semantic engines.

2.1.1 Semantic Engine of Temporal Expressions. *SEoTE*

Temporal expressions by definition have multiple granularities (finer and coarser), models, representations and other different possibilities depending on context and source such as durations and reference points. These expressions are also pervasive since every document or conversation has a high probability of contain references to particular calendar dates, clock times or duration periods (Loureiro et al. 2011). This means that temporal reference identification presents different non-trivial problems due to the inherent ambiguity and contextual assumptions of the natural language discourse, e.g. the expression "*at the beginning of* the Cainozoic era" could means a thousand of years, while the expressions "*at the beginning of* the renaissance" could means 10 years. The definition of these intervals implies to know the period of the events and what the expression *at the beginning of* means.

For the resolution of temporal references, we will follow the method proposed by Loureiro et al. (2011). In their work, authors proposed a supervised machine learning approach. The method is an instance of the well-known stacked learning paradigm proposed by Wolpert (1992). Stacked generalization is a way of combining multiple models, i.e. introduces the concept of a meta-learner or second learner. The procedure comprises four general steps: (i) split the training set into two disjoint sets; (ii) train a significant amount of base learners on the first part; (iii) test the base learners on the second part; and (iv) using these predictions as the inputs and the correct responses as the outputs, train again a higher level learner. In our case, the first learner, based on Conditional Random Fields (CRF), is used to recognize and

[2] http://www.geonames.org/

classify temporal references, and the second learner, based on Support Vector Machine (SVM) regression, is used to rank the possible candidate disambiguation for the temporal references that were initially tagged. In addition to this, we introduce a novel characteristic in this model by using the Information Retrieval techniques based on ontologies. Although the accuracy of Loureiro's method (F_1 = 0.61)[3] can be considered good enough to be applied into our project as is, we expect to improve this metric and increase the identification of the temporal expressions. We will measure such expected improvement by using the same gold-standard collections (Corporation 2004, Ferro et al. 2010, Group 2008, Group 2010, Pustejovsky et al. 2003) and textual documents to establish accurately how this modification affects the cited method. This implies that we will need to train some datasets apart from create the temporal ontology.

2.1.1.1 SEoTE proposal

In order to define what a temporal expression is and how one should be disambiguated, we follow the TIMEX2 standard. In a previous work (Nieto et al. 2010), authors made an initial attempt for the integration of TimeML into GIS. The main idea was, as in our case, to enrich GIS with temporal capabilities. However, one of the conclusions in the ongoing work was that TimeML is a powerful Mark-up language with a large semantic that goes far beyond the identification of temporal expressions, with a finer granularity, and due to this it is too much just for the identification of the expressions. For this reason and based on this conclusion, we have decided to use TIMEX2 instead of the TimeML standard (-ISO- 2007). In fact, one of the foundations of TimeML is TIMEX, although this standard implements TIMEX3.

Here we provide a very brief description of the TIMEX standard. There are six attributes defined for TIMEX2, the values of these attributes express the semantics of a temporal expression. (i) *VAL*, that contains a normalized value for the date; (ii) *MOD*, that captures temporal modifiers using values close to Allen's algebra expressions such as *before, after, less than, more than, equal or less, start, mid, end* or *approx*; (iii) *ANCHOR VAL*, that contains a normalized value for an anchoring date or time; (iv) *ANCHOR DIR*, that captures the relative direction or orientation between the VAL and ANCHOR VAL attributes, as in *within*, *starting*, *ending*, *as of*, *before* or *after* (It is used to express information about when a duration is placed); (v) SET, that identifies expressions denoting sets of times; and (vi) COMMENT, that contains any comment made by the annotator but it is ignored in the processing. The complete description of the TIMEX2 is available in (Ferro et al. 2005). As this document indicates, although one often refers to this standard as "the TIMEX2 standard", the original name stands for the TIDES program in which was originally developed. It was first documented in Ferro et al. (2000, 2001). Although Loureiro et al. only take into account the VAL and SET attributes, ignoring the remaining TIMEX2 information, for the proposed semantic engines we will consider the full set of annotations supported in the TIMEX standard since the identification of the relations would be particularly important.

[3] The F_1 rate is equal to the harmonic mean between precision and recall. Precision is the percentage of correct references identified/disambiguated by the system and recall is the percentage of references present in the test collection that are identified/disambiguated by the system.

In the methodology, the first step is the identification of the temporal expressions. In order to do this it is necessary to implement a supervised method. Nadeau and Sekine (2007) states that the currently dominant technique in NER is the supervised learning, mainly with models such as the Hidden Markov Model (HMM) or Conditional Random Fields (CRF). Supervised learning requires a pre-annotated set of documents for training a classifier capable of identifying entities in text, for doing this training data must be tagged manually[4]. For the SEoTE and the STing engines, the first learner will be based on the CRF method.

