

LA REPRESENTACIÓN DE DOCUMENTOS HISTÓRICOS: APLICACIÓN DE ESQUEMAS Y XML PARA MODELAR Y REPRESENTAR LA INFORMACIÓN Y EL CONOCIMIENTO DESCRITO EN LAS FUENTES DE PROTOCOLOS NOTARIALES.

- Autores:** M. A. Serrano-Tenllado
A. Calvo
J. A. Romero
Departamento de Informática y Análisis Numérico Universidad de Córdoba
- Resumen** El presente trabajo presenta algunos de los resultados preliminares obtenidos en la formalización de una de las fuentes históricas más valoradas en la investigación dentro de la Historia Social: los protocolos notariales. Siendo conscientes de la complejidad que tiene este tipo de fuentes se han identificado y modelado las principales estructuras descriptivas de las mismas, es decir, personas físicas, personas jurídicas, bienes, lugar y tiempo etc., así como las descripciones de las principales relaciones que intervienen en ellas. La formalización ha sido realizada en varios sistemas: CML, DTD y Esquemas. De momento, son estos últimos los que parecen reunir las mejores condiciones ya que, unido a su gran riqueza semántica que permite especificar restricciones y axiomas sobre los atributos de los elementos modelados, presentan la ventaja de la disponibilidad de herramientas informáticas que permiten su verificación y su utilización como sistema de almacenamiento de información.
- Palabras clave:** Ontologías, XML, fuentes históricas, protocolos notariales
- Abstract** The present work shows some of the preliminary results obtained in formalization of one of the most vaulted historical sources in the research into Social History: the notarial protocols. Being aware of the complexity of this kind of sources, their main descriptive structures have been identified and modelled, that is to say, individuals, legal entities, items, place and time. At the same time as the descriptions of the main relation ships which take part into them. The ontology has been made in several systems CML, DTD y Schemes. At the moment, the last ones are those which seem to have the best conditions. This is because, together with their big semantic richness, that allows to specify restrictions and axioms about the attributes of the model items, show the advantage of the availability of computer tools, which permit their verification and use as a system of information storage.
- Keywords:** Ontology, sources history, notarial protocols

1 Introducción

1.1 Génesis, motivación y objeto del trabajo

Este trabajo se enmarca dentro de un proyecto global sobre la construcción de un sistema experto en el campo de las ciencias históricas [2]. De acuerdo a la metodología CommonKads [15] una de las actividades

necesarias para lograr este objetivo es la construcción de ontologías reusables [8] [9] que representen el conocimiento estático del dominio. Este trabajo se centra en la identificación, modelado y representación del conocimiento del dominio en las fuentes históricas de protocolos notariales, analizando las principales dificultades encontradas y las soluciones aportadas para resolverlas. Se discute, asimismo, la necesidad de mantener de forma explícita y formal estas representaciones para lo que se hace uso del lenguaje CML y de la creación de un lenguaje de marcas basado en DTD y esquemas para su representación en XML [7].

1.2 Límites y justificación de los mismos

Pretender conceptualizar el mundo es algo imposible de realizar y sólo grandes proyectos de investigación como CYC [11] tienen como objetivo la obtención de tales tipos de conceptualizaciones en proyectos sobre la captura del conocimiento humano de sentido común. Para nuestros propósitos de investigación, (desarrollo de un sistema basado en el conocimiento en la investigación histórica dentro del campo de la historia social), vamos a restringir esta conceptualización al ámbito de este problema, si bien, muchos de los resultados que se obtengan aquí podrán ser reutilizados para otros fines diferentes. En cuanto a las fuentes utilizadas para este estudio hemos limitado las mismas a los protocolos notariales, no porque otras fuentes históricas no sean de interés, sino por cuestiones pragmáticas. Si los resultados que se deriven de las mismas resultan aceptables se habrá logrado implantar un método de trabajo, y la extensión a otras fuentes sería sencilla. Las fuentes notariales no son, sin embargo, fáciles de tratar y ofrecen las suficientes dificultades como para evaluar correctamente la metodología empleada.

