TENDENCIAS EN LA INVESTIGACIÓN SOBRE RECUPERACIÓN DE INFORMACIÓN JURÍDICA

María Luisa Alvite Díez*

Resumen: Se presentan las principales líneas de investigación sobre recuperación de información jurídica agrupadas en cinco bloques: investigaciones evaluativas, estudios del comportamiento del usuario en la búsqueda de información legal, aplicaciones de la inteligencia artificial en la recuperación de información jurídica, trabajos que atienden al procesamiento en lenguaje natural y, por último, la gestión de la documentación jurídica mediante el empleo de lenguajes de marcas. Se constata la limitada atención prestada a la recuperación de información jurídica en nuestro país, así como la complementariedad evidente entre algunas de las áreas de investigación reseñadas.

Palabras clave: recuperación de información, documentación jurídica, sistemas de recuperación de información jurídica, evaluación, inteligencia artificial, procesamiento en lenguaje natural, lenguajes de marcas.

Abstract: Research on the legal information retrieval is presented in five main blocks: evaluation research, papers about users behaviour, artificial intelligence techniques in legal information retrieval, studies on natural language processing and the legal information management by means of mark-up languages. We verify a limited interest in our country to legal information retrieval, and finally we consider the evident complementarity among some of the mentioned investigation areas.

Keywords: information retrieval, legal information, legal information retrieval systems, evaluation, artificial intelligence, natural language processing, mark-up languages

1 Introducción

El interés de este estudio se cifra en la Documentación jurídica y las tecnologías de la información orientadas a la recuperación de la misma, soslayamos, por tanto, el área de interés de la informática jurídica de gestión, si bien, como se apreciará, no siempre son dominios disociables.

Los primeros Sistemas de Recuperación de Información (SRI) jurídica hacen su aparición en los años sesenta en el ámbito internacional. Apunta Bing (1) como el origen de LEXIS en Estados Unidos se remonta al año 1964, fruto de una iniciativa de la *Ohio Bar Association* quien contrataría a *Data Corporation* de Dayton (Ohio) para desarrollar el sistema de recuperación. En 1968 fue adquirida por *Mead Corporation*, naciendo *Mead Data Central* que lanzó un prototipo del sistema operativo al año si-

* Área de Biblioteconomía y Documentación. Fac. de Filosofía y Letras. Univ. de León.
Correo-e: dphlad@unileon.es.
Recibido: 21-2-03

guiente. El nuevo servicio apareció en 1973 bajo el nombre de LEXIS preocupado, de modo especial, por la documentación judicial y por la creación de un sistema de recuperación fiable y robusto que, incluso en la actualidad, mantiene muchas semejanzas con la versión inicial.

Su principal competidor se desarrolló en el seno de la editorial jurídica más importante de Estados Unidos, West Publishing que sacó a la luz en 1975 su sistema WESTLAW. El software de recuperación inicial, sobre el que se realizaron desarrollos profundos, fue adquirido a la empresa canadiense *QL Systems*.

En Canadá, el desarrollo guarda relación con el proceso de Estados Unidos en virtud del uso inicial del software *QL System*, originariamente denominado QUIC/LAW que nace en 1968 de la mano del profesor Hugh Lawford. El aspecto más singular del sistema residía en que la recuperación se basaba en algoritmos de ordenación que atendían a la frecuencia de aparición de las palabras. Curiosamente, en 1974 se prefirió dar prioridad a las peticiones booleanas y a los operadores de distancia, reduciendo a simples opciones los algoritmos de ordenación.

En lo que se refiere a Europa, el sistema CREDOC, creado por los notarios de Bélgica, es, según el propio Bing (1) el SRI jurídica pionero en el viejo continente y el primero en enfrentarse a la ardua problemática de la documentación bilingüe. En general, los sistemas europeos comenzaron como sistemas especializados destinados a un pequeño grupo de usuarios a comienzos de los años setenta para, con el paso del tiempo, conformar sistemas más generales disponibles para todos los usuarios como servicios de información legal nacional.

Entre los sistemas más representativos, resulta obligado mencionar ITALGIURE surgido de la iniciativa de la Corte Suprema de Casación italiana, el Congreso de los Diputados y el Istituto per la Documentazione Giuridica de Florencia, este último caracterizado hasta el momento actual por el ingente despliegue de actividades e investigaciones en el terreno de la Documentación, lenguaje jurídico, tecnologías de la información aplicadas al Derecho, etc.

En el Reino Unido, como veremos, los primeros experimentos conducidos por Colin Tapper se remontan al año 1963. Entre 1968 y 1969 se diseñó el SRI legal STATUS en el U.K. Atomic Energy Authority en Harwell. Sin embargo, la puesta en marcha de servicios comerciales no llegaría hasta el año 1978, momento en el que la editorial Butterworths anunció su acuerdo con Mead Data Central para ofrecer el sistema LEXIS.

En Francia, la iniciativa partió del Consejo de Estado y el Tribunal de Casación en 1967, jugando a partir del año 1970 un papel fundamental el CDIJ, posteriormente CNIJ Centre National d'Informatique Juridique. Además de la profusión de sistemas más o menos especializados, se introdujo un sistema de información legal francés basado en el sistema LEXIS, en el año 1983. Como señala Moreno de la Fuente (2), la fecha clave en el desarrollo de la informática jurídica en Francia corresponde al año 1985 gracias al denominado «*Informe Leclerc*», que daría origen al sistema JURIDIAL, eligiendo Questel Plus como software de recuperación.

El inicio de la aplicación de la informática en el contexto jurídico español viene situándose en el último bienio de la década de los sesenta, con la puesta en marcha del *Proyecto Ibertrat* en 1968 y el *Plan General de Informática Jurídica* entre cuyos objetivos se hallaba el de facilitar búsquedas y recuperaciones jurisprudenciales.

Por último, en la Unión Europea la base de datos CELEX (Comunitatis Europeae

Lex) nace en 1966 y está operativa desde 1970, se empleó el software de recuperación MISTRAL e incorpora además de legislación comunitaria, jurisprudencia del Tribunal de Primera Instancia y del Tribunal de Justicia de las Comunidades Europeas, Trabajos preparatorios, Disposiciones nacionales de ejecución de las Directivas Comunitarias y Preguntas parlamentarias (3).

2 Investigación evaluativa de sistemas de recuperación de información jurídica

El campo jurídico ha sido precursor en la recuperación en línea del texto completo de los documentos y el primer experimento evaluativo sobre literatura jurídica se remonta al año 1964. En él se evaluó el sistema americano LITE, comparando la recuperación en la base de datos con las técnicas convencionales de indización manual.

En 1960 el sistema fue presentado con éxito en la Conferencia de la Asociación Americana de Abogados y sus métodos se aplicaron al establecimiento del primer servicio de información legal operativo, adquirido por el Air Force Accounting and Finance Center, en Denver, Colorado, recibiendo el nombre de Legal Information Thru Electronics (LITE).

Los resultados de este primer test, sintetizados por Tenopir y Ro (4), demostraron que el sistema automatizado a texto completo trabajaba con un porcentaje de efectividad del 92,5%, mientras que la proporción del sistema manual era del 51,6%. El índice de exhaustividad relativa de la recuperación automatizada fue del 93,5%, mientras que el del sistema manual fue del 62%.

Entre los años 1966-1967 se realizó un experimento por parte de la Joint American Bar Foundation y el International Business Machine Project, para comparar la recuperación automatizada sobre el texto completo en una base de datos que contenía 5.800 sentencias, con respecto a la recuperación manual. Se concluyó que el sistema de recuperación manual y el automatizado tenían un comportamiento similar en términos de exhaustividad y que la búsqueda manual duplicaba la efectividad en términos de precisión.

