Incorporación de la temporalidad de un corpus histórico en un SIG

M. Guerrero Nieto, A. Urrutia Zambrana, M. J. García Rodríguez,
M. A. Bernabé Poveda

Grupo Mercator - Universidad Politécnica de Madrid
Cª de Valencia, km.77 28031 Madrid
{mguerrero, adolfo.urrutia, mjosegr}@topografia.upm.es
{ma.bernabe}@upm.es

Resumen

Este artículo aborda la integración de datos históricos procedentes de textos manuscritos originales en un sistema de información geográfica. El trabajo está enfocado a la unión de la temporalidad, los SIG y los corpus históricos. Para ello, se ha utilizado el Procesamiento del Lenguaje Natural (PLN), una campo esencial para el procesado computacional del lenguaje humano, que permite la relación de los textos originales y los SIG.

En los últimos años se ha investigado en distintos sistemas de extracción y recuperación de información temporal, así como en el reconocimiento y normalización de expresiones temporales. Al amparo de estas técnicas, el presente artículo tiene dos objetivos fundamentales: por un lado, la identificación y normalización de expresiones temporales referenciales, y por otro lado, la incorporación de la variable temporal extraída de corpus históricos en los SIG.

Para la identificación de las expresiones temporales se ha utilizado el lenguaje de marcado TimeML que permite su reconocimiento y normalización. Esta anotación confiere al texto una serie de etiquetas referentes al tiempo y a los eventos, con las cuales se extrae la información temporal a través de consultas. Éste constituye el primer paso para la integración de las expresiones temporales en un SIG. La estructuración en una base de datos, obtenida a partir de la anotación lingüística TimeML garantiza la incorporación en los sistemas geográficos.

Palabras clave: TimeML, PLN, anotación lingüística, corpus histórico, SIG.

1 Introducción

Todos los acontecimientos están situados en un tiempo y un espacio; ambos resultan imprescindibles para tener una representación completa de los hechos históricos. Las disciplinas que han estudiado tradicionalmente de forma separada estas dos variables son la Historia y la Geografía, aunque tanto la una como la otra son incomprensibles sin estas dimensiones. Para analizar y representar las cuestiones espaciales relacionadas con el territorio, se desarrolló el SIG, que en esencia es una base de datos a la que se puede preguntar y que se asocia a un sistema gráfico que permite visualizar la respuesta. La posibilidad de utilizar el SIG como una herramienta de análisis y representación de los hechos históricos, es algo que se ha planteado desde la década de 1970 [1] con el fin de reexaminar y fortalecer muchos aspectos de la historia geográfica [2].

El proyecto DynCoopNet[1], en el que se enmarca esta investigación, cuenta entre sus propósitos, indagar en las dinámicas de cooperación en los dominios ibéricos que se establecieron en lo que se ha dado en llamar la Primera Edad Global. Nuestra aportación en ese proyecto es evidenciar el potencial del SIG para la confrontación y revisión de los acontecimientos históricos e incorporar herramientas capaces de realizar análisis temporal.

Teniendo en cuenta las fuentes históricas se plantearon algunas iniciativas: por un lado, la posibilidad de detectar la información textual para que ésta fuera operable informáticamente y por otro se necesitaba extraer la información temporal de los textos. Como consecuencia, se ha abierto una línea que pueden dotar a los SIG de mayor funcionalidad: el Procesamiento del Lenguaje Natural (PLN), donde se incluye la anotación semántica temporal.

La cuestión temporal ha sido investigada desde diferentes disciplinas. Desde el punto de vista computacional, el procesamiento de la información temporal ha despertado un gran interés para la comunidad científica, prueba de ellos son los numerosos workshop que se han desarrollado en el área de la creación de herramientas de extracción y análisis temporal (TERCAS [3], TANGO [4], DAGSTUHL [5], MUC [6]), en el área de los lenguajes de anotación semántica temporal (TIDES [7], TimeML [8]), de sistemas de anotación (TERSEO [9]) y en diferentes encuentros de evaluación (TERN [10], TemEval [11]). Así mismo en la rama de la información geográfica, son numerosos los trabajos que han abordado este tema. Las últimas investigaciones sobre la incorporación del tiempo en los SIG han ido modificando el almacenamiento y modelado de los datos temporales (Moving Object Data Models [12], Spatio-Temporal Object-Oriented Data Model [13], Object-Relationship Model [14]). La imprecisión de ciertos conceptos temporales como granularidad, operaciones y/o densidad temporal y lo difuso de ciertas expresiones siguen sin facilitar la gestión de los acontecimientos temporales asociados al territorio.