1.2.1 Fuentes de protocolos notariales

¿Qué es una fuente histórica? Aróstegui [1] define la fuente histórica como:

*"Todo aquel objeto material, instrumento o herramienta, símbolo o discurso intelectual, que procede de la creatividad humana, a cuyo través puede **inferirse** algo acerca de una determinada situación social en el tiempo"*

Si nos centramos en las fuentes documentales y, más concretamente, en las notariales, las fuentes históricas constituyen fundamentalmente un conjunto de documentos escritos expresados en **lenguaje natural** en las que aparecen **descripciones** de la realidad histórica que responden, en la mayoría de los casos, a situaciones *estereotipadas*, guiones [14] en el contexto de la Inteligencia Artificial, en la que las personas desempeñan unos papeles claramente establecidos. Estas descripciones también hacen referencia a objetos, lugares y tiempo que tienen relación con los personajes y objetos descritos en ellas. El historiador, a través del estudio y análisis de estos documentos, trata de **reconstruir** esta realidad. Las **descripciones de la realidad** en las fuentes históricas son imprecisas, redundantes y, a veces, poco fiables. Desde el punto de vista informático, no son más que un conjunto de

caracteres agrupados en palabras y otros símbolos a los que el historiador da significado y construye con ellos una representación más estructurada de las descripciones de la realidad, lo que le permite obtener un **conocimiento histórico** a través de procesos cognitivos que tienen lugar en su mente. Este conocimiento histórico, expresado mediante lo que llamaremos, **discurso histórico**, vuelve a ser un conjunto de símbolos expresados en lenguaje natural.

Una vez seleccionados los problemas históricos y elegidas las fuentes de investigación que le ayuden en su resolución, el historiador se plantea el problema de procesar estas fuentes. Indudablemente hay un conocimiento específico (implícito y explícito) en el historiador que relaciona los problemas planteados con las fuentes históricas necesarias para resolverlos. A este conocimiento nos referiremos como **conocimiento sobre las fuentes históricas**. Un claro ejemplo de ello, para las fuentes notariales, lo encontramos en varias monografías que tratan este tema[12] [13].

1.3 Antecedentes

La importancia de este tipo de fuentes ya ha sido puesta de manifiesto por muchos investigadores y no vamos a insistir en ello. Estos trabajos hacen hincapié en dos aspectos fundamentales: de una parte analizan la estructura de las diversas actas notariales y ponen de manifiesto la importancia que cada una de las mismas aporta a la solución de problemas históricos; de otra parte, en su análisis estructural destacan las partes básicas de la escritura, pero no profundizan de forma detallada ni hacen mención a su formalización.

En el campo de la informática el desarrollo de metodologías para la creación de sistemas de información, primero, y para el desarrollo de sistemas basados en conocimiento después, ha puesto de manifiesto la necesidad de identificar y modelar el denominado **conocimiento estático del dominio**, esto es, un conjunto de procesos, conceptos y relaciones con sus respectivos atributos y con las consiguientes restricciones, reglas y axiomas sobre ellos, que caracterizan una visión del mundo que se quiere analizar y representar. Propuestas como el modelo entidad relación de Peter Chen [3], la metodología OMT y UML en la programación orientada a objetos, la creación de ontologías mediante lenguajes como CML [15] en CommonKads etc., son algunas de las propuestas que se han desarrollado en los últimos años para hacer rentables y viables la creación de grandes sistemas de información y conocimiento. Otras opciones que también han surgido han sido los lenguajes marcados para la representación de documentos escritos. Iniciativas como SGML [7], el proyecto TEI [5], HTML y últimamente XML [7], que permite la creación de lenguajes a medida para la representación de documentos, no son más que intentos de describir el mundo de una manera formal para que sea procesable y entendible por un ordenador. Si somos capaces de trasladar a estructuras organizadas de símbolos las palabras que aparecen en lenguaje natural en un documento histórico, un ordenador podría procesar dichas estructuras y realizar inferencias con ellas.

