Una propuesta integrada de extracción de información para gobierno electrónico: estructura, referencias y evolución de los documentos jurídicos

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

Resumen Los documentos jurídicos son el soporte fundamental de las aplicaciones de gobierno electrónico. La mayoría de los procesos de recuperación de información se realizan sobre su contenido. No obstante, algunas de sus características los hacen potenciales proveedores de información adicional. La estructura de documentos como leyes, decretos, sentencias -creada en base a criterios semánticos-, las referencias entre documentos y la propia evolución de éstos (que afecta a contenido, estructura y referencias) son información valiosa que debe ser extraída de los propios documentos.
En este trabajo se presenta una propuesta cuyo objetivo es automatizar los procesos de extracción de información de documentos legislativos. Las técnicas de extracción de información se combinan para extraer la estructura de los documentos a partir de su contenido, para extraer referencias (relaciones), y se enfoca la consolidación (actualización) de textos normativos como un proceso de extracción de conocimiento a partir de la información extraida en los procesos anteriores (estructura y referencias). La evaluación (metodología y resultados) se ha hecho sobre un conjunto de textos normativos españoles. La utilización de XML en la implementación ha potenciado las posibilidades de extensión de lo desarrollado con otras posibilidades.

1. Introducción

La extracción de información consiste en extraer información específica de dominio a partir de los textos [8]. La extracción de patrones del contenido de los documentos es el modo básico de comenzar la extracción [22]. Las referencias en los documentos jurídicos se ajustan a patrones regulares [26]. Estas referencias son información valiosa, que se puede explotar de múltiples modos (búsquedas, navegación u otros). Pero además pueden ser la base sobre la que obtener información adicional, por ejemplo, la información acerca de las modificaciones entre los documentos jurídicos se encuentra en el contenido de los propios textos, asociada a referencias.

Aún existen pocos sistemas que proporcionen acceso exhaustivo a las referencias. Una de las posibles causas es la limitación que supone el esfuerzo humano que requiere realizar este trabajo manualmente. La solución parece claramente una automatización de la extracción de referencias; los procesos automáticos son, sin lugar a dudas, más baratos que los procesos manuales. Otra posible limitación es la falta de enlaces entre fragmentos de documentos. En los sistemas que utilizan enlaces para representar las referencias,

éstos suelen resolver hacia documentos completos en vez de a los fragmentos específicos referenciados en los textos. Thistlewaite demostró en 1997 [25] que la extracción de referencias a partir de contenido era factible en el entorno jurídico. Caracterizaba las referencias mediante patrones, y los almacenaba como *anchors* (anclas) de enlaces. Este trabajo se puede ampliar con la detección de referencias a fragmentos y con la captura de la semántica asociada a la referencia encontrada.

Tras ser procesada mentalmente por una persona, la información se convierte en conocimiento. Que se convierte a su vez en información tras ser transmitido bajo la forma de texto, salida de ordenador, o cualquier otro medio [3]. La extracción de conocimiento aborda varios tipos de problemas en gobierno electrónico [10]. Pero si se relaciona con las referencias, modificaciones y el modo en que se efectúan las consolidaciones (actualización) en el entorno legislativo, se puede afirmar que uno de sus retos es imitar los procesos de consolidación: ¿cómo inferir las reglas de composición de las versiones actualizadas a partir de la información que los legisladores aportan en los textos normativos? Los humanos extraemos la información de los documentos (modificaciones y referencias asociadas), extraemos conocimiento a partir de ella (estructura de las versiones consolidadas) y la usamos (consolidamos). La consolidación automática se ha abordado en pocas ocasiones [4,13]. La relación entre referencias, modificaciones y consolidación aún no está reflejada en estos procesos; la inferencia automática de la estructura de las versiones consolidadas tampoco.

La tesis que fundamenta la propuesta que se presenta en este trabajo es considerar la manipulación de estructura, referencias y la evolución de documentos como modos distintos, pero no aislados, de explotar conjuntamente los documentos y relaciones entre ellos. Las técnicas de extracción de información se combinan para extraer la estructura de los documentos de su contenido, para extraer referencias, y la consolidación (actualización) de textos normativos se resuelve como un proceso de extracción de conocimiento a partir de la información extraida previamente (estructura y referencias). La evaluación (metodología y resultados) reciben la atención principal en este documento. En cada apartado se referencia otros documentos donde se puede encontrar más información sobre las técnicas evaluadas.