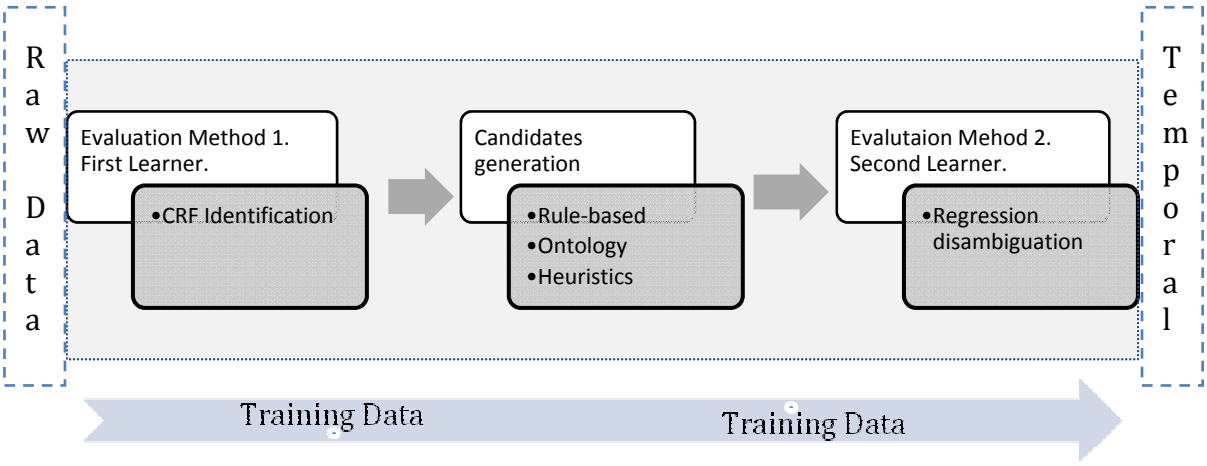

Figure 4. Schematic representation of the semantic temporal engine. The raw data could be any kind of text, e.g. input user text, or attributes of the database. The annotated could be stored in the semantic layers or be used for querying the data. The evaluation method 1 generates the initial temporal references, while method 2 is in charge of identify problems and solve it (disambiguation process).

Lafferty et al. (2001) define CRF as a framework for the construction of probabilistic models to segment and label data given in sequence. Sutton & McCallum (2012) asserts that CRF are essentially a way of combining the advantages of discriminative classification (meaning that it is based on a model of the conditional distribution p(y|o)) and graphical modelling, combining the ability to compactly model multivariate sequential data with the ability to leverage on a large number of input features for prediction. In a sequence tagging problem, such as NER, CRF carries out logistic regression over the possible sequences of tags. The conditional distribution Pλ(y|*o*) is modelled as shown in Equation 1. In other words, CRF is basically a conditional probability associated with an undirected graphical model.

$$\mathrm{P}\lambda(y|o) = \frac{1}{Z\lambda(\boldsymbol{o})} exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(\mathrm{y_j}-1,\mathrm{y_j},\boldsymbol{o},\mathrm{j})\right) \qquad (1)$$

The parameter Zλ is a constant for normalization, and it is defined as shown in equation 2.

$$\mathrm{Z}\lambda(\boldsymbol{o}) = \sum y \in Y\, exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(\mathrm{y_j}-1,\mathrm{y_j},\boldsymbol{o},\mathrm{j})\right) \qquad (2)$$

[4] There are also semi-supervised approaches, where only a small degree of supervision is needed to start the training process, as well as unsupervised learning approaches, which use techniques that rely on lexical resources, patterns and statistics to recognize the entities. (Nadeau & Sekine 2007)

With this model, the identification of temporal references starts with the procedure indicated in Figure 4. By using the BIO (*Begin*, *Inside*, and *Other*) encoding for tagging, we will identify seven types of references (classes) as in (Loureiro et al. 2011): recurrences, durations, points, ambiguous past references, ambiguous present references, ambiguous future references, miscellaneous. This implies to separate the BIO tags for each of the seven references. The classification model tags words given in a sequence according to the above tags, from which the system then generates the final results for the temporal reference recognition step.

The recognition process starts with a tokenization of the sequence to be analyzed. We use a scheme that deterministically breaks an input text into a sequence of independent tokens. This collection of tokens corresponds to the (*o*) observation sequence. What we are looking for, it is probability of classifying each symbol (o) (tokens) with a class *y* among the set of classes we have defined (Y). Tagging these tokens is a non-trivial classification problem, since if there are 7 different tags, then a sequence of length N has up to 7^N possible sequences. To get through this problem, we assume the CRF probabilistic approach.