Una vez modelado el mundo que deseamos representar podemos optar por diversas alternativas de representación simbólica: tablas relacionadas mediante las bases de datos relacionales, marcos u objetos en el entorno de la

programación orientada a objetos, elementos simples y complejos caracterizados mediante marcas a través de lenguajes obtenidos mediante XML, etc. Sí parece evidente que, si el mundo que se quiere representar es muy complejo, harán falta herramientas que faciliten la depuración, completitud y reusabilidad de tales conceptualizaciones, llamadas también ontologías[9]. Sistemas como Ontolingua [6] , Ode [8], CML [15], definición de DTD [7] y esquemas [4] son propuestas que tratan de dar solución a los problemas planteados. Muchas de estas herramientas para conceptualizar presentan facilidades para traducir las ontologías de un sistema a otro, garantizando así la portabilidad de los resultados. El presente trabajo presenta los primeros resultados obtenidos sobre la creación de una de estas ontologías o conceptualizaciones¹ para representar las fuentes de protocolos notariales. Tras unos primeros intentos de desarrollo sobre el modelo entidad-relación de Chen, CML, DTD, finalmente se ha optado por presentar una conceptualización basada en Esquemas.

2 Objetivos

El objetivo fundamental que se persigue es **la identificación, modelado y representación** del conocimiento estático del dominio en las fuentes de protocolos notariales como parte integrante de un sistema basado en el conocimiento en el campo de las ciencias históricas. En particular, se pretende determinar estructuras de modelado y representación que sirvan para la creación de estructuras de datos para el almacenamiento de la información descrita en los documentos históricos de protocolos notariales. Una vez representada dicha información se habilitarán los mecanismos de razonamiento necesarios para inferir nueva información que permita la consecución del objetivo global expuesto.

3 Metodología

Tras una lectura y estudio detallado de más de mil escrituras notariales de la más variada tipología, hemos llegado a una serie de conclusiones que, si bien pueden parecer obvias, no nos eran tan evidentes al comienzo del trabajo.

1. Las escrituras notariales recogen **descripciones** de entidades y relaciones del mundo que se desea analizar. Estas descripciones son a veces muy complejas con un alto grado de recursividad. Entre las relaciones que surgen destacamos dos relaciones importantes que revelan la estructura interna de las fuentes: la relación **es-un** y la relación **forma-parte-de**.
2. Las descripciones son repetitivas, incompletas, y no siempre gozan de una alta fiabilidad.
3. Las descripciones hacen referencia a un tiempo verbal: presente, pasado o futuro, con relación al momento en que fue otorgada la escritura.
4. Las descripciones están expresadas en lenguaje natural.

Veamos un ejemplo para apreciar lo expuesto:

¹ Hemos de advertir que las ontologías creadas no están suficientemente optimizadas y es nuestro objetivo inmediato evaluarla dentro de diferentes entornos de desarrollo de ontologías como Ontolingua y Ode.

"D. Juan Alvarez de Sotomayor y Angulo otorga en arrendamiento a Cristóbal Muñoz de Espejo 4 fanegas de tierra en el partido de la mata que son las suertes número 10 y número 11 durante dos años, uno de barbecho y otro de sementera desde 1662 en 64 reales anuales por fanega que se pagarán en San Juan. Estas tierras pertenecen al vínculo que fundó el licenciado D. Gaspar Alvarez de Sotomayor."

En este resumen de un documento observamos que mantiene una estructura que es necesario identificar y registrar.

Para facilitar la reusabilidad se optó por formalizar las principales entidades y relaciones que intervienen de forma genérica en todas las escrituras. Así, entre otras, han quedado formalizadas:

- 1) Descripción de una persona física: distinguiendo entre hombre y mujer.
 - a) Relaciones familiares, de vecindad, de ocupación de cargos, etc.
 - b) Relaciones económicas, como posesión de bienes, etc.
- 2) Descripción de la persona jurídica: Comunidades, asociaciones, fundaciones, etc.
- 3) Descripción de bienes: Inmuebles, muebles, semovientes, mobiliarios, etc.
- 4) Descripción del tiempo.
- 5) Descripción del espacio: localidad, países, reinos, etc.

Una vez formalizadas estas estructuras descriptivas fueron formalizadas las escrituras notariales. Hemos entendido que en una escritura notarial pueden registrarse varios actos jurídicos notariales que han sido igualmente formalizados.