La organización del artículo es la que sigue. La sección 2 revisa brevemente las características de los documentos jurídicos que son fundamentales para comprender la propuesta. Las secciones 3, 4 y 5 se dedican a la extracción de estructura, referencias y consolidación respectivamente. El uso de XML en la implementación es el tema de la sección 6. En la sección 7 se revisan otros trabajos relacionados y en la sección 8 se presentan las conclusiones y perspectivas de trabajo futuro.

2. Los documentos jurídicos

En esta sección se revisan brevemente los aspectos de los documentos jurídicos cruciales para comprender las estrategias de extracción de información tratadas en las secciones 3, 4 y 5. Descripciones más detalladas pueden encontrarse en manuales de técnica jurídica, como [11].

2.1. Estructura

La primera característica relevante de los documentos legislativos es su alto nivel de estructuración. La estructura de estos documentos está implícita en su contenido y es por tanto independiente del forma-

to o soporte en que se distribuyen. Los autores (legisladores) estructuran los textos siguiendo criterios semánticos (divisiones en secciones, artículos u otros que contienen la norma aplicable a situaciones determinadas); las divisiones siempre tienen tipo (`sección, capítulo, artículo, ...`) y están ordenadas: un lector puede referirse a ellas por su tipo y número de orden dentro del documento. Además, para cada clase de documento, los tipos de divisiones están estandarizadas.

2.2. Referencias cruzadas

Las referencias entre documentos son muy frecuentes en el entorno legislativo. Aparecen para incluir el texto de normas redactadas en otros documentos, para redirigir al usuario a otros documentos en busca de información adicional o para modificar textos anteriores. En cualquier caso, estas referencias, que deben ser *precisas*, se basan en la estructura mencionada en el apartado anterior.

2.3. Consolidaciones

El modo de consolidar los textos normativos es peculiar y está muy relacionado con las referencias descritas. La modificación de un texto normativo dado aparece redactada en otro posterior, en el cual se referencia el fragmento de documento que se debe modificar y se enuncia la modificación concreta que debe realizarse. Queda como responsabilidad del lector "cortar y pegar" las modificaciones oportunas sobre el texto original a fin de obtener la versión consolidada.

3. Extracción de estructura

La extracción de estructura se realiza a partir del contenido de los documentos. Se aplica a documentos digitales que, en caso de estar estructurados, tienen una estructura que no se corresponde con la estructura implícita en su contenido –documentos marcados con etiquetas no significativas o emplazadas en posiciones que no coinciden con las divisiones semánticas propias de la estructura implícita de los textos normativos–. Las divisiones implícitas se reconocen por la presencia en el texto de expresiones que indican el comienzo de cada división. Estas divisiones tienen una estructura regular y contienen vocabulario (estandarizado) que indica el tipo de la división que comienza. La estructura se extrae durante un análisis del documento. Más detalles del proceso se pueden encontrar en [18].

La implementación y experimentos se distribuyeron en las etapas siguientes:

1. Un *preprocesamiento* de los documentos previo al proceso de extracción. Esta fase consistió en transformar los documentos de entrada que no eran ascii (por ejemplo, documentos Word) a texto plano, unir en un único documento digital los textos normativos que estaban fragmentos en varios documentos electrónicos, eliminar aquello que no servía para el análisis de estructura (notas, definiciones) y crear etiquetado básico para los documentos en texto plano, que marcaba los párrafos[1].
   El resultado de este preprocesamiento (que requiere una buena dosis de revisión manual posterior) es un conjunto de 1665 documentos estructurados[2], cuya estructura lógica (etiquetado) no se corresponde

[1] Herramientas sencillas como *lex*, *yacc* fueron suficientes para esta etapa

[2] XML bien formados, como se verá en la sección 6.

aún con la estructura interna implícita en su contenido. La estructura disponible en estos documentos es la estructura de bajo nivel (párrafos o similares), no estandarizada, que también se explota en las referencias; la estructura de alto nivel (que se ajusta a divisiones estándar) es la que se extrae en la etapa siguiente.

