Explotación dinámica de relaciones en las bibliotecas digitales: aplicación a una biblioteca jurídica

M. Mercedes Martínez[1], Pablo de la Fuente[1], J. C. Derniame[2], Alberto Pedrero[3]

[1] Universidad de Valladolid, Edificio TIT, Campus "Miguel Delibes"s/n, 47011 Valladolid (Spain)
{mercedes,pfuente}@infor.uva.es
[2] LORIA, BP 239
54506 Vandoeuvre Cedex, France
derniame@loria.fr
[3] Universidad Pontificia de Salamanca, Salamanca (Spain)
apedrero@upsa.es

Resumen Las relaciones son una importante fuente de información, poco aprovechada aún en las bibliotecas digitales operativas. Uno de los modos más extendidos de explotar estas relaciones es la creación de hipertexto, por el cual los usuarios "navegan" entre documentos relacionados. Si bien ésta es una valiosa funcionalidad, no es la única posibilidad para aprovechar la información que aportan las relaciones. En este trabajo se propone expandir dicho aprovechamiento a los dos casos siguientes: consultas sobre las relaciones, y generación dinámica de documentos virtuales. Desde el punto de vista teórico, un grafo de relaciones nos permite modelar éstas y aplicar tratamientos capaces de extraer nueva información, que utilizamos en la construcción de nuevos documentos. La propuesta se completa con la implementación de dicho grafo mediante una base de enlaces XML (*xlinks*), donde se aprovecha la potencia de los estándares *XLink*, *XPointer* y *XPath* para obtener la máxima precisión, en la representación del grafo y la semántica de las relaciones consideradas. La aplicabilidad de la propuesta se ilustra sobre una biblioteca de textos legislativos.

Palabras Clave: Biblioteca digital, grafo de relaciones, documentos estructurados, explotación de relaciones, consulta de relaciones, documentos virtuales, generación dinámica de documentos virtuales, XLink, XPointer, XPath.

1 Introducción

Los enlaces son la base del hipertexto que soporta la popular navegación en la Web, y, por tanto, la forma más expandida de modelar las relaciones. Sin embargo, éstas son una importante fuente de información, capaz de aportar más

funcionalidades a las bibliotecas digitales. Por ejemplo, es lógico suponer que los usuarios de estos sistemas se interesan naturalmente por las relaciones en si mismas (preguntas cómo '*Dónde se ha referenciado mi último artículo?*' son un ejemplo). También cabe construir nuevos items de información (nuevos catálogos, documentos virtuales, etc.) a partir de estas relaciones.

La propuesta expuesta en este trabajo se ocupa de modelar los documentos y relaciones, de modo que se facilite la integración de tres tipos de explotación de estas últimas: **consultas** sobre las relaciones, creación de **hipertexto** para la navegación, y **generación dinámica** de nuevos documentos. La propuesta se plantea en dos niveles de abstracción. En el nivel más alto se utiliza para modelar la información sobre las relaciones un grafo, lo cual nos aporta un modelo formal sobre el cual diseñar las estrategias de explotación de relaciones. En un nivel de abstracción más bajo, la propuesta se ocupa de la factibilidad de la implementación de las soluciones anteriores: la información del grafo se almacena en una base de enlaces XML (*xlinks*). Ilustramos la conveniencia de nuestras propuestas sobre una biblioteca de textos legislativos, en la cual explotamos las relaciones derivadas de *referencias* entre documentos. La sección 3 comenta los aspectos relacionados con el modelado y representación de la información (*estática*), mientras que la sección 4 se ocupa de los aspectos *dinámicos* de la propuesta (tratamientos que explotan las relaciones).

2 Las relaciones entre documentos en las bibliotecas digitales y los documentos estructurados

2.1 Relaciones y su representación

Las relaciones en las cuales están implicados los documentos de las bibliotecas digitales pueden ser de varios tipos: relaciones *semánticas* (mismo autor, tema, ...), *referenciales* (relación debida a la existencia de *referencias* en un documento a otro), o se puede tratar de enlaces *explícitos* (tal es el caso de los enlaces HTML) [1][1].