4 Resultados.

De acuerdo con la metodología propuesta se han obtenido varias ontologías para representar el conocimiento estático del dominio. Entre otras, destacamos las siguientes:

1. Ontología para describir algunos **problemas históricos** en el ámbito de la historia social.
2. Ontología para la representación sobre la **realidad histórica** en dicho ámbito: sujeto de derecho, persona física, persona jurídica, bienes, bienes rústicos, etc. Esta ontología recoge las estructuras que se **rellenarán** mediante inferencias a partir de las descripciones de las fuentes.

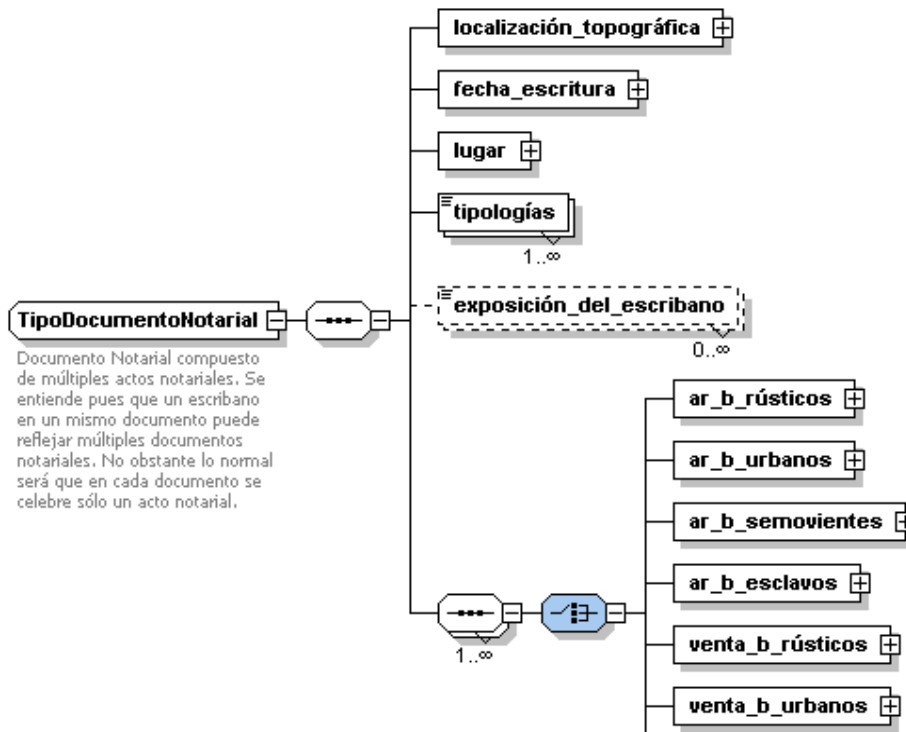


Figura 1 Estructura de una escritura notarial

3. Ontología sobre las fuentes históricas: Protocolos notariales.

Presentaremos a continuación, a modo de ejemplo, la caracterización de las descripciones de algunos elementos obtenidos en la ontología sobre los protocolos notariales.

4.1 Descripciones de la realidad en las fuentes de protocolos notariales

4.2 La escritura notarial

La figura 1 recoge la estructura general de un documento notarial en la que puede observarse que el documento se compone de una secuencia de elementos complejos que describen la localización topográfica del documento, la fecha de la escritura, el lugar y diferentes tipos de actos notariales que puede contener. Al mismo tiempo, recoge la exposición del escribano en lenguaje natural.² Finalmente se recoge una secuencia de uno o más actos notariales a seleccionar de entre todos los representados.

4.2.1 La comparecencia. Tipos de comparecencia

Para cada acto notarial se contemplan varios elementos. Las figuras 2 y 3 nos muestran la sintaxis y la descripción de un tipo de escritura de arrendamiento de bienes raíces rústicos formada por la comparecencia ante el

² Hemos de señalar que la información se recoge en muchos casos tanto en lenguaje natural como en estructuras de elementos. Esto es por la previsión que se ha hecho para que el sistema pueda incorporar interpretes de lenguaje natural que faciliten el paso de una a otra estructura.