2. La *extracción de estructura* se probó sobre los documentos obtenidos en la primera etapa y funcionó sobre 1583 documentos. Dado que este proceso esta implementado sobre parsers, los errores encontrados se deben en parte a errores que el propio parser encuentra: caracteres y entidades no reconocidas (`&icute;`, `&ordm;`, etc.), o sin un elemento raíz. La validación de los resultados se lleva a cabo en paralelo con la extracción de referencias (ver sección 4). A medida que los documentos obtenidos en esta fase son usados como entrada para la extracción de referencias se detectan posibles errores en los resultados de la extracción de estructura. Por el momento se ha probado la extracción de referencias sobre 50 documentos y, de ellos, se han detectado problemas con 7 documentos. Las conclusiones correspondientes son:

   Las divisiones en los niveles altos de la estructura que no están estandarizadas no son reconocidas correctamente, debido a la carencia de vocabulario que guíe el proceso de extracción. Por ejemplo, en algunos documentos aparecieron divisiones de tipo `norma`, que no son habituales en los textos normativos. También dió problemas el listado secuencial de varias disposiciones que comparten un único encabezamiento `disposiciones`; esta situación es poco común, ya que lo habitual es que cada disposición tenga su propio encabezamiento que incluye su título y número de disposición. Es decir, el anidamiento en la estructura resultante no siempre es tan fino como debería (como en el contenido del documento), dado que lo que deberían ser varias divisiones distintas del mismo tipo son reconocidas como una sola. Una posible solución a este problema sería introducir algún tipo de reconocimiento de formato (separaciones físicas entre párrafos) y de conceptos, de modo que se distinguiese por las variaciones en tema y separadores físicos cuándo un cambio de párrafo indica meramente un nuevo párrafo dentro de la misma disposición o cuándo corresponde al comienzo de una nueva. Este tipo de procesos, basados en factores puramente semánticos –como lo es el reconocimiento de conceptos por tema– son extremadamente difíciles de implementar (de hecho son uno de los grandes retos en áreas como Web semántico). Dado que en el caso de documentos normativos que se trata aquí son *excepciones* parece más razonable optar por una revisión y corrección manual de estos casos aislados que hacer un esfuerzo de ese nivel.

4. Extracción de referencias

La extracción de referencias implementa un conjunto de gramáticas, que a su vez se corresponden con una clasificación de las referencias. La clasificación considera simultáneamente tres criterios, que aparecen reflejados conjuntamente en las referencias encontradas en los textos.

1. **Clasificación por unidad referenciada**

   Se pueden distinguir referencias a *documentos completos* o a *fragmentos* de documentos. Las segundas se construyen añadiendo a una referencia a documento completo aquello que sirve para distinguir un fragmento dentro de dicho documento. La distinción se hace en base a la estructura interna del

documento. La referencia en la primera línea de la tabla 1 es un ejemplo; se indica el documento (*Real Decreto 1348/1985, de 1 de agosto*) y el fragmento de interés (*artículo 66*).

2. **Clasificación según el número de items referenciados**

   Se puede distinguir también entre referencias *simples* y *coordinadas*. Las referencias simples dirigen al lector hacia un único objeto (documento, artículo, ...) mientras que en las coordinadas se establece un vínculo con varios items (varios documentos, artículos, ...). La referencia de la primera fila de la tabla 1 es una referencia simple, mientras que el ejemplo de la segunda fila es una referencia coordinada: se mencionan varios apartados (*a)* y *b)*) del mismo documento (*Constitución*).

3. **Clasificación en función del grado de conocimiento del contexto necesario para resolver la referencia**

   La *resolución* de una referencia (asociación con algún item –documento o fragmento de documento en este caso–) puede depender del grado de conocimiento del contexto que tiene el lector. Se pueden distinguir las referencias *directas* (el redactor incluye toda la información necesaria para resolver la referencia) y las referencias *relativas* (la información incluida en la referencia es parcial: se necesita conocimiento adquirido en el contexto). Los ejemplos de las filas 1 y 2 de la tabla 1 son referencias directas; el ejemplo en la tercera fila es una referencia relativa, donde no se indica expresamente la Ley de la que se trata, ya que se hizo previamente en el documento.