La elección de los sistemas de anotación lingüística ha estado relacionada con el objetivo del proyecto, esto es, la incorporación de la variable tiempo, pues lo que se necesitaba era usar un lenguaje que ya estuviera siendo utilizado por herramientas geográficas y a su vez que permitiera la descripción de la variable y el intercambio de información. Los lenguajes de marcado, como por ejemplo SGML, son utilizados en los SIG para la representación espacial

[1] Proyecto financiado por el programa EUROCODES de la European Science Foundation, Dynamic Complexity of Cooperation-Based Self-Organizing Networks in the First Global Age (0-HUM2007-31128-E)

desde hace ya tiempo y han existido iniciativas para extender el lenguaje de marcado espacial sobre el dominio temporal para poder representar esta información [15].

Este artículo expone en primer lugar la identificación de expresiones temporales definidas y referenciales, mostrando cómo se organiza y describe el tiempo a través del lenguaje de marcado y cómo éste se adecua para la integración en una herramienta geográfica. En segundo lugar, expone la incorporación de la variable temporal extraída de corpus históricos en los SIG. Una característica que presenta ese corpus histórico es que procede de fuentes textuales originales. La información textual utilizada (corpus) está compuesta por cartas del comerciante castellano Simón Ruiz fechadas en el s. XVI.

Para abordar el primer objetivo se ha contado con el lenguaje de marcado TimeML que permite describir las expresiones temporales del corpus, tanto las expresiones definidas como las indefinidas; también define los eventos que aparecen y establece relaciones temporales. Estas relaciones se apoyan en el álgebra temporal de Allen [16]. Para el segundo objetivo se ha utilizado el corpus anotado, con el lenguaje de marcado especificado anteriormente, para el que se proponen una serie de procedimientos que permiten que la lingüística computacional y los sistemas de información geográfica puedan colaborar.

En la primera sección de este artículo se presenta el marco del proyecto y el objetivo del presente artículo, en la segunda parte se profundiza en el Procesamiento del Lenguaje Natural, donde se presta atención, sobre todo, a la anotación semántica temporal. Posteriormente se describe la guía de anotación utilizada para extraer información temporal y en la cuarta sección la metodología utilizada en la identificación y normalización de expresiones temporales del español, prestando además atención a la integración del TimeML en la geodatabase. En último lugar aparecen las conclusiones y se apuntan los futuros trabajos.

2 Procesamiento de Lenguaje Natural

El PLN es una subdisciplina de la Inteligencia Artificial que tiene como propósito el modelado y procesamiento computacional del lenguaje humano. En los repositorios de datos geográficos, los metadatos de un documento resultan cruciales para las consultas temporales, pero son insuficientes si se quiere consultar por la duración de acontecimientos u obtener otras fechas que no sean la fecha de publicación de dicho documento. Con la incorporación de la técnicas del PLN, se pretende utilizar las expresiones temporales lingüísticas de todo el documento, ya sean estas explícitas o implícitas, con el fin de conocer cuándo ocurren los eventos, cuánto duran, en qué periodo se dan, etc., extrayendo la información temporal, no sólo de los metadatos, sino de todo el texto. A continuación se describirán algunas de las técnicas utilizadas y se profundizará en el estado de la cuestión, donde se revisarán las principales aportaciones en estos temas.

2.1 Extracción y recuperación de información

Entre las técnicas de procesado de documentos cabe mencionar dos: los sistemas de recuperación de información (IR) y los de extracción de información (IE). En este artículo sólo hablaremos de la segunda técnica, puesto que se parte de un conjunto de documentos identificados. En los sistemas IE se intenta extraer aquellos hechos relevantes previamente establecidos presentes en los documentos. Algunas de las tareas propias de este último sistema son: reconocimiento de nombres de personas, organizaciones, lugares o expresiones temporales. Para evaluar estos sistemas de IE se crearon las conferencias MUC (*Message Understanding*

Conference) en cuya convocatoria MUC-6 se comenzaron a aplicar las etiquetas TIMEX para el reconocimiento y normalización de las expresiones temporales [6].

2.2 Anotación lingüística

Un documento en lenguaje natural puede ser marcado con el fin de identificar la información que se necesite; para ello enriquecemos el texto con etiquetas. Con éstas, la estructura de un documento se hará explícita, ya que otorgan información adicional. Las etiquetas sirven para describir fragmentos del texto cuya representación no tendrá un orden riguroso.

De entre los tipos de lenguaje de marcado destacan SGML y XML, que están diseñados para marcado descriptivo o semántico. Los lenguajes de marcado son usados por la Lingüística Computacional para incrustar la etiqueta en el texto mediante la anotación que se le está añadiendo al documento. Las ventajas de la anotación lingüística son: en primer lugar, la posibilidad de utilizar técnicas de aprendizaje y creación de herramientas que lleven a cabo la etiquetación automática de corpus, por otro lado, la reutilización del corpus anotado se convierte en una fuente de investigación científica para la investigación y el desarrollo futuros.