escribano, el objeto de arrendamiento y las condiciones de arrendamiento. La figura 4 nos muestra la estructura de esta comparecencia, que es específica de los arrendamientos. En ella observamos los diferentes papeles que pueden desempeñar los comparecientes en la escritura. La figura 5 nos muestra la estructura de un compareciente. Como podemos observar este puede comparecer en su propio nombre y/o en representación de una persona física o jurídica. Cada uno de estos elementos contiene la descripción de una persona física: hombre o mujer, así como de personas jurídicas como mayorazgos, capellanías, u otras que también han sido caracterizadas, de forma que podemos afirmar **que cada palabra del documento queda encajada en un elemento de la estructura**. Aquellas descripciones que por su naturaleza no puedan ser de momento estructuradas, quedan siempre recogidas en elementos de descripción en lenguaje natural. La figura 6 nos muestra cómo quedaría representado parte del texto del contrato de arrendamiento que expusimos anteriormente.

4.3 Discusión y conclusiones

```
<xsd:complexType name="TipoArBrr">
  <xsd:annotation>
    <xsd:documentation>Arrendamiento de bienes raíces rústicos</xsd:documentation>
  </xsd:annotation>
  <xsd:sequence>
    <xsd:element name="comparecencia" type="not:TipoCompAr"/>
    <xsd:element name="objeto_arrendamiento" type="not:DesBienesRústicos"/>
    <xsd:element name="condiciones_arrendamiento" type="not:CondicionesArrendamiento"/>
  </xsd:sequence>
</xsd:complexType>
```

Figura 2 Representación de la estructura de un arrendamiento

Una de las principales dificultades en el desarrollo de estas ontologías es la garantía de completitud. El dominio de la historia social es tan amplio y variado que las ontologías obtenidas han de contemplarse dentro de una perspectiva pragmática útil para la resolución del problema planteado. No obstante, a pesar de esta falta de completitud las ontologías han demostrado ser útiles en la resolución de los problemas para los que fueron creadas.

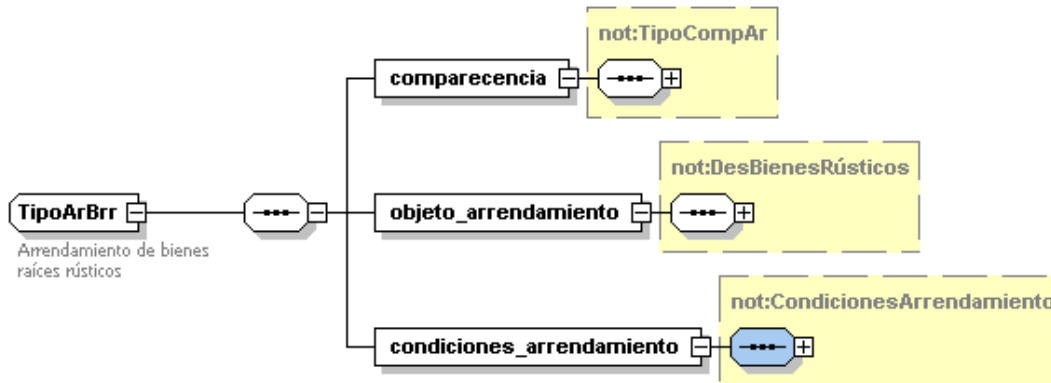


Figura 3 Representación de un arrendamiento

Se dispone hoy día de metodologías y herramientas informáticas que permiten plantear la posibilidad de organizar las palabras que encierra un documento escrito en unas estructuras de las que después puedan derivarse inferencias de interés para la investigación histórica. Los primeros intentos de llevar esta idea a cabo mediante bases de datos relacionales dejaron ver las dificultades que ello encerraba, fundamentalmente por la amplísima variedad y complejidad de los diferentes tipos de documentos. Además con las bases de datos relacionales la información queda dispersa entre las distintas tablas y pierde semántica, o queda ésta encubierta por la complejidad de relaciones que hay que establecer. El uso de esquemas y la utilización de XML parece ser una prometedora vía para la representación de este tipo de documentos. Si a esto le unimos el uso de ontologías más ricas, que recojan aquella semántica que no haya podido recogerse en las especificaciones del esquema, tendríamos así organizada la información en un conjunto de elementos estructurados, que al implementarlos sobre un lenguaje de programación podrían ser tratados como objetos a los que podrían asociarles propiedades y métodos. Todo ello favorecería la búsqueda de información e inferencias de otras nuevas.