Tipo de referencia	*Ejemplo*
Interna	*artículo 66 del Real Decreto 1348/1985, de 1 de agosto*
Coordinada	*apartados a) y b) de la Constitución*
Relativa	*... a los enumerados en los apartados 3, 4 y 5 de la mencionada Ley*

Cuadro 1. Tipos de referencias en documentos legislativos.

4.1. Evaluación

Los experimentos se han hecho sobre un conjunto de 50 textos normativos españoles[3], seleccionados entre los documentos obtenidos del proceso de extracción de estructura (sección 3). Estos textos son heterogéneos en tipo (leyes, decretos, circulares, ...) y en procedencia (creados por distintos legisladores); se buscó intencionadamente heterogeneidad para cubrir la mayor variedad posible en el modo de expresar las referencias.

Metodología

La extracción y su evaluación se organizaron en las siguientes etapas:

1. En primer lugar, se hicieron dos revisiones manuales consecutivas de los documentos. En la primera se buscaron las referencias presentes en los documentos. En la segunda se verificó el conjunto de referencias, incorporando aquellas que habían "escapado" a la extracción inicial. Después de ambas revisiones el conjunto de referencias presente en los 50 documentos estaba completo y disponible en una colección aislada.

[3] Ver en Conclusiones la razón de esta cantidad.

2. Un examen de las referencias encontradas llevó a la definición de la clasificación presentada anteriormente y de los patrones y gramáticas correspondientes. Los detalles de las gramáticas están disponibles en [21].
3. El tercer paso fue la implementación de la extracción y su puesta en funcionamento sobre los 50 documentos seleccionados. Se generaron como datos de salida: una base de enlaces -que es el objetivo de tal extracción- y la colección de referencias extraidas automáticamente. La utilidad de estas últimas es la validación y evaluación de este proceso: su comparación con las extraidas manualmente en la primera etapa permite obtener las conclusiones comentadas en el apartado **Resultados**.
4. Los enlaces generados se utilizaron en una aplicación para generar hipertexto navegacional. Una captura de pantalla de esta aplicación se muestra en la figura 1.

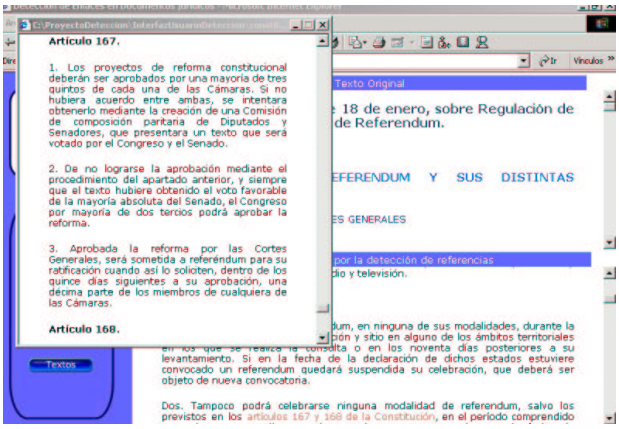

Figura 1. Navegación a través de referencias en un cliente Web.

Resultados

La evaluación se llevó a cabo con referencias hechas por identificador de documento y/o fragmento dentro de documento -las referencias por título o tema son difíciles o imposibles de ajustar a patrones y su reconocimiento queda bajo el ámbito de las aplicaciones de lenguaje natural avanzadas, que no son nuestra área de investigación prioritaria ni el objetivo del trabajo aquí presentado–. Se detectaron 1772 referencias en los 50 documentos, de las cuales 1690 se resolvieron (asociaron con algún documento o fragmento de documento) sin problemas. Dieron problemas expresiones ambiguas. En algunos casos la cadena extraída automáticamente no coincidía con la extraída manualmente (normalmente, la primera era más extensa, esto es, menos precisa, que la segunda). Por ejemplo, la referencia "*el primer artículo y siguientes*", que es ambigua, fue reconocida por la aplicación como una enumeración cuyo segundo elemento se extendía hasta el carácter de puntuación más cercano (un punto). Se conjugaron los dos problemas comentados: la cadena extraída era demasiado extensa y la redacción del legislador confundió a la aplicación, que interpretó como enumeración lo que era un rango.