Once identified the temporal expression, next step is disambiguation. For doing this, the methodology includes three steps: (i) the proposal of a set of disambiguation candidates by using a generative approach based on rules and the ontology of temporal terms, (ii) scoring possible candidates using a Scalable Vector Machine (SVM)[5] regression model, and (iii) selecting the best candidate. The SVM regression is based on the overlap between the true temporal period associated with the expression and the temporal periods of the candidates. This way, the selected candidate will be closest to the real period. For the definition of the candidates, the identification is based on (i) pattern comparison and (ii) ontology comparison. Once the candidates have been identified the highest probability will be function of the both comparisons. Patterns define rules that may refer to a small lexicon of names, units, and numeric words, these patterns could be use both in the text and the ontology. It is necessary also to define a tagger element for the VAL and SET TIMEX2 elements. In general terms, the process involves evaluating the rule`s pattern against the text of the temporal expression and, in case of a successful match, executing the tagger element to identify the candidate already disambiguated. As usual in TIR, deictic and anaphoric terms imply a higher level of complexity and additional treatments must be included as identified in (Nieto et al. 2010). In order to compare expressions properly it is necessary a temporal reference system, we will use the Gregorian calendar and ISO 8626 standard (this will be presented in Temporal component, see Section 2.2 below).

A temporal reference will be correctly identified only if it exist an exact match with the corresponding temporal reference given (*o*) in the test collections (corpus and ontologies), and correctly disambiguated only if the assigned disambiguation is an exact match with the one that is given in the TIMEX2 annotation (i.e., the VAL and SET attributes have the same values). Thus, correct disambiguation can only occur when a correct recognition has first taken place. The final step of the method involves selecting the candidate with the largest estimated overlap as the result for the disambiguation. Loureiro et al. (2011) asserted that the achieved results of the

[5] Support Vector Machine is a set of machine learning approaches used for classification and regression. This method was developed in the mid 90`s by Vapnik (2000) at the AT&T Bell Labs.

stacked proposed method are of an acceptable quality to be used in the subsequent processing stages of many different types of Temporal Information Retrieval applications, in our case, the construction of SEoTE and its integration into a GIS. We expect to improve the accuracy in our semantic engine by adding the ontology-based recognition layer and enriching the rules, pattern and heuristics. Even, it could be valid to modify the standard annotated corpus to increase the recognition of valid expressions that have been wrong annotated in the gold-standard collections.

2.1.2 Spatio-temporal semantic engine. *STing*

By using the same techniques is possible to identify geographic references, and places with a temporal scope. As in the temporal expression recognition process, to identify geographical references it is not a trivial task and it could be even considered a harder challenge, e.g. some lat-long coordinates refer to a single location but it can be recognized as different geographic references across the centuries: Byzantium (667 BC), Constantinople (330 AD), Istanbul (1930). One city or another will be right depending on the temporal context of the data. As this, there are multiple scenarios that require strong support and comprehensive knowledge-repositories such as gazetteers. The ambiguity of geographic references is high, e.g. place names often have other non geographic meanings, different places can be referred to by the same name, and the same places are often referred to by different names (Martins et al. 2010). In general, the concepts we will use for the geographic entity recognition come from the Geographic Information Retrieval field. The approach followed for the definition of the spatio-temporal semantic engine is also identification/disambiguation. The main aim of the STing engine is to identify geographic references included in user query expressions. Nonetheless, the engine could be classified as a general purpose semantic application.

System accuracy for the identification of geographic references is higher than in the TIR systems. F_1 scores around the 90% were already reported by Tjong-Kim-Sang and Meulder (2003). In the last years techniques have evolved in process speed although accuracy has not improved too much. Although Initial approaches for the identification of entities (NER) were based on manually constructed patterns and/or dictionary lists of entity names, the current trend in NER is to use machine-learning approaches, relying mainly on features extracted from training data. Machine learning approaches are more attractive in that they are trainable and adaptable (Martins et al. 2010). Different techniques such as Support Vector Machines, Maximum Entropy, Hidden Markov Models (HMM), and the already mentioned CRF, mainly support GIR. The expected geographic semantic engine will be based on a stacked method with two learners CRF and SVM, just as in the temporal engine (see Figure 4).

While the identification of the entity can be made to rely entirely on internal features of the documents, place reference disambiguation requires always external knowledge in the form of a dictionary (e.g. a gazetteer) for translating place names into geospatial footprints (Siabato et al. 2008). In our case, we will use the GeoNames gazetteer service (Wick 2011), which includes over 8 million geographic references. An important characteristic of this service is that entities evolve through time, as we exemplified above. In an early stage of this research, we were thinking to use a personal gazetteer in which we included GeoNames and other gazetteers

services (Manguinhas et al. 2009). Nonetheless, this proposal is out-of-date and GeoNames service has grown and evolved significantly making it by far a better option for the disambiguation of geographic references. In his doctoral dissertation, Leidner (2007) surveyed different approaches for identifying place references on textual documents. He concluded that most methods usually rely on gazetteer matching for performing the identification, together with a set of heuristics.