En Inglaterra, la primera investigación sobre texto completo fue liderada por el abogado Colin Tapper y se conoce como el *Oxford Experiment*. Se construyeron dos bases de datos para el experimento, una de ellas contenía informes del Alto Tribunal (2 millones de palabras) y la otra sentencias administrativas (1 millón de palabras). Se compararon los resultados, en términos de exhaustividad y precisión, con respecto a la búsqueda a través de un índice manual. Se halló un índice de exhaustividad del 70% y del 29% de precisión en la recuperación automatizada frente a un 49% de exhaustividad y un 92% de precisión en la recuperación manual. El sistema a texto completo producía, por tanto, valores de precisión inferiores (4).

Sin lugar a dudas, la investigación de mayor proyección y trascendencia ha sido la realizada por Blair y Maron, se trata de un experimento a gran escala para evaluar la eficacia de un sistema de recuperación a texto completo, cuyas conclusiones fueron publicadas en 1985, pero cuyo debate y estudio se ha prolongado hasta el momento actual. Enumera Blair (5) dieciocho trabajos que han estudiado o comentado con detenimiento la evaluación realizada.

Blair y Maron (6, 7) evaluaron el sistema STAIRS (Storage and Information Retrieval System) de IBM en un entorno concebido para el apoyo al proceso judicial. La

base de datos estaba integrada por diversos tipos de documentos: informes, escritos, correspondencia, memorias, transcripciones, etc. No contenía, por tanto, exclusivamente documentos jurídicos tipificados como tales, reunía unos 40.000 documentos y más de 350.000 páginas de texto que se emplearon para la defensa de un largo proceso colectivo. Sus resultados no son solamente una crítica al sistema STAIRS sino, más bien, una crítica a los principios en los que se basan los SRI de texto completo.

En el experimento participaron dos abogados que constituían la parte principal de la defensa en el juicio para el que se empleó el sistema. Los juristas generaron 51 peticiones diferentes de información que fueron trasladadas en preguntas formales por otras dos personas no juristas, ambas familiarizadas con el caso y con experiencia en el uso del sistema STAIRS. Estas personas realizaban las búsquedas en la base de datos hasta que encontraban un conjunto de documentos que creían podían responder a la petición de información. Se enviaban los documentos a los abogados quienes evaluaban los mismos ordenándolos de acuerdo a la siguiente escala: «vital», «satisfactorio», «marginalmente relevante» o «irrelevante» de acuerdo a la petición inicial.

Las conclusiones resultaron inesperadas:

— Escaso porcentaje de exhaustividad, situado en un 20%, mientras que la precisión alcanzaba una media del 79%.
— No se hallaron diferencias significativas entre los abogados en cuanto a la habilidad en la búsqueda.
— Tampoco fueron halladas diferencias significativas entre la capacidad de búsqueda de los abogados y de los intermediarios no juristas.
— Escasa capacidad de los abogados para anticipar las palabras o frases que deberían usarse para recuperar documentos relevantes y evitar los irrelevantes.
— La variabilidad de las palabras y frases empleadas en el sistema para tratar la misma materia era realmente excepcional e imprevisible.

Explican Blair y Maron (6) la baja exhaustividad como característica consustancial a los sistemas de recuperación a texto completo, en los cuales la recuperación por materias es dificultosa a pesar de las investigaciones sobre indización automática y, en general, sobre procesamiento en lenguaje natural. Además, el valor de la exhaustividad disminuye a medida que aumenta el tamaño de la base de datos o, expresado de otro modo, el esfuerzo requerido para lograr el mismo nivel de exhaustividad aumenta a medida que aumenta el tamaño de la base de datos, a menudo, apuntillan Blair y Maron, con más rapidez que el tamaño de la base de datos.

Por tanto, del estudio se desprende que la recuperación no alcanza niveles satisfactorios y que es necesario emplear estrategias sofisticadas de búsqueda, incluso reconociendo que no se han logrado recuperar todos los documentos relevantes del sistema.

A pesar de que el experimento se desarrolló hace casi dos décadas, la recuperación por texto libre continúa basándose, en general, en la equiparación de términos, del mismo modo que el sistema STAIRS. Así lo manifiesta Blair (5), para quien una de las razones fundamentales que justifican el interés ininterrumpido por el estudio es que los sistemas comerciales continúan empleando técnicas de recuperación similares a las evaluadas en STAIRS.

Berring (8) aprecia tales implicaciones teóricas en la utilización de sistemas de re-

cuperación para el acceso a la literatura legal que habla de un «nuevo paradigma», entendiendo que las bases de datos jurídicas suponen una ruptura con respecto a los sistemas tradicionales impresos y tienen obligatoriamente que influir en la práctica del Derecho.

Compara este autor los sistemas LEXIS y WESTLAW con el tradicional *West's American Digest System,* basado en rígidos índices y subíndices y, aun valorando el salto cualitativo que supone para el usuario la libertad de la búsqueda en el texto completo, obviando las restricciones de los índices impuestos por los juicios de una determinada editorial, concluye que para una búsqueda eficaz en la esfera conceptual es irremplazable la indización humana y añade que solamente estos sistemas se pueden considerar herramientas de búsqueda ultimadas si se añaden tesauros implementados por indizadores profesionales.

En 1986 Dabney (9) reflexiona sobre las implicaciones de los resultados de Blair y Maron en el desarrollo de lo que denomina sistemas CALR (Computer-Assisted Legal Research). Compara someramente el sistema STAIRS, diseñado para el apoyo a un proceso judicial y los sistemas LEXIS y WESTLAW, típicos sistemas CALR, concluyendo que son mayores las semejanzas entre los dos modelos que las diferencias. Insiste igualmente Dabney en la dificultad de predecir el término de búsqueda, aspecto subrayado por Blair y Maron, y divide el problema de la equiparación de las palabras en tres categorías: sinónimos, palabras ambiguas y expresiones complejas; la mayor dificultad no reside en la «mera» imprecisión del lenguaje sino en el carácter analítico de la ciencia jurídica, es decir, el Derecho implica ideas, ideas que no se hallan directamente relacionadas con una palabra determinada. A menos que un concepto legal dado pueda ser representado de forma fiable por una única palabra o conjunto de palabras, el concepto será invisible para la persona que busca en un sistema de texto libre.

Gerson (10) remite a la evaluación de Dabney de 1993 sobre el rendimiento de la recuperación de los sistemas jurídicos, como paradigma de investigaciones evaluativas que pretenden demostrar que los resultados de Blair y Maron se pueden generalizar a los sistemas de recuperación jurídica basados en el modelo booleano. En el experimento de Dabney se usaron 23 Comentarios de materias legales incluidas en los *American Law Reports* para proporcionar las preguntas de la prueba y los conjuntos de respuestas. El estudio trató de localizar todas las sentencias mencionadas dentro del marco de los Comentarios empleando LEXIS y WESTLAW.

Los resultados del test mostraron en LEXIS un índice del 26,4% de exhaustividad y un 11,5% de precisión, mientras que WESTLAW alcanzó un 32,2% de exhaustividad y un 12,4% de precisión. Se observaron limitaciones en los sistemas LEXIS y WESTLAW para la recuperación por materias, concluyendo que la exhaustividad solamente puede ser mejorada a costa de sacrificar la precisión.

Ahora bien, como aduce Dabney (9), las necesidades de información jurídica no se restringen al acceso por materias, muy al contrario, existen otros medios previstos por los sistemas que atienden a un gran número de peticiones: fecha, número de documento, artículos legales estudiados, tribunal, citas del documento, etc. y que se hallan adecuadamente resueltos por los denostados sistemas booleanos.