2.3 La anotación lingüística temporal

Los diferentes lenguajes de marcado con información temporal desarrollados en los últimos 10 años para representar la información incluida en lenguaje natural poseen distintas etiquetas y valores que están condicionados al objetivo de cada lenguaje.

- STAG [17] es una guía para anotar eventos y tiempos en textos periodísticos, cuya etiqueta para la información temporal es TIMEX.
- TIDES [7] está constituido como estándar de anotación de expresiones temporales. Desarrollado para marcar estas expresiones temporales de un documento e identificar el valor de la expresión temporal (TIMEX2).
- TimeML [9] anota eventos, expresiones temporales y relaciones temporales para ordenar cronológicamente los eventos. El corpus está compuesto por 183 artículos de la prensa americana [18]. Este sistema es el elegido en el marco de nuestra investigación.

2.4 Lenguaje de anotación temporal TimeML

La anotación semántica temporal TimeML es una especificación lingüística para anotar eventos y expresiones de tiempo. Ofrece una sistematización para la extracción y representación de información temporal así como para el intercambio de información. Nació con el fin de anotar artículos periodísticos aunque, como se comprobará, se puede extender a otro tipo de información textual. Posee las siguientes propiedades: interpreta las expresiones temporales, marca el tiempo de los eventos y ordena los eventos con respecto a otros a través de un anclaje temporal. TimeML desarrollado en 2002 [3] [4], está siendo consolidado como estándar ISO (ISO WD 24617-1:2007), y es compatible con la forma ISO 8601 que especifica la notación estándar para almacenar fechas. Hay que señalar que ha sido aprobado como lenguaje de anotación para TempEval, cuyo objetivo es evaluar los sistemas automáticos en el análisis semántico de textos [11].

TimeML combina y extiende características de otros estándares de anotación temporal como los mencionados anteriormente: STAG y TIDES. En TimeML, las expresiones de tiempo están marcadas con la etiqueta TIMEX3, lo que representa una mejora respecto a las etiquetas anteriores.

Para el tratamiento de los diferentes *timexes* existen distintos lenguajes de anotación y un corpus anotado para la lengua inglesa, TimeBank [18]. También existen herramientas automáticas de anotación temporal TARSQI [19] y TERSEO [8] e incluso otras tareas asociadas a la información temporal, como las ontologías Cycorp, KSL 1991, DAML 2002, OWL time 2006 [20]. Sin embargo, la mayoría de esos recursos no pueden utilizarse para el español o se han quedado obsoletos, por lo que sería necesaria la creación de corpora españoles anotados en TimeML y el desarrollo de herramientas específicas.

2.5 Descripción y características del TimeML

Este lenguaje de marcado contiene tres etiquetas básicas: TIMEX3, EVENT, SIGNAL y tres subtipos de link: TLINK, ALINK y SLINK. Se procede a explicar brevemente cada una de las etiquetas:

- TIMEX3 es usada para marcar expresiones temporales: *4 de noviembre del 2009, ayer, a las 6 de la tarde, este sábado, el próximo año.*
- EVENT es usada para marcar eventos mencionados de un texto: *ocurrir, creer, estudiar, empezar.*
- SIGNAL es usada para anotar señales temporales: *antes, después, durante.*
- TLINK es usada para marcar las relaciones temporales: *Luisa fue a Murcia del 4 al 6 de noviembre* (se relaciona la información temporal con el evento *ir*).
- ALINK es usada para anotar las relaciones aspectuales: *María empezará a presentar su artículo a las 12 de la mañana* (el verbo *empezar* está mostrando una fase del evento).
- SLINK es usada para anotar relaciones de modalidad o evidencialidad: *Juan dijo que iría a Murcia en noviembre.* (muestra el tipo de relación que hay entre los dos eventos: *decir* e *ir*).

TimeML ofrece la posibilidad de expresar distintas granularidades. Posee cuatro tipos para la expresión de tiempo (TIMEX3):

- DATES se utiliza para expresiones que se refieren a un calendario: *a 3 de agosto de 2005, el domingo pasado, ayer por la mañana,* etc.
- DAY TIMES es utilizada para una expresión temporal que es menor a un día: *esta tarde, a las tres menos veinte.* La distinción entre estos dos tipos de tiempos tiene especial atención por la diferencia entre la granularidad de las expresiones.
- DURATION es usada para describir una duración en el tiempo: *durante cuatro días, hace dos años.*
- SET es usada para expresiones que se repiten en el tiempo: *dos veces a la semana, cada ocho días.*

El lenguaje natural no tiene una única manera de expresar una granularidad específica, sino que puede haber diferentes expresiones temporales que refieran al mismo granulo. La granularidad es el nivel de detalle con el que se mide el tiempo y se puede decir que el lenguaje natural no tiene una forma canónica de expresar el tiempo [21], lo que sí que se sabe es que las expresiones temporales lingüísticas difieren en la granularidad pero que se ajustan al calendario gregoriano. Se pueden encontrar equivalencias entre el lenguaje natural y este calendario, al menos en las lenguas occidentales.