Some remarkable examples of GIR systems are the MetaCarta RSS GeoTagger[6] and the Yahoo! Placemaker[7]. Martins et al. (2010) presented a comparison of these commercial services against a proposed machine learning approach for the recognition and disambiguation of place references based on HMM. Experiments with labelled datasets in three different languages showed that the machine learned method outperforms the two commercial state-of-the-art systems.

2.1.2.1 STing Proposal

The approach that we will use in this engine is the same that we presented for the temporal one: stacked learning based on CRF and SVM. In this case, the (SVM) regression to rank a set of possible disambiguation candidates relies on the gazetteer service.

For recognition task, once again the first step is the tokenization of the text. The candidates (recognized tokens) must be tagged following general structures such as the BIO encoding or BMEWO+[8] encoding. For the geographic entities we will check which encoding system is more appropriate. The recognition problem must include a method for solving the non-geo/geo and geo/geo problems.

The disambiguation process involves (i) the identification of candidates by querying the gazetteer, (ii) the classification of possible candidates using the SVM regression's scores, and (iii) the identification of the highest scoring candidate. Following Martins et al. (2010) methodology, the scoring criteria used in the second step are based on estimating the distance between the true geospatial footprints and the geospatial footprints of the candidates by using algorithms such as the convex hull and the Levenshtein distance. The selected candidate will be the top scoring, which corresponds to the one with the least estimated distance in terms of geographic reference approximation. Due to try to identify geographic references by exact location is not possible at all, mainly for the precision of the different datasets, the disambiguation process will take into account a radio of tolerance, the value of this radio must be determined by comparing precision of the involved datasets.

The datasets for training data will be the ones defined in the Conference on Computational Natural Language Learning -CoNLL- (Tjong-Kim-Sang & Meulder 2003). Both editions of the CoNLL NER evaluation, the organizers provided a set of data files for each language (Spanish in the 2002 edition and English in 2003).

---

[6] http://labs.metacarta.com/rss-geotagger/

[7] http://developer.yahoo.com/geo/placemaker/

[8] LingPipe's BMEWO+ encoding distinguishes the following set of tokens: begin-of-entity (B); mid-entity (M); end-of-entity (E); non-entity (O) and single-token entities (W). There are also sub-classifications and constraint rules.

2.1.3 Implementation

These engines are expected to be implemented through the CRF and SVM algorithms defined in LingPipe[9] and the Weka Machine Learning toolkit[10]. These libraries are robust and portable enough to implement the engines as though software pieces to be included into the GIS software.

2.2 Temporal component

The temporal component is the most important part of this research due to the whole developed concepts fall on it. In this section we introduced very briefly the temporal model that we intend to define and to establish.

2.2.1 Temporal model. *The cube of Time*.

Here we present some basic and initial elements about what we have called *The Cube of Time*. This concept differs from the space-time prism defined by Hägerstraand (1970) and Lenntorp (1976) and which is commonly used for spatial analysis (e.g.Miller 1991). Unlike the space-time prism which is composed by two spatial axes and one temporal (Z), the cube of time (C*t*) comprises three temporal axes: database recording time, object changing time, and object creation time (see Figure 5). The point defined by the three temporal coordinates is called the *Timing Point*. A sequence of *timing points* describes the evolution of the object through time. This sequence, which can be described mathematically, shows the evolution of an object through time by following a linear function (we will define if the linear function is good enough or it will require a different type of function). A set of functions permit to establish relations between objects and to identify for instance how a single change in a specific object can affect another or others. The aim of this cube, and therefore of the model, is to provide mathematical definitions that describe the evolution of single objects through time and permit to analyze and identify how they are correlated.

Even though Miller (2005) has established some mathematical definitions for the space-time path and prism, space-time lifelines, bundles, intersections, and other components of the space-time prism framework, and the definitions we will propose for our model are quite linked to Miller's proposal since both of them relate spatial objects in the background; the prism and the cube are quite different in concept mainly because our functions express purely temporal relations and not space-time relations.

In the late nineties and based on the concept of *bitemporal element* (BTE) (Snodgrass 1992), Worboys (1998) defined the spatio-bitemporal model for geographic information. In this model he considered *event-time* and *database-time* along two orthogonal axes. He defined the *ST-simplex* primitives (0-simplex, 1-simplex, 2-simplex) from which *ST-complexes* and *ST-objects* were derived. Every simplex can be combined with a bi-temporal element to form an ordered pair space-time, this pair defines geographic objects with both spatial and bi-temporal extents.