5. Consolidación

La consolidación consiste en la automatización del proceso de inferencia (extracción) de las reglas de composición (estructura) de las versiones actualizadas. Esta estructura se infiere a partir de la información

que aportan las modificaciones entre documentos. Estas modificaciones, que aparecen siempre asociadas a referencias entre documentos, son relaciones entre ellos, cuya información se almacena en una base de enlaces, al igual que la concerniente a las referencias.

El algoritmo de inferencia está explicado en [19]. Utiliza como entrada la versión inicial del documento que se pretende consolidar, los documentos modificadores (contienen la redacción de la modificación) y la base de enlaces con las relaciones de modificación entre modificadores y documentos modificados.

5.1. Evaluación

En la implementación se hace simultáneamente la inferencia de la estructura y la composición automática de la versión consolidada. Así se puede evaluar los resultados, comparando dos versiones: la inicial, que estaba disponible en las colecciones de la biblioteca, y la generada automáticamente. Si la consolidación es correcta, las diferencias entre ambas son las modificaciones.

De momento se han realizado cinco consolidaciones, con varias modificaciones implicadas en cada una de ellas (4 en el caso menos numeroso y 20 en el que más). La evaluación se ha hecho comparando la consolidación obtenida automáticamente con los documentos disponibles (versión inicial y modificadores). Es decir, para verificar la consolidación automática se realizó una consolidación manual.

Los resultados obtenidos llevan a la siguiente conclusión: la bondad de la consolidación depende directamente de la bondad de los enlaces almacenados. Allí donde había errores en los enlaces (por ej., paths que referenciaban un fragmento distinto del que deberían), la consolidación fue incorrecta (se insertó en la versión consolidada fragmentos de texto inadecuados). Sin embargo, allí donde los enlaces eran correctos, la consolidación obtenida también lo era.

6. XML en la implementación

XML es el lenguaje escogido para representar los documentos jurídicos en este proyecto (ver en Trabajos Relacionados más sobre el uso de XML en este entorno). También se ha utilizado para almacenar los metadatos y otras informaciones relacionadas; la base de enlaces que almacena la información sobre referencias y modificaciones es una base de enlaces XML. La implementación de los procesos de extracción se ha hecho como una aplicación de análisis que trabaja sobre un parser XML cualquiera, usando la interfaz de acceso a documentos SAX; esto es aplicable a la extracción de estructura (sección 3) y de referencias (sección 4). En cuanto a la consolidación se ha implementado simultáneamente la extracción de estructura y la composición automática de la versión consolidada, reduciendo ambos procesos a una hoja de estilo, que puede ser procesada por cualquier procesador estándar de hojas de estilo XML.

Es destacable la importancia de los lenguajes de direccionamiento de fragmentos dentro de documentos (XPath, XFragment) para obtener el grado de precisión necesario a la hora de almacenar la información sobre referencias y modificaciones. Esta precisión es altamente deseable en el caso de referencias e imprescindible en el caso de consolidaciones. Para las referencias, porque permite direccionar a los usuarios (por ejemplo, en la aplicación de navegación mostrada en la figura 1) a la porción de información concreta que les interesa, evitándoles navegar por documentos extensos buscando el artículo, frase u otro que

quieren consultar. Para la consolidación, porque, como se comentó en los resultados de la sección 5, un direccionamento inadecuado de los fragmentos conlleva una consolidación incorrecta.

7. Trabajos relacionados

La extracción de estructura se puede enfocar desde varios puntos de vista: extracción de estructura física [27,24], extracción de la estructura lógica general (DTD en el caso XML) de una clase de documentos [27,24,7], o extracción de la estructura lógica de un documento [16,23]. La extracción de la sección 3 pertenece a esta última categoría. Su diferencia más importante con las restantes soluciones es que el análisis se guía por la aparición de expresiones y vocabulario que indican el comienzo de cada nueva división, mientras que en las otras propuestas se trata de reconocer *conceptos* dentro de los párrafos.