Las observaciones que se infieren de los estudios anteriores inciden en que un SRI legal efectivo debe generar una alta exhaustividad, debe recuperar un alto porcentaje de documentos relevantes. Centrándose en la recuperación de sentencias —no se debe

olvidar la trascendencia de la jurisprudencia en los sistemas legales anglosajones—Berring (8) establece como el rasgo más distintivo de los profesionales jurídicos americanos en la búsqueda de información legal, su obsesión por localizar todas las sentencias relevantes para un caso particular y Dabney (9) subraya como punto de partida en un operador jurídico la intención de no esperar encontrar un único documento que resuelva una petición de búsqueda, sino la consideración de los casos similares útiles. En consecuencia, los juristas están dispuestos a examinar muchas sentencias con la finalidad de no perder ningún documento relevante.

En opinión de Gerson (10), los resultados de Blair y Maron sobre la recuperación de información a texto completo no pueden ser generalizados a los SRI legal WESTLAW y LEXIS, debido a que la documentación tratada en el estudio no era exclusivamente jurídica, como se ha apuntado más arriba, y muchos de los documentos estaban redactados de modo «informal». Los sistemas propiamente jurídicos, añade este autor, se caracterizan por incluir documentos consistentes, de formato y tono uniforme y escritos de modo formal.

Además, este autor abandona el estándar que priorizaba la exhaustividad como indicador básico del rendimiento y procede a la evaluación de los modelos de recuperación por relevancia de los sistemas LEXIS Y WESTLAW, empleando como medida de rendimiento la precisión y no la exhaustividad. Argumenta, basándose en las consideraciones de Burson (11), que las expectativas de los juristas se cifran en herramientas que logren una alta precisión adecuándose al núcleo del tema buscado. Gerson (10), añade que dentro del contexto de la estructura organizativa exhaustiva propia de la documentación judicial, consabida en el entorno de los usuarios de información jurídica, la mayor parte de los mismos quieren emplear los SRI legal para encontrar rápidamente unos pocos documentos en concreto, anhelan solamente un pequeño conjunto de documentos. Por tanto, son usuarios que esperan que el conjunto recuperado contenga un número elevado de documentos relevantes.

Admitiendo que el tamaño ideal del conjunto de documentos recuperados dependerá del usuario particular, existen evidencias que señalan la cifra de 20 o 25 documentos como la más adecuada. Así lo han considerado los sistemas LEXIS y WESTLAW en sus métodos de recuperación por relevancia. Gerson en su investigación toma la cifra de 20 documentos y considera como medida de rendimiento apropiada el porcentaje de documentos relevantes contenidos en ese conjunto expresada del modo siguiente: el número de documentos recuperados relevantes dividido por el menor del número total de documentos relevantes o por 20.

La metodología de Gerson en lo que atañe a la elaboración de las preguntas es similar a la empleada por Dabney, usando como base de la prueba los Comentarios de los *American Law Reports* para evaluar los modelos no booleanos de LEXIS (FREESTYLE) y WESTLAW (WIN: Westlaw Is Natural), buscando temas legales complejos.

Los resultados de la investigación mostraron la efectividad de los dos sistemas de acuerdo a las medidas de rendimiento propuestas. FREESTYLE alcanzó un rendimiento del 31%, esto es, devolvía una media de 6 sentencias relevantes y WIN obtuvo un 37%, 7 documentos relevantes de los 20 devueltos. La posición media del primer documento fue de 2,5 en WIN y 2,3 en FREESTYLE, es decir, los usuarios han de ojear dos o tres sentencias para localizar una ajustada directamente al caso. Concluye Gerson señalando que este rendimiento dependerá, en última instancia, de las cir-

cunstancias del usuario concreto, pero servirá de ayuda para determinar si se estima oportuno o no emplear los métodos de ordenación por relevancia de los sistemas evaluados.

El desarrollo de Internet en los últimos años ha propiciado una preocupación evidente por la RI legal en particular y, consecuentemente, han aparecido criterios para la evaluación y control de calidad de los sistemas jurídicos dispuestos en la red.

La profesora Newman (12) llama la atención sobre la ingente labor de Facultades de Derecho de numerosas universidades y otras instituciones públicas que se han visto involucradas en la publicación de fuentes jurídicas de indudable valor, hoy accesibles al lado de sitios web de escasa calidad y dudosa autoría, que pueden causar un gran perjuicio a muchos usuarios que acceden a información legal sin contrastar la fiabilidad de la fuente.

Propone esta autora la urgente necesidad de aplicar criterios rigurosos para la evaluación de sitios web jurídicos, los fundamentales son contenido y consistencia de la base de datos, autoría, diseño amigable e inclusión de herramientas de búsqueda potentes. Se muestra partidaria de que los sistemas estén basados en el modelo vectorial y ofrezcan al usuario posibilidades de ponderación.

La clara tendencia al empleo de Internet por parte de los juristas como un recurso más de información, unido al crecimiento de la información jurídica disponible en la web, ha convertido a las herramientas de búsqueda en Internet en herramientas vitales de búsqueda legal, lo que ha motivado la realización de estudios de evaluación del rendimiento de los buscadores en la recuperación de información exclusivamente jurídica. Barmakian (13) realiza un doble estudio, de un lado, analiza la efectividad de 15 buscadores en la recuperación de items conocidos relacionados con el ámbito jurídico y, de otro, evalúa el rendimiento de 10 buscadores en la recuperación de búsquedas jurídicas por materias. Los buscadores seleccionados fueron buscadores generales, con la excepción de dos motores especializados en materias legales: LawCrawler y LawRunner.

Los aspectos más interesantes que se desprenden de los resultados de la evaluación reflejan:

— Contrariamente a lo que cabría esperar, el rendimiento de los buscadores jurídicos fue peor que el de los de carácter general en la búsqueda por items conocidos.
— El rendimiento en la búsquedas por materias jurídicas, sin embargo, resultó decepcionante. El escaso nivel de relevancia indica que los motores de búsqueda no son todavía alternativas viables a los SRI jurídica comerciales para la búsqueda por materias.
— El comportamiento de los motores de búsqueda LawCrawler y LawRunner fue superior al resto en las búsquedas temáticas, sin embargo, la relevancia de los resultados no logró un mínimo adecuado.
— El mayor nivel de solapamiento se da entre los dos motores especializados. Google y Fast son los motores que devuelven un mayor número de resultados «únicos» y un menor nivel de solapamiento con respecto al resto de buscadores.

Tomando como punto de partida la aparente necesidad de herramientas de recuperación web específicas para determinadas comunidades de usuarios, como los juris-

tas, en el caso que nos ocupa, Dempsey *et al*. (14) diseñan y evalúan a pequeña escala software de recuperación jurídica en entorno web. *LIBClient*, con el motor de búsqueda *IRISWeb,* el sistema permite la búsqueda a texto completo en lenguaje natural sobre las páginas recogidas por medio del motor *IRISWeb*. Los resultados en la recuperación resultan realmente alentadores y creemos que el desarrollo de herramientas de segunda generación, adecuadas a tareas de recuperación especializadas, puede alcanzar una gran expansión en los próximos años.

Hanft (15) expone la necesidad de idear un nuevo modelo de búsqueda de información jurídica en lo que él denomina «edad electrónica». Este nuevo patrón ha de enfrentarse a los retos de la sobreabundancia informativa, accesibilidad de datos, (no exclusivamente documentación jurídica sino también información estadística, acceso a registros públicos, documentación administrativa, prensa, etc.), validez de la información, volatilidad de la misma y pérdida de visión periférica necesaria para el estudio de los asuntos legales. La clave de este modelo radica, en opinión de este profesor, en la confianza en fuentes secundarias de alta calidad dispuestas de modo conceptual y con una cobertura completa.