Como puede observarse en la clasificación de los TIMEX3 algunas expresiones lingüísticas son deícticas (*ayer*, *a las 3 de la tarde*, *a diez días vista*, *desde el 28 del mes pasado*), esto es, necesitan del conocimiento del momento narrativo en el que se enmarcan para poder precisar el intervalo de tiempo comprendido por la expresión. El corpus que se ha utilizado permite usar los metadatos temporales con el fin de poder determinar en qué momento ocurren los eventos y así poderlos situar en una línea de tiempo. Esto se consigue con el atributo *AnchorTime* que permite establecer un eje temporal.

La estructura del tiempo se aprecia en la Figura 1. donde se recoge la definición del tipo de documento (DTD), a modo de descripción gráfica, una descripción de estructura y sintaxis de un documento XML o SGML. Su función básica es la descripción del formato de datos, para usar un formato común y mantener la consistencia entre todos los documentos que utilicen la misma DTD. De esta forma, dichos documentos, pueden ser validados, al conocerse la estructura de los elementos y la descripción de los datos que trae consigo cada documento. En este gráfico se observan tres de los elementos que componen las etiquetas del TimeML.

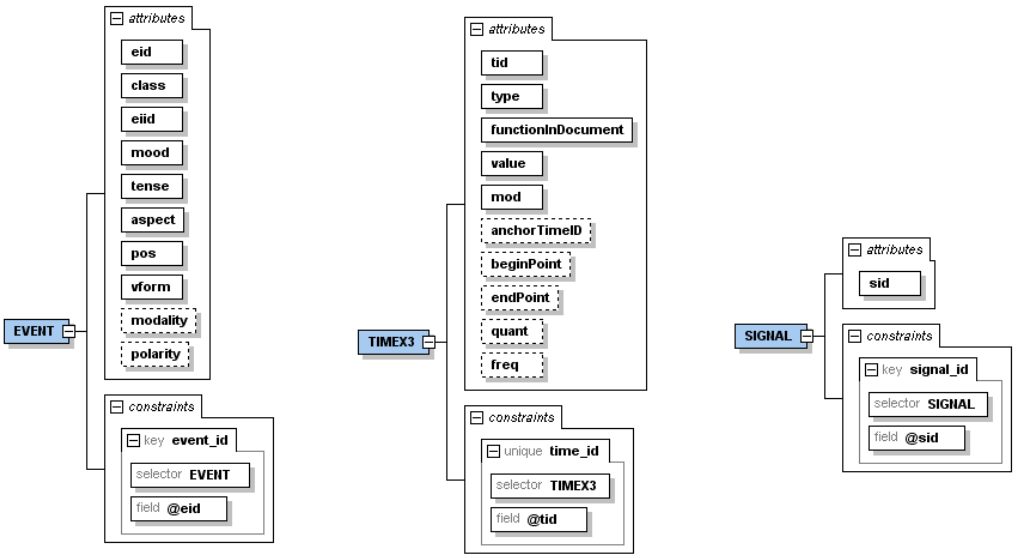

Figura 1. Esquema de TimeML (formato DTD), donde se observan las etiquetas y los atributos de EVENT, TIMEX3 y TLINK. Fuente: Elaboración propia.

3 Metodología para el procesado de expresiones temporales de un corpus histórico

El corpus utilizado procede de una selección de 20 cartas del comerciante español Simón Ruiz fechadas en el s. XVI. El objetivo de esta metodología es la incorporación en un SIG de la variable temporal descrita a través de un lenguaje de marcado temporal y procedente del texto original.

Se ha dividido en tres pasos:

1. Procesado lingüístico del corpus, el cual contiene las tareas de tonkenización, lematización y POS-tagger.

2. Identificación y normalización manual de las expresiones temporales con TimeML.

3. Incorporación de TimeML dentro de una geodatabase: conceptos informáticos en los que se basa la integración y diseño de la estructura de la base de datos.

3.1 Procesado del corpus

Con el fin de extraer la información temporal del corpus, propone seguir el siguiente proceso:

1. *Tokenizador:* consiste en descomponer una expresión compleja en otras más sencillas para trabajar con éstas últimas por separado. El sistema debe ser capaz de segmentar en vocablos y oraciones.

2. *Lematización:* se construyen todas las palabras pertenecientes a la familia lingüística de cada lema, usando para ello la clase morfológica.

3. *POS-tagged:* este método etiqueta morfológica y lexicalmente cada una de las palabras del texto con la información que se dispone en un lexicón, diccionario de dominio específico.