[9] http://alias-i.com/lingpipe/
[10] http://www.cs.waikato.ac.nz/ml/weka/

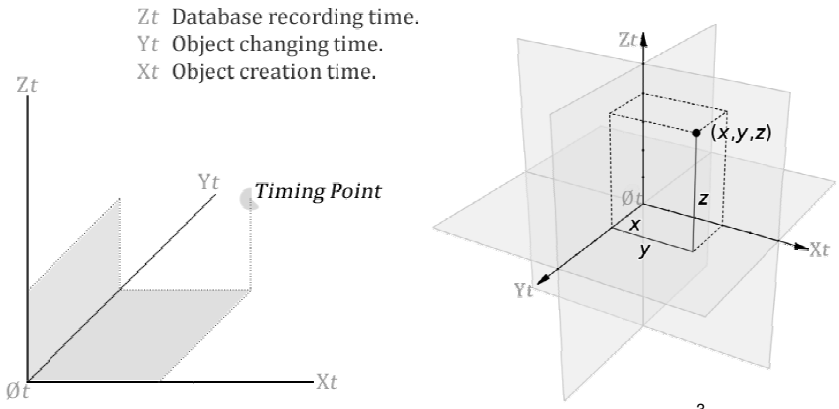

Figure 5. Three temporal Axes of the Time Cube. The cube is described in a R^3 Euclidean space (E^3). Each coordinate of the *Timing Point* is related to a single object's state. The origin θ*t* can be positioned in any range of R. Although the functions that represent the evolution of the objects can be circumscribed in any octant, due to any *database recording time* would never be less than the first recorded value θ*t*(z), the functions will be always over the plane Z*t*=0, and therefore, into the four upper octants.

This is a significant difference between our proposal and Worboy's model due to C*t* is an abstract composite for establishing temporal relations and do not include neither spatial information nor elements. In fact, it is not possible to state comparisons between these approaches. Nonetheless, it is necessary to bear in mind the bi-temporal model because it contains two of the three axes we have defined in our cube. Although it seems quite logic to use only two axes and to omit the *object creation time* axis mainly because this coordinate could be considered as the first event of change, hence not to think the temporal composite as a cube but as two orthogonal axes just as Worboys (and Snodgrass) defined, we will show how by defining the three times (axes) independently and by treating the *Timing Points* in different and parallel planes will be possible to model temporal relations in a different way and to make temporal analysis easier, richer and more flexible. The main objective is to enrich temporal descriptions and temporal relations of the objects (features). On the other hand, this requires more complex definitions due to vectors that describe the behaviour of the object are defined in R^3, but this added complexity level is just for the model definition and do not affect neither implementation nor performance.

As we stated in the definition of this research project (Siabato & Manso-Callejo 2011), how to perceive and model time is a mainstay. In this study, the mechanical and mathematical models will be considered due to through them all geographic features perceived by our senses (rivers, ways) or abstractly modelled (airways) may be ordered and described from a temporal viewpoint. Moreover, it is possible to define elements such as measure, interval, dimension, modification, instant, and position among others. Particularly the applied perspective will be the Newtonian, where time is envisioned as an independent dimension though widely related and similar to the spatial dimension. This approach is used in view of the flexibility it provides for independent handling of the temporal component allowing that space and time can be measured independently (relativist perspective (Ott & Swiaczny 2001)).[11] In general, (i) the three axes that define the cube are mechanical-temporal axes; (ii) the three planes defined by the axes are bi-temporal planes; (iii) the origin of the cube (θ*t*) is determined by the first record in the system (e.g. Spatial database), what implies that the cube in the very first record will be a square (XZ bi-

[11] A complete discussion of types of time in GIS was presented by Frank 1998.

temporal plane). Finally, the time registered in the axes follows the Gregorian calendar for the identification of calendar days and the ISO representation of dates and times (-ISO- 2004). We will evaluate to adapt Coordinated Universal Time for the Timing Points values, this will permit evaluate and operate them easier.[12]

2.3 Spatial component

An initial description of the model for the spatial component is available in (Siabato & Manso-Callejo 2011). The proposed incremental spatial storage model is also based on mathematical foundations that determine *born tasks*, *evolutional tasks* and *dying tasks*. These three set of tasks are also defined in the temporal model. In this sense, temporal evolution and spatial evolution will be modelled independently but they are quite related as in any spatio-temporal model. So, we will integrate the whole concepts in a metamodel, that's the main reason we assumed the mechanical model described above. As presented in (Siabato & Manso-Callejo 2011, p.404), the spatial model is based on the +/-δ*t* operator. We do not extend this concept due to the fact that it has been described quite enough in such work.