Las relaciones se pueden representar mediante un grafo *hipertexto*, utilizado para permitir a los usuarios navegar entre los documentos relacionados a través de los hiperenlaces [8]. Otra opción para modelar estas relaciones consiste en la utilización de *metadatos*, como *Dublin Core* [6]. El entorno jurídico es un buen ejemplo de utilización de hipertexto para modelar las relaciones entre documentos como hiperenlaces [2,7,11,12].

2.2 Documentos estructurados y relaciones entre *fragmentos* de documentos

Existe un nivel adicional de complejidad en las relaciones: se puede tener en cuenta que los recursos relacionados no son siempre documentos completos, sino

[1] A su vez las relaciones *semánticas* y *referenciales* pueden considerarse *implícitas*, frente a las relaciones *explícitas* debidas a la existencia de marcas que explicitan la existencia de un enlace (ver [1]).

que en muchos casos se trata de *fragmentos* de documentos. Localizar estos fragmentos se simplifica cuando se trabaja con documentos estructurados.

Los documentos *estructurados* [3] son aquellos que tienen una *estructura lógica*, que describe el documento como una composición de piezas o *fragmentos* que albergan otras piezas, las cuales a su vez pueden albergar más piezas. Esta estructura lógica se puede representar usando un modelo *arborescente* [10]: se representa un documento por un árbol tal que los nodos internos representan los fragmentos (elementos) y el texto (contenido) se encuentra en las hojas. El uso de etiquetas, entremezcladas con el contenido del documento, para marcar el principio y final de cada uno de estos elementos es un modo de almacenar la información sobre esta estructura. De entre los lenguajes de etiquetado disponibles, el estándar XML se viene imponiendo gracias a su extensibilidad y legilibilidad.

3 El grafo de relaciones y su obtención

En esta sección nos centramos en la obtención del grafo de relaciones que proponemos utilizar. Su construcción se hace a partir de las *referencias* detectadas en los textos de los documentos. La aparición de referencias en los documentos a otros documentos es frecuente; por ejemplo, los documentos científicos referencian otros documentos donde se tratan aspectos similares. Un tratamiento completamente automático de las relaciones (desde su detección hasta su explotación) requiere, en lo que a obtención del grafo se refiere:

- **Detectar las relaciones**. Esta etapa se lleva a cabo mediante un análisis de los documentos, en la cual se reconocen las referencias entre documentos.
- Ser capaces de **obtener un grafo** equivalente, cuyos nodos se correspondan con los fragmentos de texto referenciados (o que referencian). Para ello, incluimos una etapa de *normalización* de documentos, que nos aporta los que serán nodos del grafo de relaciones.
- Herramientas que permitan **localizar** los fragmentos de información que nos interesan y almacenar la información sobre el grafo de relaciones en un soporte digital. Utilizamos para esto los lenguajes *XLink*, *XPointer* y *XPath*.

3.1 Documentos estructurados y grafos: normalización de documentos

Algunos documentos tienen una estructura implícita muy bien definida. Por ejemplo, un *texto normativo* está compuesto por elementos de tipo *título*, *capítulo*, ..., *artículo*, existiendo además unas reglas de inclusión que garantizan que un *título* nunca aparecerá en el interior de un *capítulo* o *artículo*, etc. Esta estructura resulta útil a los profesionales que los manipulan y se explota en las referencias a otros textos.

Disponer de una copia digital de los documentos en la cual se refleje la estructura semántica utilizada por los usuarios supone una ventaja cara a la manipulación de documentos [5,9]. En nuestro caso, además, encontramos una ventaja