El procesador automático que contiene las diferentes tareas descritas es *Tree-Tagger* [22]. La elección del analizador se ha debido a la facilidad de incorporación de léxico nuevo dentro de su lexicón. La dificultad en cuanto a la lematización consiste en el procedimiento, ya que debido a las características específicas de los textos (español del s.XVI): ausencia de normalización ortográfica, uso de abreviaturas y aglutinaciones, etc., se hace necesaria la lematización semiautomática. El POS-tagger también se ha realizado de forma semiautomática pues el léxico que posee es del español actual. El *Tree-Tagger* permite la incorporación de léxico nuevo de forma manual. El hecho de que esta herramienta no sea capaz de adivinar etiquetas de palabras que no aparezcan en su lexicón, es lo que ha conducido al procedimiento semiautomático.

3.2 Anotación temporal del corpus

TimeML contiene una guía para identificar y normalizar expresiones temporales, eventos y relaciones temporales. Actualmente esta guía existe para el inglés, chino e italiano, aunque está en curso la adaptación para el español [23] [24]. Al tratarse de un corpus histórico del español, ha sido necesario adaptar la guía de acuerdo a la variedad lingüística. Es pertinente señalar que por el momento no ha habido trabajos enfocados en la especialización de correspondencia antigua en castellano en XML ni en otro formato que incluya información temporal. Hasta ahora, el terreno en el que se han movido los trabajos sobre extracción temporal se han enmarcado en textos periodísticos o textos legales [25].

Tras la identificación y normalización de las expresiones temporales en castellano renacentista, el siguiente paso ha sido la anotación en TimeML. Para ello se cuenta con la gramática (DTD) que contiene todos los elementos, valores ya tributos de los que está compuesto el TimeML (ver Figura 1).

A continuación se muestra un ejemplo del corpus de la normalización de estas expresiones temporales en lenguaje TimeML, donde aparecen los valores de la guía. Se ha elegido un ejemplo del corpus donde aparezcan TIMEX3, EVENT y TLINK.

"La de v.m. de 15 deste he recebido en este dia"

```
<TIMEX3 tid="tid12" type="DATE" value="1567-05-15" anchorTimeID="tid11">15
deste</TIMEX3>
<EVENT eid="eid28" aspect="PERFECTIVE" mood="NONE" pos="VERB" vform="NONE"
class="OCCURRENCE" tense="PRESENT" stem="RECIBIR">recebido</EVENT>
<TIMEX3 tid="tid13" type="DATE" value="1567-05-28" anchorTimeID="tid11">este
dia</TIMEX3>
<TIMEX3 tid="tid11" type="DATE" value="1567-05-28" >28 de mayo de 1567</TIMEX3>
<TLINK relType="INCLUDEs" lid="lid31" timeID="tid13" relatedToEventInstance="eid28"/>
<TLINK relType="BEFORE" lid="lid31" timeID="tid12"relatedToEventInstance="eid28"/>
```

El tipo de información temporal que podemos encontrar en las cartas es muy variado y rico debido a la retórica de la época y al tipo de documento, con expresiones temporales del tipo: "*de pocos días a esta parte*", "*muchos días ha*" o "*a ocho deste*". Como puede observarse en el ejemplo, las expresiones lingüísticas utilizadas en las cartas pueden ser deícticas, esto es, necesitan del conocimiento del momento narrativo en el que se enmarcan para poder precisar el intervalo de tiempo comprendido por la expresión. El corpus utilizado permite utilizar los metadatos temporales con el fin de poder determinar en qué momento ocurren los eventos y así poderlos situar en una línea de tiempo.

Para ordenar los eventos del corpus se puede hacer de dos maneras: extrínseca e intrínseca. La manera extrínseca es ordenar las cartas de Simón Ruiz teniendo en cuenta sólo las fechas de publicación del documento, esto es, los metadatos. Intrínsecamente consiste en ordenar todo las expresiones temporales que aparezcan en el documento. Al ser éste un proceso más sofisticado es necesario acudir a todo este procesado de la información del corpus.

A continuación se presenta una muestra gráfica del ordenamiento temporal (ver Figura 2) que se puede llevar a cabo siguiendo el TimeML. La etiqueta que marca las relaciones temporales es el TLINK, basándose éstas en las trece relaciones binarias del álgebra temporal de Allen.

Los TLINK representan las relaciones temporales existentes entre dos eventos, dos tiempos o entre un evento y un tiempo. En la Figura 2. se ilustra un ejemplo de TLINK. En el ejemplo, el evento sería "*he recebido*" que va acompañado de dos expresiones temporales "*de 15 desde*" y "*en este dia*". Las relaciones temporales entre estos tres elementos se marcan con la etiqueta TLINK de la manera que se puede ver en la Figura 2.