2.3.1 Spatio-temporal analysis

Finally, to provide the system with analysis capabilities we rely on Allen's interval algebra (Allen 1981, Allen 1983). For doing this, we will develop an xQuery library to exploit the model, the layers and the proposed storage data model. This library could be considering as third component over and above the semantic engines.

3 Preliminary conclusion

Although this paper just presents the formal methodology that we will follow up into our research, it is possible to state some conclusions. Bearing in mind that the overall goal of this research is to significantly enhance geographic data (i.e. making it more intelligent and flexible) in order to improve current spatio-temporal analysis methods, and to provide GIS with semantic tools that allow system interaction following the flow User←System←→Data; in this paper we have briefly described the methodology and foundations on which we will relay to achieve such objective.

After analyze different proposals and literature reviews, it is clear that for the semantic engines (SEoTE and STing) the Machine Learning approach turns in the best option for the identification and disambiguation of temporal expressions and geographic references. We will follow the stacked method by taking into account two learners based on the Conditional Random Fields (CRF) and Support Vector Machine (SVM) regression. We decided to use these learners due to in the reported works this approach has achieved the highest quality score F_1. Furthermore, these techniques are well-known and there exist a set of algorithms and robust software libraries on which we will stand to create the engines. These libraries guaranty a tough system as well as its portability and scalability. Although the aim of our research is not to create new methods for identifying or disambiguating the expressions, we expect to improve the F_1 score by introducing a new ontology-based

[12] The formal definition of the temporal model will be submitted for peer-reviewed to the ICCSA 2012 Conference. This model corresponds to third paper of this series.

layer. About this, it is not possible to conclude something until implement the prototype and measure the resolution of references (temporal and geographic) using the same gold-standard document collections and conditions under were tested the proposals on view. We intend to create general purpose semantic engines, in this sense, their integration into the GIS will be just a single use case through which such engines will be validated in a geographic context and system.

Regarding this matter in the case of the STing engine, and due to once included in the GIS software the geographic context will be already defined in somehow, we will have to deal with this specific characteristic and, even more importantly, to use it for the disambiguation process. In fact, this should be considered as the main source for this process since the system supposes to have in most cases trust data, and these data will define free of error the geographic scope of the system in that scenario.

Although Louveriro et al. concluded that for addressing more advanced temporal processing a next step in their proposal could be the use of a more powerful mark-up language as TimeML; based on our previous results we strongly believe that simple annotation is more practical and useful for the integration of GIR and TIR methods in GIS and other systems. Due to this, we will use simple TIMEX2 annotations. Probably, TIMEX3 would be considered if the spatial and temporal integration does not report the expected results.

The described temporal model plays a fundamental role in the research. We think that representing the dynamic behaviour of geographic data through parallel planes, on which each feature has “drawn” its evolution, will permit us to define a comprehensive set of operations to analyze temporal data in a completely novel way. Moreover, once defined the foundations of *The Cube of Time*, it will possible to create scenarios for n-dimensional analysis by creating new mathematical definitions for a Universal cube (U-C*t*) in which the basic cubes are contained into a Universal set. This will permit to analyse how changes in one system impacts or changes the behaviour of another. Nonetheless, much more have to be done about this proposal to be considered as a general model, once again, just until implementation tasks have been completed the real impact and reach will be analyzed.

By putting together the former position paper and this new work, we have tried to describe the whole process, from pinpointing the basic problem to describing and assessing the solution. In the Timebliography web site (Siabato & Manso-Callejo 2012a), there are available over 150 references in the sections *Semantic, IR and GIR* and *Annotating Time* related to the development of the semantic engines. There are 50+ references related to Allen’s interval algebra and time intervals in the section *AI and Logic*. Ongoing work and next steps couldn´t be others than implementation of the engines and definition of the mathematical foundations of the models.

Acknowledgments

This work was partially supported by the Doctoral Program of the Technical University of Madrid -UPM- (Grant ref.CH/056/2008) and by the UPM Training and Mobility of Researchers Programme (Resolution 23/02/2011). The work is also supported through SInteliGIS, a research project at INESC-ID financed by the Funcação para a Ciência e Técnologia (grant PTDC/EIA-EIA/109840/2009). Willington Siabato is on leave from UPM and he is currently a visiting researcher at INESC-ID, a research lab associated to the Technical University of Lisbon.

References

Ahn, D., Rantwijk, J. v. & Rijke, M. d. 2007, 'A Cascaded Machine Learning Approach to Interpreting Temporal Expressions', in *HLT 2007: The Conference of the North American Chapter*, ACL pp. 420-427.