3 Comportamiento del usuario en la búsqueda de información jurídica

Desde sus orígenes, la Documentación se ha preocupado de estudiar el comportamiento del usuario en la búsqueda de información, ciñéndose, en principio, a las pautas de actuación del usuario estudiante o científico. A principios de la década de los ochenta del pasado siglo, las investigaciones comienzan a particularizar específicamente en la actitud de los profesionales, entendidos como usuarios que buscan información en el desempeño de su ocupación o trabajo diario.

Se han propuesto varios modelos en el comportamiento del usuario en la búsqueda de información —teoría del Sense-Making de Dervin, modelos de Wilson, Kuhlthau, Ellis, etc.— y convenimos con Wilson (16) entendiendo de modo global el comportamiento ante la información como el conjunto de actividades que una persona puede dedicar a la identificación de su necesidad de información, la búsqueda de esa información y la utilización y transferencia de la misma. Por tanto, existe una relación directa entre el comportamiento humano en la comunicación y el comportamiento ante la información y una vinculación efectiva, igualmente, con el campo de la interacción hombre-máquina.

En lo que respecta a los profesionales jurídicos, señalan Leckie, Pettigrew y Sylvain (17) se han realizado exiguos estudios, y éstos se han centrado en temas éticos y de responsabilidad profesional, soslayando mencionar aspectos relacionados con las necesidades y el uso de la información de este colectivo. Se apunta de modo somero, en el citado trabajo, que la búsqueda de información por parte de los juristas se halla mediatizada por una compleja interacción de variables personales y contextuales:

— Contexto de la organización en la que trabaja.
— Área de especialización jurídica.
— Grado de experiencia del jurista que determina la mayor o menor necesidad de acudir a determinados tipos de búsqueda legal.
— Nivel de formación en el uso de fuentes jurídicas y estrategias específicas de recuperación de información.

— Falta de amigabilidad y exhaustividad de los SRI legal. Se alude a problemas de cobertura y acceso que estos sistemas llevan arrastrando desde sus orígenes y que persisten, incluso han aumentado ante la necesidad de seleccionar el producto adecuado entre la plétora de servicios en el mercado, ya sea en línea o en CD-ROM, con diferente cobertura pero con información potencialmente relevante.

Es difícil, según este estudio, establecer modelos adecuados del comportamiento en la búsqueda de información de los abogados frente a otros profesionales estudiados en el mismo. Se refieren Leckie, Pettigrew y Sylvain a diversas tentativas americanas y canadienses, como la desarrollada en este último país por el propio Departamento de Justicia estructurando ordenadamente las actividades llevadas a cabo por un jurista, pero sin explicitar en el modelo las implicaciones de la búsqueda de información en esas actividades.

Concluyen hablando de estos modelos lineales como una simple ayuda para entender las complejidades de la investigación legal y permitirnos entrever las dificultades a las que se enfrentan los juristas en la recuperación de información. Sin embargo, solamente proporcionan una visión parcial de la amplia gama de actividades de búsqueda de información emprendidas por la mayor parte de los abogados en la práctica diaria de su trabajo.

Berring (8) incide en las habilidades del jurista en el manejo de SRI como componente modificador del comportamiento del mismo ante la búsqueda de información legal. Insiste, por ello, en la importancia de la formación integrada en las propias facultades de Derecho y no en dependencia de los cursos de formación de las bibliotecas correspondientes o de los programas de familiarización con los sistemas que realizan las principales empresas del sector.

Tanto es así, que para Berring existen limitaciones del propio usuario final sin conocimientos suficientes sobre el uso de los sistemas empleados o sin renovar las habilidades puntuales en el manejo de sistemas que sufren modificaciones constantes. A esto se añadiría como agravante la engañosa facilidad de uso de las bases de datos que crean en el jurista una «falsa sensación de competencia».

El trabajo más profundo efectuado hasta el momento es el de Kuhlthau y Tama (18) realizado dentro del programa de investigación basado en el modelo del «Proceso de Búsqueda de Información» *(Information Search Process, ISP)* desarrollado por Carol Kuhlthau en un conjunto de estudios previos. El modelo se inscribe dentro de la aproximación cognitiva y aplica tareas complejas que requieren búsqueda, recogida, interpretación y uso de información sobre un amplio periodo de tiempo.

El objetivo era alcanzar un mejor conocimiento de la variedad de tareas en las que se desenvuelven los abogados, como un grupo particular de trabajadores de la información, saber cómo utilizan la información para realizar su trabajo, cuál es el papel que juegan los intermediarios en el proceso de búsqueda y empleo de la información y explorar qué fuentes, sistemas y servicios podrían ser de utilidad.

Kuhlthau y Tama concluyen insistiendo en la propuesta de sistemas *«just in time»* y *«just for you»*. Los abogados participantes en este estudio mostraron la necesidad de sistemas de apoyo al proceso de construcción, no sistemas que ofrecen meramente una respuesta correcta, reclaman flexibilidad y control del proceso en manos del usuario. El problema recurrente es el de la terminología, muestran una gran falta de

confianza en sistemas que obligan a introducir una palabra clave y devuelven «sólo palabras». Las bases de datos jurídicas se muestran útiles para resolver tareas rutinarias o peticiones específicas, pero su utilidad es dudosa para búsquedas poco específicas o para tareas complejas.

Partiendo del modelo propuesto por Leckie, Pettigrew y Sylvain sintetizado anteriormente, Wilkinson (19) realiza un trabajo exploratorio sobre el comportamiento de los abogados en la búsqueda de información, analizando 154 entrevistas a abogados en ejercicio de Ontario. El estudio forma parte de una investigación interdisciplinar sobre la profesión jurídica patrocinada por el Social Science and Humanities Research Council de Canadá y el Westminster Institute for Ethics and Human Values. Los resultados, en nuestra opinión, cuentan con puntos desconcertantes que sugieren la urgencia de emprender trabajos empíricos que profundicen con rigor en el conocimiento de las necesidades reales de información y el comportamiento y uso de la misma por parte de estos profesionales.

4 Sistemas basados en el conocimiento legal y sistemas legales expertos

Un sistema experto es una herramienta de software basada en técnicas de Inteligencia Artificial (IA). En sus inicios, los objetivos eran simular y quizá reemplazar el razonamiento humano en marcos diversos. En estos momentos los sistemas expertos se han replegado a retos "más modestos".

Los sistemas basados en el conocimiento suelen ser sistemas híbridos que combinan varios esquemas de representación del conocimiento. Este tipo de sistemas puede también incrustarse en otras aplicaciones principales, posibilitando el uso de técnicas de razonamiento de IA empleadas conjuntamente con técnicas tradicionales de procesamiento de información. Se utiliza en ocasiones, como sinónimo, el término de «sistemas inteligentes de apoyo a la toma de decisiones». Tradicionalmente, tanto los sistemas expertos como los basados en el conocimiento se han venido desarrollando dentro del campo de la «Ingeniería del conocimiento».

La IA aplicada al ámbito jurídico supone la confluencia de varias disciplinas, lo que complica sobremanera el seguimiento de sus desarrollos, a los expertos jurídicos se han de unir ingenieros del conocimiento, programadores, analistas de sistemas, usuarios finales, etc. Para Erdelez y O'Hare (20) se trata de una campo en continuo progreso desde los años ochenta, en el que lo esperable es caminar desde la recuperación de información textual hacia sistemas basados en el contenido de los documentos y en el uso de agentes inteligentes.