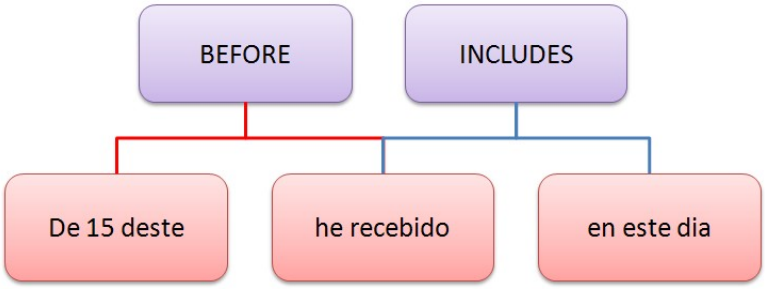

Figura 2. Ejemplo gráfico de TLINK. Fuente: Elaboración propia.

Para la detección de expresiones temporales se ha utilizado la aproximación por patrones. Consiste en la identificación de expresiones temporales de manera manual, posteriormente se generan patrones con el fin de generalizar esas reglas. Una vez conseguido esto, se pretende incorporar esos patrones de manera automática. Fase aún por desarrollar.

3.3 Incorporación de TimeML como parte de una geodatabase

Una vez anotado el corpus en TimeML, se continúa con la integración del texto en el sistema de información geográfica. Los SIG tienen diferentes fuentes de datos: geodatabases, tablas en Ms Access, tablas en Ms Excel (con ciertas restricciones), etc. Sin embargo, los archivos XML, aunque se utilizan para el intercambio de datos entre las bases de datos de diferentes SIG, no forman parte de estas fuentes.

A pesar de este hecho, la DTD de TimeML proporciona una estructura estable y predecible, por lo que se podría diseñar una base de datos relacional para almacenar la información contendía en los atributos de cada uno de los elementos del archivo. Se requiere la creación de un algoritmo de mapeo entre ambas estructuras (BD y Corpus) para poder guardar y extraer la información libremente. Tal herramienta, como también se especifica en el método, podría ser implementada como un módulo interno del gestor de la base de datos o como un componente de software independiente [26].

Debido a lo anterior, las dos entidades (geodatabase y DTD) serían prácticamente idénticas. La única diferencia sería el tipo de fichero sobre el cual estarían constituidas. Esto facilitará la introducción de la información, además que las expresiones anotadas del corpus no sufrirían ningún cambio. El XML y la geodatabase se habrían convertido en las dos caras del almacenamiento de las expresiones temporales. Es pertinente mencionar que el XML no es inherente al TimeML, ya que este puede convertirse en otros formatos.

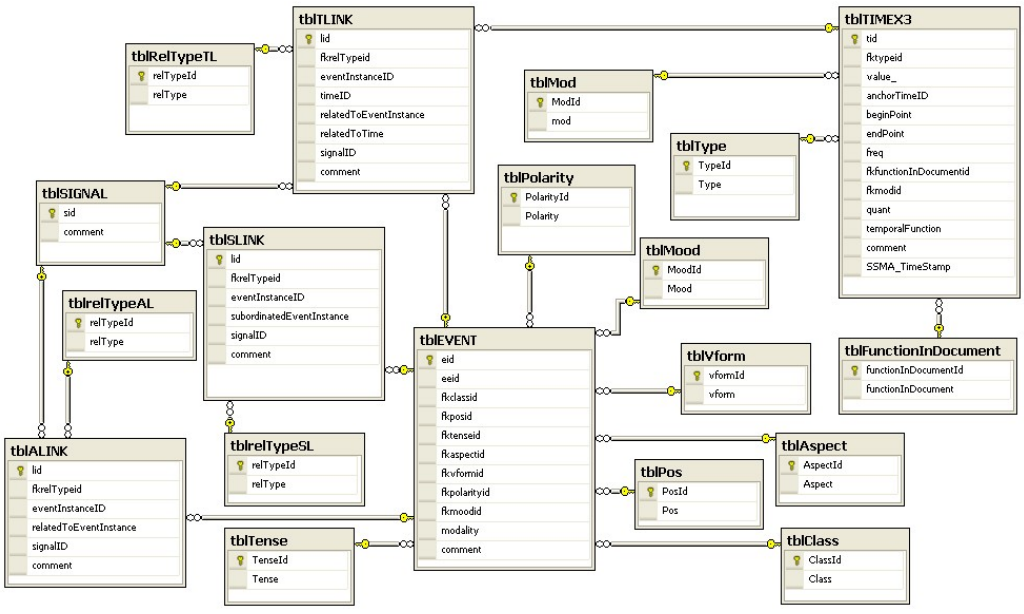

Figura 3. Estructura de la geodatabase basada en la DTD de TimeML, cada una de las etiquetas contiene su tabla correspondiente. Fuente: Elaboración propia.