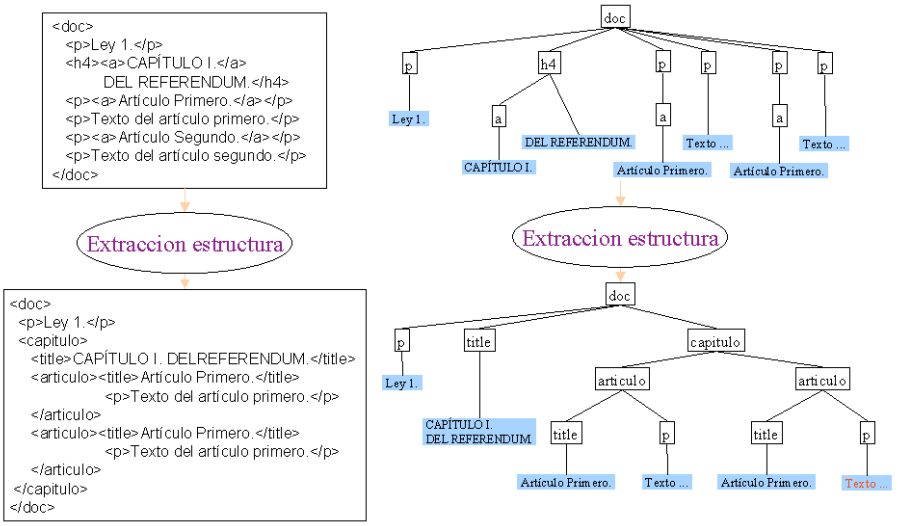

Figura 1. Ejemplo de normalización de documentos.

adicional: reconocer, representar y manipular las relaciones derivadas de las *referencias* entre documentos son tareas cuya complejidad (y factibilidad) depende del modelo usado para representar los documentos implicados en las referencias. Si este modelo es lo más cercano posible al que utilizan los usuarios que "crean" dichas referencias, los tratamientos automáticos que las manipulan se ven facilitados.

Un proceso de **normalización** nos permite conseguir este objetivo. En caso de que el documento que se integra a la biblioteca no tenga la estructura lógica deseada, aplicamos una transformación en la cual obtenemos una nueva copia cuyo contenido (texto) es el mismo, pero cuya estructura lógica refleja la estructura semántica utilizada por los usuarios. Los documentos resultado son XML. Los árboles asociados a estos documentos normalizados constituyen un grafo *estructural* (en él se encuentra representada la *estructura* de los documentos), al cual incorporamos posteriormente las relaciones semánticas detectadas a partir de las referencias en los textos.

La figura 1 muestra un ejemplo (simplificación, para aportar claridad a la figura, de un caso real), donde se ha aplicado este proceso de normalización. A la izquierda se pueden ver el documento de entrada al proceso de transformación (superior) y el resultado de este proceso (inferior): un documento con el

mismo contenido pero distinta estructura lógica. A la derecha se encuentran los árboles correspondientes a sendos documentos, cuya comparación evidencia que la estructura lógica ha cambiado completamente.

La normalización se basa en un análisis del *contenido* del documento. Se tiene en cuenta que la estructura semántica que pretendemos obtener en la salida se encuentra implícita en los textos: ciertas expresiones y vocabulario indican el comienzo de cada fragmento. Se aprovecha también el conocimiento acerca de la DTD a la cual debe ajustarse el documento de salida. La combinación de ambos factores permite normalizar los documentos que posteriormente se manipularán. Más detalles sobre este proceso se pueden encontrar en [13].

3.2 Referencias, relaciones semánticas y grafos de relaciones

Las referencias establecen una relación entre el texto que referencia y el que es referenciado, que se puede representar por un arco que une los dos recursos relacionados (vértices). Sin embargo, es posible extraer más información de una referencia. En algunos casos, la referencia se realiza para establecer algún otro tipo de relación adicional. Un ejemplo muy frecuente se da en los textos legislativos: se referencia un determinado *artículo* o fragmento de una ley, para indicar posteriorme cómo modificarlo; se establecen así dos relaciones con semántica diferente: la *referencia* y la *modificación* asociada.

Vértices del grafo de relaciones y su localización La referencia a un determinado elemento dentro de un documento implica que dicho elemento y todos los incluidos en él están afectados por la relación (por ejemplo, una referencia a la primera sección de este trabajo, concierne a todos los párrafos, figuras, etc. incluidos en dicha sección).

En el caso de los documentos estructurados, tanto la porción de documento que contiene la referencia, como los fragmentos referenciados, son localizables por su posición relativa en la estructura lógica del documento (camino desde la raíz del árbol hasta el nodo correspondiente).

Relaciones heterogéneas = enlaces con tipo Si se incluyen en el grafo estructural los arcos que representan las relaciones de *referencia* y *modificación*, se obtiene un grafo cuyos arcos están etiquetados en función de la semántica de la relación (*estructural*, *referencia*, *modificación*).