Curran y Higgins (21) señalan que las técnicas de Inteligencia Artificial suponen el reto más ambicioso emprendido hasta el momento para mejorar el proceso de búsqueda legal. No se trata solamente de que el sistema devuelva documentos relevantes, sino de proponer una guía de cómo emplear los mismos. En la línea de lo señalado más arriba, apuntan estos autores cómo inicialmente este tipo de aplicaciones se desarrolló con la intención de proporcionar soluciones a los problemas jurídicos como lo haría un experto legal; sin embargo, tales sistemas han reconducido sus objetivos hacia la incorporación de conocimiento jurídico para proveer conocimiento legal, como guía o apoyo a la toma de decisiones por parte de los profesionales jurídicos, de modo que consideran más adecuado referirse a «Sistemas basados en el conocimiento legal» o «Sistemas de apoyo a la toma de decisiones legales».

La clasificación más exhaustiva es la presentada por Bench-Capon (22), quien establece las siguientes categorías de sistemas basados en el conocimiento en el ámbito jurídico:

1. Sistemas legales clasificados por tarea:

1.1. Sistemas de asesoramiento dirigidos a abogados: entre los sistemas de este tipo cabe mencionar: *BNA (British National Act System), Latent Damage Advisor, HYPO, CABARET,* en los que se incorporan como base de conocimiento otros casos legales precedentes. Estarían incluidos en esta categoría los sistemas para la preparación de documentos estándar, este grupo creemos que podría contener los numerosos sistemas de «Formularios» aparecidos en España desde finales de los noventa, en algunos casos integrados con bases de datos tradicionales.
1.2. Sistemas de asesoramiento dirigidos al público, como *DHSS Demonstrator Advice Systems* o el sistema desarrollado por Arthur Andersen con gran éxito, *RPFA (Retirement Forecast and Advice System).*
1.3. Sistemas destinados a la judicatura, ejemplos como *Local Office Demostrator.* En Holanda se han usado sistemas como *JURICAS,* desarrollado por la Universidad Erasmus de Rotterdam con la intención de servir de ayuda a los jueces en casos rutinarios y *TESSEC,* desarrollado por la Universidad de Twente aplicado al área de la Seguridad Social.
1.4. Otras tareas, como sistemas tutoriales de aprendizaje por ordenador, incluye Bench-Capon en esta categoría los sistemas de ayuda a la redacción de disposiciones legales y apunta el sistema *ExpertiZe* como el modelo más interesante, basado en reglas para evaluar la consistencia de la legislación y determinar los efectos de la misma a través de un proceso de simulación. Particularmente interesante es el uso de la simulación con una base de datos estadística para predecir el coste de los cambios legislativos propuestos. Se incluyen, por último, los sistemas de adquisición de conocimiento legal, cuyos proyectos más representativos son *FLEXICON*, parte del proyecto *FLAIR* de la University of British Columbia, proyecto dirigido a la recuperación general de documentos jurídicos, se basa en la indización automática de jurisprudencia; *ILAM* intenta tratar de forma semiautomática la legislación fiscal italiana confiando en la forma y la regularidad encontrada en los textos escritos en estilo jurídico y *ACAT*, semejante al proyecto anterior, que trabaja con disposiciones legales en francés. Este área de adquisición de conocimiento legal guarda una estrecha vinculación con los desarrollos que veremos sobre procesamiento en lenguaje natural concitando una expectación enorme.

2. Sistemas legales clasificados según su formalización:

2.1. Sistemas basados en la producción de reglas, fue la tendencia inicial seguida en el diseño de los sistemas expertos
2.2. Sistemas basados en la programación lógica, como el mencionado *BNA*, *LEGOL* basado en el álgebra relacional o *ESPLEX.*
2.3. Sistemas estructurados basados en objetos, incluye en esta categoría las redes

semánticas, marcos y programación orientada a objetos. Este modelo se ha seguido mayoritariamente para la representación de casos judiciales.

2.4. Lenguajes de representación legal especializada.

3. Sistemas legales clasificados por el método de razonamiento:

3.1. «*Black Letter systems»,* término empleado para referirse a aquellos sistemas más simples, con escasas pretensiones y dudosas ventajas derivadas de su ayuda.

3.2. Sistemas expertos, tratan de implantar habilidades o pericia jurídica procedentes bien de un experto jurista, bien de fuentes escritas o de la combinación de ambos. El modelo existente tiende a quedarse en un razonamiento puramente deductivo, pese a todo, se ha mostrado su gran utilidad en la práctica (*DHSS*, *JURICAS*, etc.).

3.3. Sistemas de razonamiento basados en casos, el mecanismo básico no es la deducción sino la equiparación. Se recuperan aquellos casos similares al caso que se está estudiando y se aplica el principio del tratamiento de casos semejantes de un modo similar. Uno de los sistemas más sofisticados de este tipo es *HYPO*.

3.4. Sistemas que construyen el razonamiento legal, el interés reside en generar un argumento basado en casos, más que un caso que es considerado suficientemente cercano para justificar una decisión. Estos programas han sido objeto de un enorme interés teórico y académico y han dado lugar a técnicas incorporadas en otro tipo de sistemas.

3.5. Sistemas de recuperación o sistemas de recuperación conceptual, centrados en el perfeccionamiento de la habilidad para recuperar información relevante. Los máximos exponentes son LEXIS Y WESTLAW.

Galindo y Lasala (23), proponen una clasificación alternativa de los sistemas jurídicos de IA, basada en una ontología temática que se subdivide atendiendo a la función o actividad jurídica a la que se dirigen: sistemas orientados a la aplicación —ayudan a decidir sobre la fundamentación jurídica de casos concretos—, sistemas orientados a la interpretación, sistemas orientados a la construcción de dogmas, y sistemas orientados a la construcción de teorías normativas.

Coinciden Curran y Higgins (21) y Bench Capon (22) en señalar que la mayor parte de los sistemas en este campo han adoptado técnicas basadas en uno de los dos paradigmas teóricos legales dominantes:

— Sistemas sustentados en reglas *(Rule-based systems),* implican la adopción de un punto de vista basado en el positivismo, entendiendo el Derecho como un conjunto determinado de normas.

— Sistemas sustentados en casos *(Case-based reasoning, CBR),* inciden en el reconocimiento de que un importante componente del razonamiento legal se halla en la identificación de casos que sientan precedentes en el ordenamiento legal. Sobre este tipo de sistemas se han efectuado un mayor número de aplicaciones.

Llama la atención Bench-Capon (22) sobre el interés de algunos sistemas como

PROLEXS que emplean el paradigma más adecuado dependiendo de la fuente legal tratada.

Resulta necesario, cuando menos, mencionar los trabajos de Matthijssen (24, 25, 26) llevados a cabo en el Center for Law, Administration and Informatization de la Universidad de Tilburg (Holanda). El aspecto central, y de mayor interés para nosotros, es la RI jurídica, sin embargo, profundiza asimismo en el necesario apoyo a tareas de redacción de documentos jurídicos.

Estudia cómo se ha de representar el conocimiento legal en un SRI para superar los problemas detectados. A pesar de la existencia de varios lenguajes de recuperación, del uso de diferentes modos de indización y modelos de relacionar una pregunta con un índice (booleano, espacio vectorial, probabilístico), existe una característica común a todos los sistemas de recuperación de información jurídica, en los que es preciso traducir la necesidad de información en forma de conceptos legales en una pregunta que debe ser formulada por medio de conceptos técnicos de la base de datos. Matthijssen se refiere a este obstáculo o problema denominándolo «vacío conceptual». A este problema se ha de añadir la escasez de conocimientos sobre las estructuras de almacenamiento de las bases de datos y el funcionamiento de los sistemas de recuperación. El resultado suele provocar una pérdida de la mayoría de la información contextual que determina la necesidad de información en ese proceso de formulación de la pregunta y la búsqueda resultante suele ser, a menudo, demasiado genérica.