Quedan aún por valorar las posibilidades de los nuevos mecanismos de recuperación y filtrado de información en estos sistemas (XSLT), así como su operatividad cuando se enfrenten a estructuras de documentos complejas o recursivas. En nuestro estudio el programa utilizado XML Spy 3.5 presenta ciertas limitaciones de tiempo de respuesta cuando se usan esquemas en vez de DTD para especificar la estructura de los documentos. A pesar de estas dificultades, valoramos positivamente esta herramienta, ya que permite una fácil y rápida modelización del sistema de información.

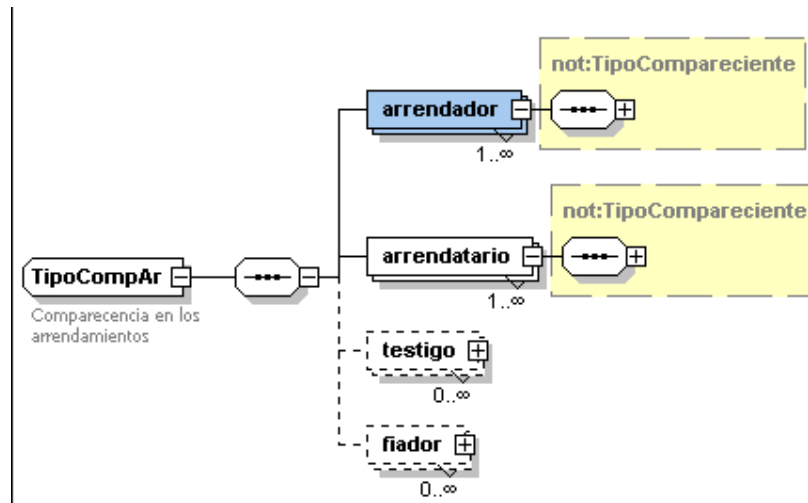


Figura 4 La comparecencia en la escritura de arrendamiento

Nuestra futura investigación en esta línea estará, además de depurar y ampliar las ontologías desarrolladas, en valorar este mecanismo de representación en XML frente a las bases de datos orientadas a objetos, crear herramientas para facilitar la transcripción y entrada de datos de los documentos e incluir métodos para la interpretación restringida del lenguaje natural que faciliten, así mismo, la transcripción. En síntesis, abordar la representación de documentos escritos en estructuras organizadas de símbolos fáciles de crear y a partir de las cuales puedan realizarse de forma automática inferencias de interés en las ciencias históricas.

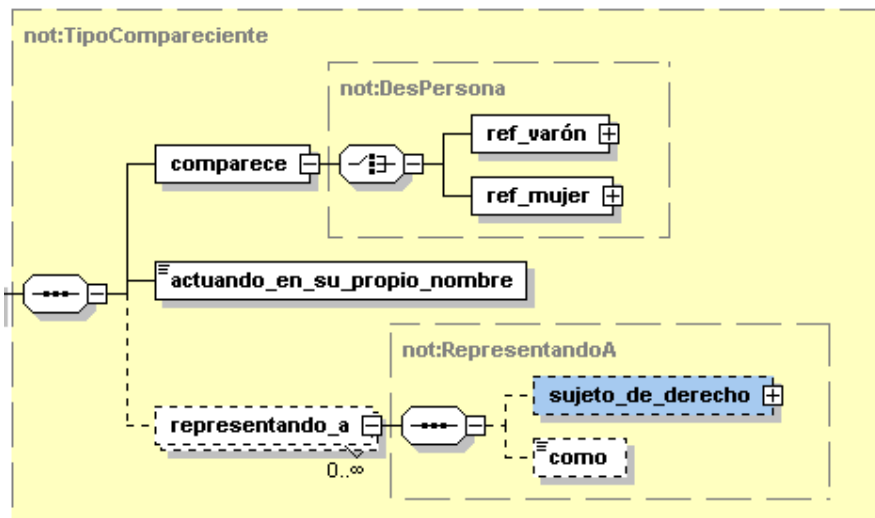


Figura 5 Estructura de un compareciente ante un escribano

```

.
.
<arrendador>
<comparece>
<ref_varon>
<descripcion/>
<datos_personales>
<idref>a124</idref>
<clave_personal>Don Juan Álvarez de Sotomayor y Angulo</clave_personal>
<tratamiento>Don</tratamiento>
<nombre>Juan</nombre>
<apellidos>Álvarez de Sotomayor y Angulo</apellidos>
</datos_personales>
</ref_varon>
</comparece>
<actuando_en_su_propio_nombre>true</actuando_en_su_propio_nombre>
<representando_a>
<sujeito_de_derecho>
<persona_juridica>
<mayorazgo>
<descripcion/>
<fundador>
<ref_varon>
.
.
.