Cardinalidad Los vértices del grafo pueden ser a la vez origen y destino de varios enlaces. Es normal que un documento participe en varias relaciones heterogéneas (bien porque contiene referencias a varios documentos, bien porque es referenciado desde varios puntos, o porque se dan ambas circunstancias) y lo mismo puede decirse de los elementos de los documentos (una determinada porción de documento puede ser referenciada desde varios puntos). Por ejemplo, los artículos de la *Constitución* están referenciados en múltiples documentos.

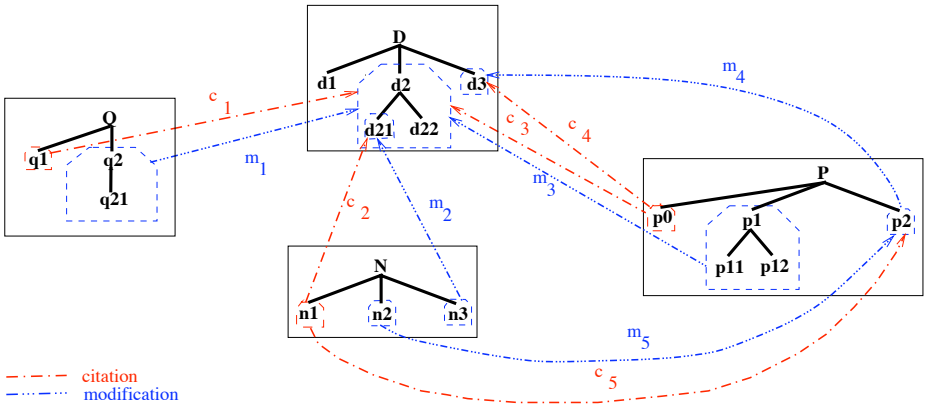

Figura 2. Grafo de relaciones con enlaces heterogéneos: *estructurales*, de *referencia*, y de *modificación*.

3.3 Un ejemplo

El ejemplo de la figura 2 muestra un grafo de relaciones, donde participan cuatro documentos, con raíz respectivamente en D, Q, N and P. Los árboles de la figura representan la estructura lógica de estos documentos. Cada nodo se corresponde con un elemento XML; el texto (que debería aparecer en las hojas) no se ha representado, para que la figura sea más clara. En este grafo encontramos tres tipos de relaciones:

1. Las relaciones **estructurales** entre los elementos. Son las relaciones de *inclusión* entre elementos. En la figura están representadas mediante líneas gruesas.
2. Las **referencias** a (porciones de) D . Se han representado mediante líneas punteadas, y etiquetado c_i. Por ejemplo, desde el elemento q_1 del documento Q se referencia el elemento d_2 de D; en el grafo esta relación está representada por el arco etiquetado c_1.
3. Las relaciones de **modificación** que indican cómo modificar el documento D. Los arcos correspondientes (punto-punto-punto-raya) están etiquetados m_i. Cada uno de estos arcos indica que el fragmento de texto representado por el nodo destino del arco (el elemento de D alcanzado por el arco) debe ser reemplazado por aquel fragmento de texto representado por el nodo origen del enlace.

3.4 XLink, XPath, XPointer

Los estándares asociados a XML, *XLink* [17], *XPointer* [15], *XPath* [14], proporcionan las herramientas necesarias para direccionar los fragmentos de los documentos y modelar enlaces tan complejos como se desee. La asociación de XLink

```
<ENLACE>
<ORIGEN
   xlink:href="l13-1986.xml#xpointer(descendant::disposicion[1]/articulo[1])"
   xlink:role="substitution"      date= "1981" doctype= "norma"   />
<DESTINO
   xlink:href= "lo2-1980.xml#xpointer(child::articulo[1])"
   xlink:role="target"            date= "1986" doctype= "norma"   />
<ARCO    xlink:from="substitution" xlink:to="target"
         xlink:show="undefined"    xlink:actuate="undefined"/>
</ENLACE>
```

Figura 3. Ejemplo de xlink. El enlace expresa una relación de modificación (*sustitución*).

y XPointer permite obtener enlaces mucho más potentes que los tradicionales enlaces HTML: se puede asociar semántica a los enlaces (decir *porqué* existen estos enlaces) crear enlaces multidireccionales (los enlaces HTML son unidireccionales obligatoriamente) asociar múltiples recursos en un mismo enlace, y crear bases de enlaces independientes de los documentos (los enlaces HTML siempre se entremezclan con el contenido del documento).