El autor presenta como solución un prototipo, al que denomina *ARMOR* (*ARgument MOdel based Retrieval system*), ideado para la búsqueda de información en el área específica del procedimiento administrativo. La idea esencial es que un SRI jurídica ha de adaptarse a las necesidades de información del usuario y en el preciso momento en el que es solicitada dicha información. Para ello ha de superarse el índice simple, que solamente proporciona palabras, e ir hacia un modelo más sofisticado, agrupando las palabras en materias, ampliando las relaciones entre los términos por medio de un tesauro y creando, finalmente, un hiperíndice que representa la información atendiendo a su contenido y a su estructura y que permite interrelacionar legislación y jurisprudencia. Matthijssen lo denomina «*tarea basada en hiperíndice*», concebido éste como un puente para superar ese vacío conceptual existente entre los usuarios y las bases de datos jurídicas.

Por su parte, Galindo y Lasala (23) proponen la utilización de tecnología de IA para el desarrollo de lo que denominan «Sistemas Inteligentes de REcuperación de Documentación Jurídica (SIREDOJ)», en cuyo planteamiento incluyen comprensión del lenguaje natural e integración de distintos sistemas de bases de datos con sistemas expertos. En este entorno presentan el prototipo ARPO-2, diseñado para utilizar argumentos legales relativos a incumplimiento de contratos de obra.

Curran y Higgins (21) realizan una interesante recapitulación sobre la RI legal construida sobre dos tipos de sistemas fundamentales. De un lado, los SRI jurídica «tradicionales», las herramientas sin duda más empleadas, fallan, en su opinión, a causa de la falta de estructura adecuada, dado que para indizar información jurídica es preciso tener en cuenta los conceptos legales más que tomar como base las palabras clave. De otro lado, los sistemas legales expertos o las diversas aplicaciones de IA legal fallan a nivel filosófico/teórico y a nivel práctico, a pesar de la intensa y laboriosa investigación llevada a cabo.

Precisamente Erdelez y O'Hare (20) apuntan como una de las posibles razones que

explican la falta de éxito comercial de los sistemas legales expertos, la estructura abierta de los conceptos legales. Curran y Higgins (21) van más allá y argumentan la inexistencia de productos genéricos comerciales por lo costoso del desarrollo de estas aplicaciones, fácilmente deducible por lo señalado más arriba, y por la dificultad enorme para su mantenimiento y puesta al día, aspecto fundamental en cualquier aplicación informática e inasumible en una materia como el Derecho, viva y cambiante.

Con el objetivo de aumentar la rapidez de la investigación legal sin pretender simular el razonamiento jurídico, proponen Curran y Higgins una vía alternativa, un modelo de SRI basado en Java. Partiendo de principios de la inteligencia artificial (clasifican sentencias y doctrina en términos de los «factores» presentes en dichas fuentes) y de una indización en la que se atiende a conceptos y principios jurídicos, muestran un prototipo a pequeña escala que pretende vertebrar una RI legal inteligente.

Finalmente, hemos de mencionar el papel fundamental desempeñado en la investigación y desarrollo de los sistemas aquí referidos y de herramientas sofisticadas para el tratamiento y la RI legal en general, instituciones como la International Association for Artifical Intelligence and Law (IAAIL), Foundation for Legal Knowledge Based Systems (Jurix), así como las sucesivas conferencias auspiciadas por dichas entidades, International Conference on Artificial Intelligence and Law, JURIX: International Conference on Legal Knowledge-Based systems, sin olvidar las revistas más específicas en este ámbito, *Artificial Intelligence and Law* y *Journal of Information Law and Technology (JILT)*.

5 Procesamiento en lenguaje natural (PLN)

El procesamiento en lenguaje natural constituye un área de investigación que estudia la forma en la que el texto introducido en lenguaje natural en un sistema informático puede ser manipulado y transformado del modo más adecuado para un mejor tratamiento. Partiendo de la evidencia de que los humanos nos comunicamos por medio del lenguaje natural, procede deducir que esta es la forma, en principio, más fácil y efectiva para que interactúen hombre y máquina.

El PLN cuenta con disciplinas fuertemente relacionadas, ciencias cognitivas involucradas en el desarrollo de teorías psicológicas sobre el lenguaje humano, y, principalmente, la Lingüística generativa, la Inteligencia Artificial y la Lingüística computacional.

Las técnicas del PLN se efectúan mediante diversos análisis, ocupando cada uno de ellos distintos niveles: análisis morfológico, sintáctico, semántico y pragmático. El enorme crecimiento de las bases de datos a texto completo y los problemas de recuperación intrínsecos a las mismas, han sugerido a muchos investigadores la posibilidad de introducir estas técnicas para optimizar los resultados mediante la expresión de las búsquedas en lenguaje natural, evitando los problemas derivados de los lenguajes controlados y del empleo de los diversos operadores facilitados por los lenguajes de recuperación concretos, por medio de la incorporación de palabras semánticamente relacionadas y de formas flexivas de modo que los términos permitan ampliar la pregunta automáticamente. En un modelo ideal se restringirían los documentos a aquéllos que respondiesen al sentido de la búsqueda y no a la equiparación de una palabra.

Como observan Pérez-Carballo y Strzalokowski (27), estos procedimientos pueden

ser utilizados de modo eficiente a gran escala y pueden tener un impacto significativo en la RI para superar la inadecuación de los métodos puramente cuantitativos. Son muchos los experimentos sobre RI empleando PLN realizados en los últimos años con resultados prometedores, entre otros los trabajos de TREC (Text Retrieval Conferences, trec.nist.gov) en este sentido y, sin embargo, algunos investigadores como Sparck Jones (28) ponen en duda su eficacia para las tareas relacionadas con la recuperación común. En una línea semejante se manifiestan Arampatzis *et al*. (29), afirmando que las técnicas disponibles en estos momentos de PLN adolecen de falta de precisión y eficacia y, aun más, se requieren más investigaciones para demostrar que la estructura sintáctica puede sustituir adecuadamente al contenido semántico.

Se puede decir, generalizando, que los SRI se adaptan a la representación textual mientras que los sistemas expertos, como hemos visto, tratan de representar el conocimiento, los primeros emplean el lenguaje como forma primaria de representación del conocimiento pero exclusivamente como almacén de palabras, soslayando cuestiones como la sinonimia, homonimia o polisemia, que inciden de modo evidente en la escasa precisión en la recuperación de información jurídica. El significado exacto de un término jurídico vendrá determinado en muchas ocasiones por el contexto en el que éste se emplee.

Convenimos con Schweighofer (30), en que una gran oportunidad para mejorar los SRI legal radica en el estudio del lenguaje jurídico, ofreciendo ayuda a los usuarios atendiendo a los significados semánticos y pragmáticos. Argumentando la necesidad de emplear técnicas de desambiguación, puso en marcha el proyecto KONTERM entre los años 1992-1996 desde el Instituto de Derecho Internacional Público, de la Universidad de Viena. El objetivo consistía en proporcionar una aplicación híbrida de métodos de representación del conocimiento legal de apoyo a los juristas en la gestión de grandes cantidades de información jurídica contenida en documentos en lenguaje natural.