Allen, J. F. 1981, 'An interval-based representation of temporal knowledge', in *7th international joint conference on Artificial intelligence*, ed. P. J. Hayes, William Kaufmann, San Francisco-CA-USA, pp. 221-226.

Allen, J. F. 1983, 'Maintaining knowledge about temporal intervals', *Communications of the ACM*, vol. 26, no. 11, pp. 832-843.

Corporation, T. M. 2004, 'Time Expression Recognition and Normalization Evaluation', in *TERN-2004 Evaluation Workshop*, The MITRE Corporation, Bedford-VA-USA.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. & Wilson, G. 2005, *Standard for the Annotation of Temporal Expressions -TIDES-*, The MITRE Corporation, McLean-VG-USA.

Ferro, L., Gerber, L., Mani, I., Sundheim, B. & Wilson, G. (2010/10/15), *ACE Time Normalization (TERN) 2004 English Evaluation Data V1.0*, [Online], Linguistic Data Consortium, Available from: <http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T18> [05-02-2012].

Ferro, L., Mani, I., Sundheim, B. & Wilson, G. 2000, *TIDES Temporal Annotation Guidelines – Draft Version 1.0*, The MITRE Corporation, McLean-VG-USA.

Ferro, L., Mani, I., Sundheim, B. & Wilson, G. 2001, *TIDES Temporal Annotation Guidelines Version 1.0.2*, The MITRE Corporation, McLean-VG-USA.

Frank, A. U. 1998, 'Different Types of "Times" in GIS', in *Spatial and Temporal Reasoning in Geographic Information Systems*, eds. M. J. Egenhofer & R. G. Golledge, Oxford Press, New York-USA, pp. 40-62.

Goldberg, D. W., Wilson, J. P. & Knoblock, C. A. 2009, 'Extracting geographic features from the Internet to automatically build detailed regional gazetteers', *International Journal of Geographical Information Science*, vol. 23, no. 1, pp. 93-128.

Group, T. R. (2010/12/22), *Advanced Question Answering for Intelligence -AQUAINT-*, [Online], NIST, Available from: <http://www-nlpir.nist.gov/projects/aquaint/> [05/02/2012].

Group, T. W. (2008/01/15), *TimeML Corpora*, [Online], Available from: <http://www.timeml.org/site/timebank/timebank.html> [04/02/2012].

Hägerstraand, T. 1970, 'What about People in Regional Science?', *Papers in Regional Science*, vol. 24, no. 1, pp. 7-24.

-ISO-, I. O. 2004, *ISO 8601:2004 - Data elements and interchange formats -- Information interchange -- Representation of dates and times*, International Organization for Standardization -ISO-, Geneva - Switzerland.

-ISO-, I. O. 2007, *Language resource management – Semantic Annotation Framework (SemAF) – Part1: Time and events*, International Organization for Standardization -ISO-, Geneva - Switzerland.

Kornai, A. & Sundheim, B. 2003, 'Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References', in *HLT-NAACL 2003*, Association for Computational Linguistics, Morristown-NJ-USA.

Lafferty, J. D., McCallum, A. & Pereira, F. C. 2001, 'Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data', in *Eighteenth International Conference on Machine Learning -ICML`01-*, eds. C. E. Brodley & A. P. Danyluk, Morgan Kaufmann Publishers, San Francisco-CA-USA, pp. 282-289.

Leidner, J. L. 2007, *Toponym Resolution in Text. Annotation, Evaluation and Applications of Spatial Grounding of Place Names*, Doctoral Dissertationthesis, University of Edinburgh.

Lenntorp, B. 1976, *Paths in space-time environments: a time-geographic study of movement possibilities of individuals*, Royal University of Lund, Lund-Sweden.

Loureiro, V., Calado, P. & Martins, B. 2011, 'A Machine Learning Method for Resolving Temporal References in Text', in *EPIA2011 - 15th Portuguese Conference on Artificial Intelligence*, Associação Portuguesa Para a Inteligência Artificial, Braga-Portugal, pp. 745-759.

Lutz, M. & Klien, E. 2006, 'Ontology-based retrieval of geographic information', *International Journal of Geographical Information Science*, vol. 20, no. 3, pp. 233-260.

Manguinhas, H., Martins, B., Borbinha, J. & Siabato, W. 2009, 'The DIGMAP Geo-Temporal Web Gazetteer Service', *e-Perimetron. International web journal on sciences and technologies affined to history of cartography and maps*, vol. 4, no. 1, pp. 9-24.

Mani, I. & Wilson, D. G. 2000, 'Robust temporal processing of news', in *38th Annual Meeting on Association for Computational Linguistics*, ed. H. Iida, Association for Computational Linguistics, Stroudsburg-PA-USA, pp. 69-76.