```

Figura 6 La representación del documento en XML

BIBLIOGRAFÍA

- [1] ARÓSTEGUI, J. *La investigación histórica: Teoría y Método*. Crítica, Barcelona, 1995.
- [2] CALVO, A., SERRANO, M. A., ROMERO, J. A. La Ingeniería del Conocimiento en la Investigación de la Historia. Seminario Internacional fuentes y métodos para la Historia Rural (ss. XVIII-XX). *Contar, Representar, Interpretar. Bases de datos y análisis Histórico*. Córdoba, 8-11, 2000.
- [3] DE MIGUEL, A., PIATINI, M. *Concepción Y Diseño de Bases de Datos. Del Modelo E/R al Modelo Relacional*. Rama, Madrid, 1993.
- [4] DAVID C. FALLSIDE (eds.), *Xml schema part 0: Primer w3c candidate recommendation*. <http://www.w3.org/TR/xmlschema-0/>, October 2000.
- [5] DERROSE, S. XML and the TEI. In MILONAS, E. y RENEAR, A. (Eds.), *Text Encoding Initiative: Anniversary conference; 10th*. November 1997, Providence, RIV (Norwell, MA, USA, and Dordrecht, The Netherlands, 1999),

vol. 33 of COMPUTERS AND THE HUMANITIES 1999; 2, Kluwer Academic Publishers Group, pp. 11-30.

[6] FARQUHAR, A., FIKES, A., R., RICE. The ontolingua server: A tool for collaborative ontology construction. *International Journal of Human-Computer Studies*, 1997, vol. 46, nº 6, pp. 707-727.

[7] GOLDFARDB, C. F., PRESCOD, P., *Manual de XML*. Prentice-Hall, Madrid, 1999.

[8] GÓMEZ-PÉREZ, A. ET AL. Towards a Method to Conceptualize Domain Ontologies. *Workshop on Ontological Engineering*. ECAI'96. 1996. Pags. 41-51

[9] GRUBER, T. R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In GUARINO, N., POLI, R., (Eds.), *Formal Ontology in Conceptual Analysis and Knowledge Representation V*. Kluwer Academic Publishers. Deventer, The Netherlands. 1993.

[10] IDE, N., MCGRAW, T., WELTY, C. Representing TEI documents in the classic knowledge representation system. In *Proceedings of the Tenth workshop of the Text-Encoding Initiative V*, 1997, pp. 26-29.

[11] LENAT, D. B., GUHA, R. V. *Building Large knowledge-Based Systems*. Addison-Wesley, 1990.

[12] PORRES, R. ET AL. *Aproximación metodológica a los protocolos notariales de Alava* (Edad Moderna). Universidad del País Vasco, Zaráuz, 1996.

[13] ROEL, A. E. ET AL. La Historia Social de Galicia en sus fuentes de Protocolos. In *Tipología documental de los protocolos gallegos*. Universidad de Santiago de Compostela, Santiago de Compostela, 1981, pp. 11-37.

[14] SCHANK, R. C., ABELSON, R. P. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ, 1977.

[15] SCHREIBER, G. ET AL. *Knowledge Engineering and Management. The CommonKADS Methodology*. The MIT Press, Cambridge, 1999.

[16] SCHREIBER G., ET AL. CML: The commonKADS conceptual modelling language. In STEELS, L., SCHREIBER, G. Y VAN DER VELDE, W. (Eds.), *Proceedings of the 8th European Knowledge Acquisition Workshop : A Future for Knowledge Acquisition V* (Berlin, Sept. 1994), , vol. 867 of LNAI, Springer, pp. 1-25.