Aprovechando estas propiedades, hemos creado una base de enlaces (*extended*) XLink (xlinks) que nos aporta las siguientes ventajas:

- Las modificaciones a los enlaces no afectan en modo alguno a los documentos. No es necesario tener permiso de escritura en los documentos relacionados, para crear nuevos enlaces o modificar los que ya existen.
- Los enlaces están caracterizados por atributos que modelan la semántica que nos interesa. Así, les asociamos un *tipo* que indica el tipo de relación, y otros atributos, como la *fecha* del enlace, útiles para consultas y tratamientos posteriores.
- Por último, disponer de una base de xlinks, supone disponer de un conjunto de datos XML, lo cual aprovechamos en la funcionalidad presentada en el apartado 4.1.

El ejemplo de la figura 3 es un xlink que contiene información sobre un nodo *origen*, *destino* y el *arco* que los liga. La semántica de la relación se modela en los atributos de los elementos del *xlink*. El primer *artículo* dentro de la primera *disposición* del documento *l13-1986.xml* sustituye al primer *artículo* de *lo2-1980.xml*.

4 Explotación de las relaciones: consultas y composición dinámica de documentos

En este apartado tratamos tres modos de explotar las relaciones: las consultas que conciernen las relaciones (incluida su semántica) la creación de hipertexto navegacional y la composición dinámica (inteligente[2]) de nuevos documentos.

4.1 Consultas sobre las relaciones

Las consultas sobre relaciones son aquellas en las que se busca información sobre ellas mismas. Preguntas como '*Cuál es la jurisprudencia que se basa en esta ley?*', o '*Cuáles son las leyes referenciadas en esta jurisprudencia?*', son ejemplos donde el interés se centra en la existencia (o no) de relaciones.

Las respuestas a preguntas como éstas se pueden obtener mediante búsquedas en una base de datos XML (los *xlink* son datos XML) basada en una confrontación de valores en los atributos de los enlaces.

4.2 Hipertexto navegacional

La creación de hipertexto navegacional consiste en la obtención de un grafo hipertexto, utilizando para ello el subgrafo de relaciones de referencia obtenido del grafo de relaciones. Utilizamos las relaciones de referencia, ya que son las que están en el origen de cualquier otra relación. Dada la situación actual (no existen navegadores capaces de soportar XLink y XPointer[3]) la implementación pasa por la transformación de los documentos XML en equivalentes HTML en el momento en que son enviados por el servidor a los navegadores. Esta transformación es un proceso que se lleva a cabo mediante la utilización de hojas de estilo XSLT (*XSL Transformations*) [16].

4.3 Composición dinámica de documentos

El último tipo de explotación del que nos ocupamos es la composición dinámica de nuevos items de información, que ilustramos con la composición de documentos virtuales (concretamente, versiones históricas).

Una nueva versión histórica de un documento se obtiene aplicando a la versión inicial de ese mismo documento un conjunto de modificaciones, las cuales se encuentran expresadas a su vez en el contenido de otros documentos. Un ejemplo puede verse en la figura 4. El documento de la esquina superior izquierda (*Real Decreto 685/1982*) es objeto de varias modificaciones, cuya redacción se incluye en el documento situado en la zona inferior de la figura (*Real Decreto 775/1997*).

[2] Se entiende aquí por *inteligente* la capacidad del algoritmo para *deducir* automáticamente las reglas de composición de los nuevos documentos creados.

[3] Tanto XLink como XPointer aún no son recomendaciones estables del W3C en el momento de la redacción de este trabajo (junio 2001) lo cual hace suponer que su soporte por parte de los navegadores no llegará hasta que se consiga dicha estabilidad.