Markowetz, A., Brinkhoff, T. & Seeger, B. 2005, 'Geographic information retrieval', in *Next generation geospatial information: from digital image analysis to spatio temporal databases*, eds. P. Agouris & A. Croitoru, Taylor & Francis, London-UK, pp. 5-17.

Martins, B., Anastácio, I. & Calado, P. 2010, 'A Machine Learning Approach for Resolving Place References in Text', in *Geospatial Thinking*, eds. M. Painho, M. Y. Santos & H. Pundt, Springer-Verlag, Berlin - Germany, pp. 221-236.

Miller, H. J. 1991, 'Modelling accessibility using space-time prism concepts within geographical information systems', *International Journal of Geographical Information Systems*, vol. 5, no. 3, pp. 287-301.

Miller, H. J. 2003, 'What about people in geographic information science?', *Computers, Environment and Urban Systems*, vol. 27, no. 5, pp. 447-453.

Miller, H. J. 2005, 'A Measurement Theory for Time Geography', *Geographical Analysis*, vol.37, no.1, pp.17-45.

Nadeau, D. & Sekine, S. 2007, 'A survey of named entity recognition and classification', *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3-26.

Nieto, M. G., Rodriguez, M. J., Zambrana, A. U., Siabato, W. & Bernabé, M. Á. 2010, 'Incorporating TimeML into a GIS', *International Journal of Computational Linguistics and Applications -IJCLA-*, vol. 1, no. 1-2, pp. 269-283.

Ott, T. & Swiaczny, F. 2001, *Time-integrative Geographic Information Systems - Management and Analysis of Spatio-Temporal Data*, Springer-Verlag, Berlin - Germany.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R. J., Setzer, A., Radev, D. R., Sundheim, B., Day, D., Ferro, L. & Lazo, M. 2003, 'The TIMEBANK corpus', in *Corpus Linguistics 2003 Conference*, eds. D. Archer, P. Rayson, A. Wilson & T. McEnery, UCREL, Lancaster-UK, pp. 647-656.

Siabato, W. & Manso-Callejo, M. Á. 2011, 'Integration of temporal and semantic components into the Geographic Information through mark-up languages. Part I: definition', in *Computational Science and Its Applications - ICCSA 2011*, eds. B. Murgante, O. Gervasi, A. Iglesias, D. Taniar & B. O. Apduhan, Springer-Verlag, Berlin - Germany, pp. 394-409.

Siabato, W. & Manso-Callejo, M. Á. (2012a/01/07), *Timebliography. Bibliography for spatio-temporal trends in Temporal GIS -TGIS-*, [Online], UPM, Available from: <http://spaceandtime.wsiabato.info> [07/01/12].

Siabato, W. & Manso-Callejo, M. Á. 2012b, 'Timebliography: a Survey of Literature about Space and Time in GIS', *Geography Compass*, vol. 0, no. 0, p. 0-0.

Siabato, W., Fernández-Wyttenbach, A., Martins, B., Bernabé, M. Á. & Álvarez, M. 2008, 'Análisis semántico del lenguaje natural para expresiones geotemporales', in *Jornadas Técnicas de la IDE de España - JIDEE 2008-*, Cartográfica de Canarias S.A., Tenerife - España.

Sinton, D. F. 1978, 'The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study', in *First International Advanced Study Symposium on topological data structures for Geographic Information Systems*, ed. G. Dutton, Harvard University, Cambridge-MA-USA, pp. 1-17.

Snodgrass, R. T. 1992, 'Temporal databases', in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, eds. A. U. Frank, I. Campari & U. Formentini, Springer-Verlag, Berlin - Germany, pp. 22-64.

Sutton, C. & McCallum, A. 2012, 'An Introduction to Conditional Random Fields for Relational Learning', in *Foundations and Trends in Machine Learning*.

Tjong-Kim-Sang, E. F. & Meulder, F. D. 2003, 'Introduction to the CoNLL-2003 shared task: language-independent named entity recognition', in *Seventh Conference on Natural Language Learning -CoNLL-2003-*, eds. W. Daelemans & M. Osborne, ACL, Morristown-NJ-USA, pp. 142-147.

Vapnik, V. N. 2000, *The Nature of Statistical Learning Theory*, Springer, New York-NY-USA.

Wick, M. (2011/09/12), *GeoNames*, [Online], Available from: <http://www.geonames.org/> [05/02/2012].

Wolpert, D. H. 1992, 'Stacked generalization', *Neural Networks*, vol. 5, no. 2, pp. 241-259.

Worboys, M. F. 1998, 'A generic model for spatio-bitemporal geographic information', in *Spatial and Temporal Reasoning in Geographic Information Systems*, eds. M. J. Egenhofer & R. G. Golledge, Oxford University Press, New York-NY-USA, pp. 25-39.