El traspaso de información, mostrado en la Figura 4., indica cómo cada columna de la estructura TIMEX3, vista desde TimeML, se corresponde con una columna en la tabla tblTIMEX3 dentro de la geodatabase.

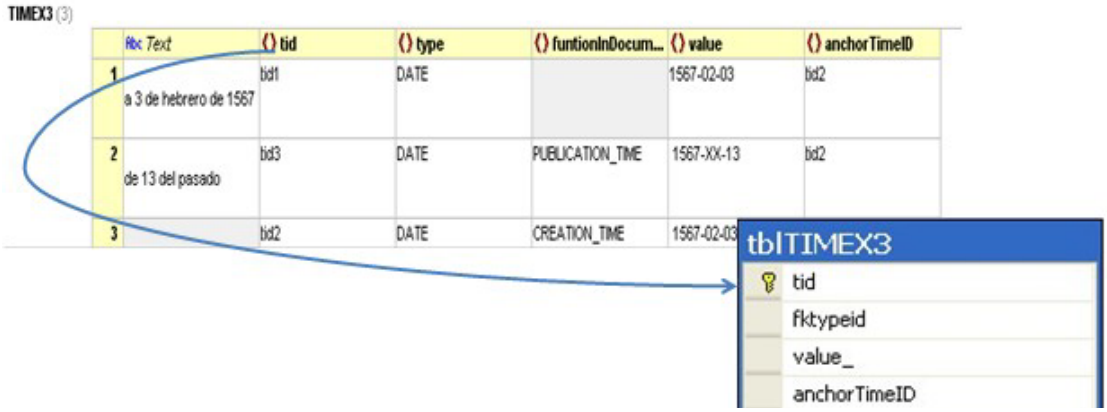

Figura 4. Muestra el proceso de trasladar la información anotada en TimeML hacia la geodatabase. Fuente: Elaboración propia.

Finalmente, con la información dentro de la geodatabase, la representación del corpus anotado dependerá de las características de éste. Por ejemplo, si se tratase del texto de la bitácora de un capitán de barco, se hará énfasis en la representación de las corrientes oceánicas, los vientos y tormentas, sin embrago, si el corpus estuviera compuesto por textos que describieran viajes por tierra, los detalles de la presentación serían sustancialmente distintos, dando importancia a otros aspectos. A este respecto se puede añadir, en cuanto a la representación de la anotación lingüística de corpus, que es una línea futura de investigación.

4 Conclusión y trabajos futuros

Se ha diseñado una metodología para el reconocimiento y la normalización de expresiones temporales siguiendo las especificaciones del TimeML, presentando el procedimiento a seguir y apuntando la unión de dos ámbitos para el desarrollo de la temporalidad en el SIG. Asimismo se ha presentado la metodología para la conexión entre la anotación lingüística de corpus y los sistemas de información geográfica. También se han expuesto las ventajas funcionales del propósito de integrar textos en lenguaje natural y la representación de la temporalidad en un SIG. De igual modo, se han descrito las ventajas de utilizar el TimeML debido a su carácter estándar, su formato como base de datos, su aplicabilidad a cualquier lengua, pues posee una gramática definida, y sobre todo sus propiedades, ya que permite el ordenamiento de los eventos en una línea de tiempo. TimeML, como otros lenguajes de marcado, permite el tratamiento masivo de información textual.

Se han descrito algunas de las limitaciones que se han observado para llevar a cabo la propuesta:

- Los ficheros XML no forman parte de las fuentes de datos de las cuales un SIG pueda leer directamente.
- Para lograr la representación de la información textual etiquetada, el sistema de información geográfica deberá tener una base de datos espacio-temporal que le permita almacenar y consultar la información proveniente del corpus, es decir, se hace necesario un SIG temporal que refleje el TimeML.

- Escasez de corpus etiquetados en TimeML para lenguas diferentes al inglés, lo que impide la utilización de técnicas de aprendizaje automático, ocasionando el uso de la anotación semiautomática y manual.

- Se necesita la adaptación del TimeML al castellano antiguo para la identificación de expresiones temporales en este tipo de textos.

Entre los trabajos futuros:

- Se pretende el reconocimiento, normalización y cuantificación de expresiones temporales en corpus históricos del español, así como la integración de anotación lingüística temporal y espacial.

- Se prevé la creación de una herramienta de análisis que permita la utilización de expresión temporales en el momento de especificar la consulta dentro de un SIG espacio-temporal, así como la ampliación de las consultas en SQL con expresiones temporales difusas y nombres propios temporales.

- Integración de las consultas de las expresiones temporales en un SIG.