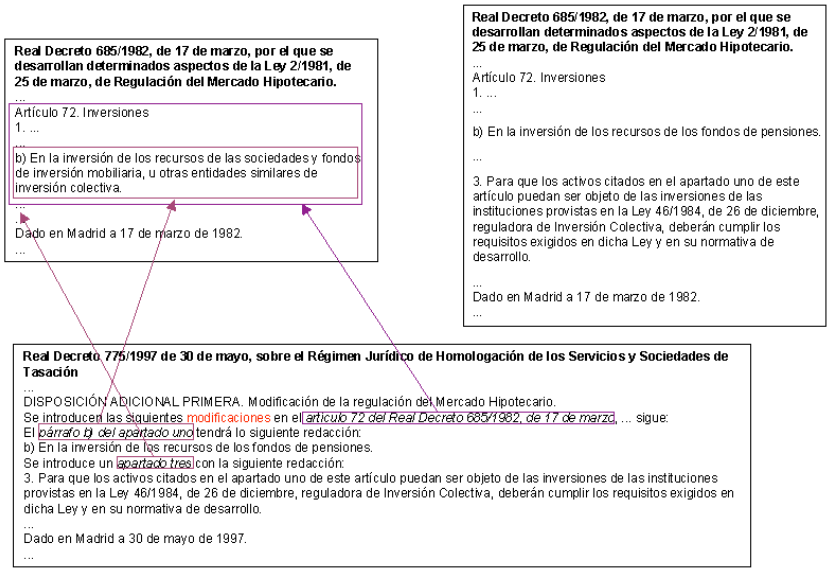

Figura 4. Ejemplo de composición dinámica de documentos legislativos.

Se referencian los fragmentos que se debe modificar y se indica cómo hacerlo. La aplicación de estas modificaciones da lugar a una nueva versión del *Real Decreto 685/1982*, cuyo texto se observa en la esquina superior derecha de la figura.

La generación automática que presentamos en este apartado evita a los usuarios verse obligados a realizar una composición manual -recorrer todos los documentos implicados en las modificaciones, cortando y pegando manualmente los fragmentos de texto-, que, en los casos de numerosas modificaciones, puede ser una tarea complicada.

Dicha generación se basa en un recorrido de un subgrafo de *modificaciones* que afectan a la versión inicial del documento, el cual se extrae del grafo de relaciones. En este recorrido se deducen *automáticamente* las reglas de composición de la nueva versión, y se realiza simultáneamente dicha composición. La explicación de este algoritmo, y las soluciones a los posibles conflictos que pueden ocurrir durante la deducción de las reglas de composición se puede encontrar en [13].

5 Trabajos relacionados

El *hipertexto* ha sido aceptado generalmente como un modo ideal de modelar las relaciones [2], mediante hiperenlaces. En el entorno jurídico se pueden encontrar varios ejemplos [11,9]. La solución que proponemos en este trabajo asume dicha utilidad, pero expande la explotación de relaciones a otras funcionalidades, que van más allá de la navegación. Simplificamos así la consulta de relaciones, que pasa de ser un proceso de navegación a una consulta de una base de datos XML. Por otro lado, la generación dinámica de documentos virtuales es algo que únicamente hemos visto propuesto en un caso. Arnold-Moore y colaboradores [4,5] proponen una composición automática de documentos -también en el contexto jurídico-; sin embargo, en ningún momento indican que las reglas de composición en que se apoyan se extraigan a partir de relaciones, lo cual es un punto fundamental de nuestra propuesta.

6 Conclusiones

En este trabajo hemos resaltado la valía de las relaciones como una potente fuente de información, que puede ser utilizada para expandir las funcionalidades de las bibliotecas digitales. Hemos elevado los enlaces que representan las relaciones a la categoría que, en nuestra opinión, los portadores de tan valiosa información merecen: items de primer nivel en la biblioteca. El modelo propuesto (grafo de relaciones) es flexible en su potencial aprovechamiento, lo cual se ha demostrado sobre tres explotaciones diferentes; además, es aplicable a cualquier tipo de relación, sin verse limitado a las relaciones referenciales consideradas en este trabajo. Hemos puesto de relieve las ventajas de disponer de una base de *xlinks* independiente de los documentos (*extended, out-of-line*) donde se aprovecha la potencia de los lenguajes *XLink*, *XPointer* y *XPath*. La consulta de relaciones se beneficia de la potencialidad de los lenguajes XML, sobre todo teniendo en cuenta los avances que cabe esperar en lo que a consulta de datos XML se refiere, una vez que existe un lenguaje de consulta estándar [18]. La generación dinámica de documentos virtuales en base a relaciones, deduciendo automáticamente las reglas de composición de estos documentos, es original, y esperamos expandir su aplicación en breve a otros tipos de documentos virtuales.