El proyecto KONTERM ha continuado sirviendo de base para trabajos punteros en el ámbito de la RI jurídica. Merkl, Schweighofer y Winiwarter (31) empleando como corpus 75 sentencias y una lista de unos 250 descriptores tomados de la base de datos europea CELEX, hacen uso de redes neuronales en combinación con técnicas estadísticas para la construcción de un tesauro legal por medio de un análisis connotativo y la creación, igualmente, en el nivel de los documentos, de un espacio en el que se clasifican los mismos por medio de criterios de similaridad.

Pietrosanti y Graziadio (32) se decantan por el uso de técnicas de PLN como recurso clave para solventar las limitaciones que presentan los SRI jurídica. Estos inconvenientes derivan, en primer lugar, de la ausencia de información contextual, ideal para ser combinada con conceptos que mejoren de modo esencial la precisión en la recuperación y, en segundo lugar, de lo que denominan «problema económico» que atañe a la tarea manual de análisis de contenido e indización centrada en la extracción de información auxiliar apropiada para codificar diversos aspectos relevantes del contenido del texto, fundamentalmente referencias cruzadas y conceptos que pertenecen al esquema de clasificación, tesauro, etc. Esta ardua actividad se halla asimismo expuesta a un sustancial grado de subjetividad y sometida a posibles errores derivados de la naturaleza humana de la tarea y del ingente volumen de información a tratar.

Señalan estos investigadores la imposibilidad en estos momentos de poder emplear técnicas de PLN aplicadas al dominio jurídico de forma global para el procesa-

miento de textos a gran escala, pero, sin embargo, sí resulta viable utilizar técnicas de PLN «superficial» aplicadas a áreas específicas. Las pautas de trabajo se fijaron, por tanto, en tres aspectos: énfasis en el modelo de caracterización contextual de palabras y conceptos para mejorar la precisión de la RI, uso de herramientas semi-automáticas para la adquisición de componentes de información legal y experimentación del potencial de las técnicas de búsqueda e indización para el desarrollo de apoyo avanzado en la redacción de documentos.

Pietrosanti y Graziadio diseñaron y desarrollaron, como paso inicial, en 1994 el prototipo *NaviLex: Navigation on Lex*, dirigido a usuarios expertos que trabajaban en el Departamento legislativo del Banco de Italia como encargados de la redacción y mantenimiento de las disposiciones legales bancarias. Emplearon inicialmente Toolbook™ para el tratamiento de legislación bancaria, para evolucionar hacia un entorno en Visual Basic basado en Fulcrum™ incluyendo, además, una base de datos fiscal.

El lenguaje jurídico presenta, por un lado, aspectos favorables, susceptibles de grandes oportunidades para el uso de técnicas basadas en el PLN que dimanan de la naturaleza de los textos jurídicos, fundamentalmente la disposición tipográfica, las expresiones formales y recurrentes y el empleo de un vocabulario especializado, características ya señaladas anteriormente, que han llevado a Pitrosanti y Graziadio a hablar de un «sublenguaje estructural», definido tácitamente mediante un tipo especial de estructuras regulares que implícitamente quedan definidas bajo la expresión «estilo jurídico». Por otro lado, presenta grandes retos imbricados en la reconocida complejidad de la documentación jurídica, textos en los que predominan frases largas y ambiguas con multitud de referencias cruzadas y anafóricas.

Conviene reconocer algunos esfuerzos importantes que emanan de proyectos de I+D de la Unión Europea, como *NOMOS* (*Knowledge Acquisition for Normative Reasoning Systems*) proyecto del programa Esprit II, del que arrancó Navilex, *RENOS* (*Reduction of Noise and Silence in Full Text Retrieval Systems for Legal Text*), proyecto en el que se combinan medios estadísticos con análisis morfológico y lingüístico para implementar un módulo semiautomático para la identificación de términos legales, o *COBALT* (*Construction, augmentation and use of Knowledge bases from natural language documents*), dedicado a textos de carácter financiero y en el que se observaron posibilidades claras para el ámbito jurídico que permitieran explotar un modelo híbrido en el que se aplicaran tecnologías del PLN y de IA.

Son numerosos otra serie de proyectos centrados más en la descripción automática o semi-automática que en la RI y cuyas ventajas constatadas, en estos momentos, aún no pueden ser trasladadas a bases de datos extensas, entre los más citados pueden figurar *FLEXICON*, *SALOMON*, o el mencionado en el epígrafe anterior *HYPO*; asimismo, consideramos sugestivo el proyecto de Yeap (33) en el que aplica análisis semántico al Derecho de familia llamando la atención sobre su aplicación al lenguaje oral recogido en informes verbales llevados a cabo en este tipo de procesos.

Por último, uno de los proyectos de investigación en curso en el Istituto per la Documentazione Giuridica italiano que, sin duda, profundizará en el conocimiento de los documentos legales, lleva por título: «*Analisi strutturale e semantica dei documenti normativi, giudiziari e amministrativi*», y su objetivo no es otro que individualizar una representación lo más rica posible desde el punto de vista semántico y estructural de este tipo de documentos.

6 Lenguajes de marcado en la gestión de la documentación jurídica

La estructura de los documentos jurídicos se ha aprovechado en el diseño de los tradicionales SRI para tratar de mejorar su efectividad mediante la creación de diversos campos de búsqueda que respondiesen a las peculiaridades de este tipo de documentación. Schweighofer (30) cita como ejemplo más extremo de esta aproximación la base de datos CELEX de la Comunidad Europea diseñada con 10 índices principales y 80 campos.

Las marcas se han utilizado de un modo simple como ayuda en el procesado informático; en el sentido del ejemplo anterior, permiten la separación de campos de una base de datos, con el tiempo dieron paso a sistemas más complejos, como los procesadores de texto y, con pretensiones más ambiciosas, aparecieron los lenguajes de marcas que posibilitan el uso del marcado de los documentos con fines documentales.

Un lenguaje de marcado viene determinado por un conjunto de reglas que permiten fijar el tipo de marcas que se utilizarán, las marcas permitidas en cada una de las partes del documento, la forma de distinguir el texto del documento de las marcas y, por último, la gramática y sintaxis que rigen el empleo de las mismas.

A finales de los años sesenta del siglo anterior, tres investigadores contratados por IBM, Charles Goldfarb, Ed Mosher y Ray Lorie, recibieron el encargo de diseñar un sistema de edición, almacenamiento, búsqueda y gestión de documentos legales construyendo un sistema de formateo estructural al que, en un principio, se denominó GML. En 1986 se convertiría en un estándar, SGML *(Standard Generalized Markup Language),* metalenguaje de etiquetado de texto convertido en norma ISO 8879.

A pesar de la enorme potencialidad que ofrece SGML su uso ha quedado relegado a la publicación, gestión e intercambio de documentos electrónicos en grandes instituciones. Ha sido HTML, una aplicación del lenguaje SGML desarrollada inicialmente por Tim Berners-Lee, que indica como se han de codificar los documentos para su distribución en la web, el lenguaje con mayor presencia en la red y todo ello a pesar de sus notorias limitaciones.

Sin embargo, creemos que conviene destacar, por su trascendencia en la RI jurídica, cómo gracias al fenómeno web, el desarrollo del hipertexto (aunque su invención sea anterior al nacimiento de esta malla mundial multimedia), se erigió en una tecnología útil y nos atrevemos a afirmar que imprescindible para dar solución adecuada al problema de las relaciones entre los documentos jurídicos.

En 1996 comenzó su andadura XML *(eXtensible Markup Language)* respaldado por el W3C *(World Wide Web Consortium)* con la intención de diseñar un lenguaje de marcas capaz de integrar la simplicidad de HTML con la potencia de SGML. Constituye, por tanto, una versión abreviada de SGML viable en el entorno web, elimina parte de las operaciones sintácticas de SGML, pero proporciona estructura a la información.