La extracción de referencias ha sido tratada por los investigadores de dos áreas: la creación automática de hipertexto [1,25,7] y *reference linking* [6,14]. Al igual que en nuestro caso, consiste en un análisis del documento durante el cual se reconocen patrones (referencias) y se resuelven (asocian con algún item del universo de trabajo). Las experiencias en la construcción automática de hipertexto se han aplicado también sobre documentos jurídicos, utlizando gramáticas. Nuestro trabajo completa estas propuestas: con una clasificación de referencias más detallada, y con el incremento de la precisión (extracción de referencias a fragmentos de documentos) en su tratamiento.

La consolidación de los documentos normativos preocupa a los investigadores desde hace tiempo [26,15]. No obstante, las propuestas de consolidación automática son recientes [5,13]. Estas propuestas suponen un esfuerzo importante para facilitar la manipulación de la legislación por parte de sus usuarios. En esta línea se encuadra nuestra investigación, que da un paso adelante incorporando la inferencia automática de la estructura de composición de las versiones consolidadas. La extracción de conocimiento no estaba aún incorporada en estos procesos.

La adopción de XML para representar documentos jurídicos es tan general, que se puede afirmar que es "el" lenguaje del gobierno electrónico sin temor a equivocarse. Existen numerosas DTDs desarrolladas expresamente para estos documentos y cada vez lo utilizan más administraciones para sus documentos (por ejemplo, el proyecto europeo Eulegis). Algunos de los trabajos más relevantes en los que se han aprovechado las posibilidades de XML son [5,20,12,17]. Previos a ellos las posibilidades de la norma SGML habían sido exploradas por Agosti [2] y Finke [9]; estos trabajos merecen ser destacados por su carácter pionero y la apertura de posibilidades que supusieron para investigadores posteriores.

8. Conclusiones y trabajo futuro

Se ha presentado una propuesta para la extracción automática de información de documentos jurídicos. Presta especial atención a los documentos y sus relaciones. La relación entre estructura, referencias (relaciones) y consolidación conecta los tres procesos de extracción presentados y da originalidad a la propuesta. Los resultados obtenidos muestran que la relación entre estos tres aspectos es aún más importante de lo que se suponía inicialmente (ver, por ejemplo, la dependencia de la consolidación de la bondad de las referencias detectadas comentada en la sección 5).

En este documento se han resaltado la evaluación y resultados. La consulta sobre las soluciones de diseño y algorítmicas está disponible en los documentos referenciados en las secciones respectivas o en [18].

La implementación y evaluaciones se han hecho sobre una colección de documentos normativos españoles. En aquellos procesos que requieren un alto nivel de intervención humana para la evaluación (extracción manual de referencias y consolidación manual) se ha limitado el número de pruebas, no por restricciones de los procesos automáticos, sino de disponibilidad de personal y tiempo. La dependencia entre consolidación y resultados de la extracción de referencias ha motivado que esta evaluación sea la menos avanzada: se optó por esperar a tener una base de enlaces extensa y robusta resultado de la etapa anterior para proseguir con las consolidaciones.

La evaluación prosigue. Se incorporarán además documentos de la Unión Europea en las pruebas futuras. La mejora de las interfaces de usuario es una de las líneas de actuación previstas para un futuro cercano. La extensión con otras posibilidades de tratamiento de hipertexto es una línea de trabajo cuyo planteamiento y diseño está en curso, si bien su implementación y validación se prevé a medio/largo plazo.

Agradecimientos

La alumna Sandra Muñoz Mínguez implementó la extracción referencias. El Dr. Dámaso Javier Vicente Blanco, del departamento de de Derecho Mercantil, Derecho del Trabajo y Derecho internacional privado guió el análisis de los documentos jurídicos.

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. 316–25.
3. Alavi, M., and Leidner, D. E. Knowledge Management Systems: Issues, Challenges, Benefits. *Communications of AIS 1*, 2 (Feb. 1999), 2–41.
4. Arnold-Moore, T. Automatic Generation of Amendment Legislation. In *Sixth International Conference on Artificial Intelligence and Law, ICAIL'97* (Melbourne, Victoria, Australia, 1997), pp. 56–62.
5. Arnold-Moore, T. Connected to the Law: Tasmanian Legislation Using EnAct. *Journal of Information, Law and Technology 1* (2000).
6. Bergmark, D., and Lagoze, C. An Architecture for Automatic Reference Linking. In *5th European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2001* (Darmstadt, Germany, September 4-9 2001).
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. 449–58.
8. Cowie, J., and Wilks, Y. *Handbook of Natural Language Processing.* Marcel Dekker, New York, 2000, ch. Information Extraction, pp. 241–60.
9. Finke, N. TEI Extensions for Legal Text. In *Text Encoding Initiative Tenth Anniversary User Conference* (Providence, Rhode Island, USA, Nov. 1997).