Agradecimientos

Los autores de este trabajo agradecen al Dr. Jordi Barrat i Esteve, del área de Derecho Constitucional de la Universidad de León, su colaboración en los aspectos relacionados con el modelado de la información jurídica.
Este trabajo se ha llevado a cabo gracias a la financiación del proyecto CICYT TEL99-0335-C04.

Referencias

1. Agosti, M., and Allan, J. Methods and tools for the construction of hypertext. *Information Processing and Management 33*, 2 (1997), 129–271.

2. Agosti, M., Colotti, R., and Gradenigo, G. A two-level hypertext retrieval model for legal data. In *14th ACM-SIGIR International Conference on Research and Development in Information Retrieval* (Dipartamento di Elettronica e Informatica, Universita'di Padova, Oct. 1991), Chicago, IL USA, pp. 316–325.
3. André, J., Furuta, R., and Quint, V. Structured documents: What and why? In *Structured Documents* (1989), J. André, R. Furuta, and V. Quint, Eds., Cambridge University Press.
4. Arnold-Moore, T. Automatic Generation of Amendment Legislation. In *Sixth International Conference on Artificial Intelligence and Law, ICAIL'97* (Melbourne, Vicoria, Australia, 1997), ACM, pp. 56–62.
5. Arnold-Moore, T., Fuller, M., Kent, A., Sacks-Davis, R., and Sharman, N. Architecture of a content management server for XML document applications. In *1st International Conference on Web Information Systems Engineering (WISE 2000)* (Hong Kong, June 2000).
6. Biagioni, S., Carlesi, C., and Castelli, D. Supporting retrieval by 'relation amog documents' in the ERCIM Technical Reference Digital Library. In *11th ERCIM Database Research Group Workshop on Metadata for Web Databases* (May 1998).
7. Choquette, M., Poulin, D., and Bratley, P. Compiling Legal Hypertexts. In *Database and Expert Systems Applications, 6th International Conference, DEXA'95* (Sept. 1995), N. Revell and A. M. Tjoa, Eds., vol. 978 of *Lecture Notes in Computer Science*, Springer, pp. 449–458.
8. Conklin, J. Hypertext: An introduction and survey. *IEEE Computer 20*, 9 (1987), 17–41.
9. Finke, N. TEI Extensions for Legal Text. In *Text Encoding Initiative Tenth Anniversary User Conference* (Providence, Rhode Island, USA, Nov. 1997).
10. Furuta, R. Concepts and models for structured documents. In *Structured Documents* (1989), J. André, R. Furuta, and V. Quint, Eds., Cambridge University Press, pp. 7–38.
11. Haider, G., Sjöberg, C. M., Quirchmay, G., and Sebald, V. The Comparative Part of the Corpus Legis Project - Using SGML for Intelligent Information Retrieval of Legal Documents. In *EXPERSYS-96, Artificial Intelligence Applications.* (1996), A.Ñiku-Lari., Ed., Technology Transfer Series, pp. 181–186.
12. Noticias jurídicas. http://noticias.juridicas.com/.
13. María Mercedes Martínez González. *Principios para la explotación dinámica de relaciones entre documentos en las bibliotecas digitales: aplicación al entorno jurídico.* PhD thesis, Dpto. de Informática (U. Valladolid, Spain), abril 2001.
14. W3C, the World Wide Web Consortium. *XML Path Language (XPath)*, Nov. 1999. W3C Recommendation. http://www.w3.org/TR/1999/xpath.
15. W3C, the World Wide Web Consortium. *XML Pointer Language (XPointer)*, Dec. 1999. W3C Working Draft. http://www.w3.org/TR/xptr.
16. W3C, the World Wide Web Consortium. *XSL Transformations (XSLT)*, Nov. 1999. W3C Recommendation. http://www.w3.org/TR/1999/xslt.
17. W3C, the World Wide Web Consortium. *XML Linking Language (XLink)*, Feb. 2001. W3C Working Draft 21-February-2000. http://www.w3.org/TR/2000/WD-xlink-20000221.
18. W3C, the World Wide Web Consortium. *XQuery: A Query Language for XML*, Feb. 2001. W3C Working Draft. http://www.w3.org/TR/xquery.