Referencias

[1] Sack R.D.: Chronology and Spatial Analysis, Annals of the Association of American Geographers, 64, pp.439-452 (1974)

[2] Gregory, I.N.,Ell, P.S.: Historical GIS: Technologies, Methodologies and Scholarships, Cambridge Univesity Press (2007)

[3] Pustejovsky, J.: TERQAS: Time and Event Recognition for Question Answering Systems, ARDA Workshop, MITRE, Boston (2002) http://www.timeml.org/site/terqas/index.html

[4] TANGO (TimeML Annotation Grafhical Organizer): http://www.timeml.org/site/tango/index.html

[5] Dagstuhl Seminar Proceedings. Annotating, Extracting and Reasoning about Time and Events. http://drops.dagstuhl.de/opus/volltexte/2005/313/

[6] Advanced Research Projects Agency, Proceedings of the Sixth Message Understanding Conference (MUC-6) (1995), Software and Intelligent Systems Technology Office.

[7] Ferro, L., Gerber, L., Mani, I., Sundheim, B., & Wilson, G.: TIDES 2005 Standard for the Annotation of Temporal Expressions, The MITRE Corporation (2005)

[8] Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: TimeML: Robust specification of event and temporal expressions in text, In AAAI Spring Symposium on New Directions in Question-Answering (Working Papers), Stanford, CA, pp. 28–34 (2003)

[9] Saquete, E., Martínez-Barco, P., Muñoz, R., Negri, M., Speranza, M., Sprugnoli, R.: Automatic resolution rule assignment to multilingual Temporal Expressions using annotated corpora, Procedings of the Thirteenth International Symposium on Temporal Representations and Reasoning, pp.218-224 (2006)

[10] DARPA TIDES (Translingual Information Detection, Extraction and Summarization). The TERN evaluation plan: Time Expression Recognition and Normalization. Working papers, TERN EvaluationWorkshop (2004) http://timex2.mitre.org/tern.html

[11] Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: SemEval-2007, Task 15: TempEval Temporal Relation Identification, In *Proceedings of SemEval 2007*, 4th InternationalWorkshop on Semantic Evaluation, pp.75-80, Prague, ACL (2007). http://nlp.cs.swarthmore.edu/semeval/tasks/index.php

[12] Erwig, M., Guting, R.H., Schneider, M., Vazirgiannis, M.: Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases, GeoInformatica, 3(3): 265-291, (1999).

[13] Montgomery, L. D.: Temporal Geographic Information Systems Technology and Requirements: Where We are Today, Thesis, Department of Geography, Ohio State University, USA (1995).

[14] Claramunt, C., Parent, C., Spaccapietra, S., Theriault, M.: Database Modeling for environmental and Land Use Changes, Geographical Information and Planning, Chapter 20, Springer-Verlag (1998)

[15] Zipf, A., Krüger, S.: TGML - Extending GML by Temporal Constructs - A Proposal for a Spatiotemporal Framework in XML, ACM-GIS 2001,The Ninth ACM International Symp. on Advances in Geographic Information Systems, Atlanta, USA (2001)

[16] Allen, J. F.: Maintaining knowledge about temporal interval, Comunications of ACM, 26, 11, pp. 832-843 (1983)

[17] Setzer, A., Gaizauskas, R.: Annotating Events and Temporal Information In Newswire Text, In LREC 2000, pp. 1287-1294, Athens (2000).

[18] Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day,D., Ferro, L., Lazo, M.: The TIMEBANK Corpus. Proceedings of Corpus Linguistics 2003, pp. 647-656 (2003). http://www.timeml.org/site/timebank/documentation-1.2.html

[19] Verhagen, M., Mani, I., Saurí, R., Littman, J., Knippen, R., Jang, S. B., Rumshiský, A., Phillips, J., Pustejovský, J.: Automating temporal annotation with TARSQI, In 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan (2005)

[20] Hobbs, J. R., Pustejovsky, J.: Annotating and Reasoning about Time and Events, Proceedings, AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford, California (2003).

[21] Llidó Escrivá, D. M.: Extracción y recuperación de la información temporal, Universidad Jaume I, Castellón, Tesis Doctoral (2002)

[22] Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decission Trees, Proceedings of the International Conferenceon New Methods in Language Processing, Manchester, UK, pp. 44-49 (1994).

[23] Saurí, R., Pustejovsky, J.: Annotating Time Expressions in Spanish, TimeML Annotation Guidelines (in Press).

[24] Saurí, R., Batiukova, O., Pustejovsky, J.: Annotating Events in Spanish, TimeML Annotation Guidelines (in Press).

[25] Schilder, F., and Mcculloh, A.: Temporal information extraction from legal documents. In Katz et al. http://drops.dagstuhl.de/opus/volltexte/2005/318/

[26] Ramez, E.: Fundamentals of database systems, Pearson Education, 4th ed., pp. 842-856 (2004)