10. Gottschalk, P. Use of IT for knowledge management in law firms. *The Journal of Information, Law and Technology (JILT) 3* (1999).
11. Grupo de Estudios de Técnica Legislativa. *Curso de técnica legislativa GRETEL.* Serie de Técnica Legislativa I. Centro de Estudios Constitucionales, Madrid, 1989.
12. 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. 181–6.
13. Hemrich, M. A New Face for Each Show: Make Up Your Content by Effective Variants Engineering. In *XML Europe 2002* (2002).
14. Hitchcock, S., Carr, L., Jiao, Z., Bergmark, D., Hall, W., Lagoze, C., and Harnad, S. Developing services for open eprint archives: Globalisation, integration and the impact of links. In *ACM Proceedings of Digital Libraries, 2000 (DL2000)* (San Antonio, Texas, 2000).
15. Leung, R. Versioning on legal applications using hypertext. In *Proceeedings of the Workshop on Versioning in Hypertext Systems. Held in connection with ECHT '94 ACM European Conference on Hypermedia Technology* (Edinburgh, Sept. 1994), pp. 18–23.
16. Lim, S.-J., and Ng, Y.-K. WebView: A Tool for Retrieving Internal Structures and Extracting Information from HTML Documents. In *Sixth International Conference on Database Systems for Advanced Applications (DASFAA)* (Hsinchu, Taiwan, Apr. 1999), IEEE Computer Society, pp. 71–8.
17. Marchetti, A., Megale, F., Seta, E., and Vitali, F. Using XML as a means to access legislative documents: Italian and foreign experiences. *ACM SIGAPP Applied Computing Review 10*, 1 (2002), 54–62.
18. Martínez González, M. *Principios para la explotación dinámica de relaciones entre documentos en las bibliotecas digitales: aplicación al entorno jurídico / Principes d'exploitation dynamique des relations inter-documents dans les bibliothèques électroniques: application au domaine juridique.* PhD thesis, Universidad de Valladolid, Spain / Institut National Polytechnique de Lorraine, France, Sept. 2001.
19. Martínez, M., de la Fuente, P., Derniame, J., and Pedrero, A. Relationship-based dynamic versioning of evolving legal documents. In *Web-knowledge Management and Decision Support - Selected papers from the 14th International Conference on Applications of Prolog*, vol. 2543 of *Lecture Notes on Artificial Intelligence.* Springer-Verlag, 2002, pp. 298–314.
20. Mets, G. D. Consleg Interleaf: SGML Applied in Legislation. In *SGML'96: Celebrating a Decade of SGML* (Boston, 1996).
21. Mínguez, S. M. Detección de enlaces en documentos jurídicos. M.Sc. Thesis, Universidad de Valladolid, Spain, jul 2002.
22. Moens, M.-F. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law 9*, 1 (Mar. 2001), 29–57.
23. Smith, D., and Lopez, M. Information extraction for semistructured documents. In *Workshop on Management of Semi-structured Data* (Tucson, Arizona, 1997).
24. Summers, K. M. *Automatic Discovery of Logical Document Structure.* PhD thesis, Cornell University, aug 1998.
25. Thistlewaite, P. Automatic Construction and Management of Large Open Web. *Information Processing and Management 33*, 2 (1997), 161–73.
26. Wilson, E. Links and structures in hypertext databases for law. In *European Conference on Hypertext, ECHT'90* (Paris (France), 1990), A. Rizk, N. A. Streitz, and J. André, Eds., The Cambridge Series on Electronic Publishing, Cambridge University Press, pp. 194–211.
27. Xu, Y. *An Incremental Apporach to Document Structure Recognition.* PhD thesis, GMD - Forschungszentrum Informationstechnik GmbH, 1998.