Teniendo en cuenta las peculiaridades que caracterizan a los documentos jurídicos y a las bases de datos jurídicas, resulta incuestionable reconocer, como queda establecido en los trabajos de Nogales y Arellano (34) o Nogales *et al.* (35), las enormes ventajas que aporta la aplicación de la tecnología web a la documentación jurídica, es decir, la utilización de lenguajes de marcas como medio de difusión de este tipo de información en detrimento de las tradicionales bases de datos.

Algunos de los argumentos más evidentes para su aplicación pueden cifrarse en la

masiva implantación de la tecnología web, la posibilidad de aplicación a documentos de cualquier tamaño e integrando formatos diversos, la idoneidad para gestionar referencias internas y externas al propio documento en forma de hipervínculos, la posibilidad de incorporar motores de búsqueda que puedan proceder a la indización de los documentos atendiendo a las etiquetas del correspondiente lenguaje de marcas y distribución web o por medio de soportes ópticos diversos.

En nuestro país son varias las recopilaciones de documentos legislativos y jurisprudenciales que hacen uso del lenguaje de marcado con su correspondiente tratamiento hipertextual desde hace varios años: *Norm@civil* (civil.udg.es/normacivil/), proyecto del área de Derecho civil de la Universidad de Gerona que incluye legislación y jurisprudencia, *Noticias jurídicas* (www.juridicas.com) portal de Editorial Bosch que contiene normativa concordada, artículos doctrinales o guía judicial, el portal IUStel (http://www.iustel.com/), algunos boletines oficiales autonómicos, etc.

Consideramos especialmente relevantes los trabajos llevados a cabo por el CETL (Center for Electronic Text in the Law) de la Escuela de Derecho de la Universidad de Cincinnati creado con la intención de trabajar con recursos digitales jurídicos, investigar las mejores posibilidades para la representación digital de los textos legales y, por último, publicar en Internet materiales seleccionados relacionados con el Derecho. Nos interesa de modo singular la investigación centrada en el desarrollo de TEI (*Text Encoding Initiative*), modelo de metadatos basado en SGML, en los documentos legislativos. El CETL fue el productor de dos bases de datos distribuidas vía web: *Diana*, de derechos humanos y *Securities Lawyer's Deskbook.*

Quizá el proyecto más ambicioso lo constituya el *Corpus Legis project* (www.juridicum.su.se/iri/corpus/) desarrollado por el Law and Informatics Research Institute (IRI), de la Facultad de Derecho y el Departamento de Lingüística Computacional de la Universidad de Estocolmo, nacido con el fin de elaborar recursos de textos legales electrónicos para la realización de estudios jurídico-lingüísticos.

Sjöberg (36) señala entre los principales objetivos del proyecto:

— Dar respuesta a la necesidad manifiesta de mejorar los métodos de recuperación de información legal.
— Servir de apoyo a las crecientes investigaciones sobre información jurídica emprendidas desde distintas disciplinas (Tecnologías de la información, Lingüística o Derecho).
— Ofrecer una solución viable al rápido crecimiento de la información legal, a la internacionalización de la misma y la necesidad general de la armonización europea como resultado del Derecho comunitario.

Para lograr estos objetivos el punto de partida es la consideración de SGML como la herramienta adecuada para expresar estructuras paralelas, multidimensionales y una red compleja de relaciones entre las mismas, además, es posible diseñar DTDs para documentos legales y, por último, SGML es un medio para mejorar los métodos de recuperación de información jurídica.

El *Corpus Legis Project* ha generado el *Corpus Legis System*, que comprende, además del corpus textual legal en formato SGML y en otros formatos, otros ficheros asociados (declaraciones SGML, DTDs, etc.). El sistema se compone de tres aplicaciones: Panorama (navegador), PRISE (aplicación de RI) y un sistema de gestión y

publicación electrónica para el que se ha empleado Dataware II: The *Corpus Legis Application Demonstrator*.

En noviembre de 1998 se constituyó el *Legal XML* (www.legalxml.org/) *c*omo una organización sin ánimo de lucro en la que participan tanto entidades públicas como privadas, su intención es desarrollar estándares técnicos abiertos, no propietarios, para su aplicación en la documentación jurídica y en aplicaciones relacionadas. Se divide en diversos grupos de trabajo, atendiendo a los tipos específicos de documentos jurídicos. Los trabajos realizados ya han proporcionado resultados de interés, así el National Center for State Courts (www.ncsconline.org/) y el host Lexis-Nexis como patrocinador, han publicado *Concepts for a judicial XML Mamespace & Data Tag Dictionary,* el propósito del informe es definir una DTD de XML para ser empleada en los tribunales.

Afortunadamente, en nuestro país se están llevando a cabo investigaciones relevantes sobre la aplicación de los lenguajes de marcas a la documentación jurídica. Poseen especial interés los proyectos desarrollados por el Departamento de Biblioteconomía y Documentación de la Universidad Carlos III de Madrid, que atienden, en uno de los casos a la normativa de Mercosur y, en el segundo, a las Disposiciones Generales publicadas en el Boletín Oficial de la Comunidad de Madrid.

El proceso para la versión hipertextual del *Código del Mercosur* desde 1991 hasta 1998, partió de la concepción del corpus en papel como un bloque hipertextual marcado con HTML, con una estructura de ficheros y directorios de fácil manejo, y con expresión de hiperenlaces entre las distintas normas relacionadas. El proyecto fue pionero en España en el uso de HTML para marcar documentación legislativa y expresar las relaciones contenidas en la misma (37).

El segundo proyecto nació con el objetivo de desarrollar una base de datos hipertextual accesible en línea y en soporte óptico de las Disposiciones Generales publicadas en el *Boletín Oficial de la Comunidad de Madrid*. Inicialmente pensado para usar XML con una DTD propia, se decidió, a la espera de que las diversas tecnologías XML se conviertan en estándares aceptados *de facto*, emplear HTML, si bien se ha enriquecido empleando *clases* aplicadas al etiquetado de ciertas partes de los documentos, lo que permitirá, de un lado, una futura traducción de las etiquetas utilizadas a XML y, de otro, la aplicación de hojas de estilo (34, 37).

En la Universidad de Valladolid, en este caso desde el Departamento de Informática, en concreto el Grupo de investigación de Recuperación de información y Bibliotecas digitales, trabaja en el uso de XML y estándares asociados XLink, XPointer y XPath en la documentación jurídica. Martínez *et al*. (38) proponen explotar las relaciones entre textos legislativos de forma que se puedan realizar consultas sobre éstas y presentan la posibilidad de generar automáticamente documentos de la versión definitiva de la disposición legal. La consulta de relaciones pasa de ser un proceso de navegación a una consulta en una base de datos XML.

En el entorno privado de nuestro país, la empresa ISOCO ha puesto en marcha el sistema Tirant on Line, sistema de publicación electrónica aplicado al campo jurídico. La tecnología empleada se basa en la combinación de técnicas de inteligencia artificial para tratamiento de textos, técnicas avanzadas de búsqueda, seguridad en las conexiones, autentificación de usuarios y XML. El formato XML se usa para introducir y clasificar de forma automática los documentos en el sistema, permitiendo una conveniente separación entre la capa de datos, la lógica de la aplicación y el sistema de visualización.

Los propios Pietrosanti y Graziadio (32) se plantean en su trabajo una posible adopción de XML para la representación de la información de los documentos jurídicos en el sistema NaviLex, a la luz de la capacidad de XML para el intercambio, tratamiento reutilización de los documentos y posible marco de representación de técnicas estándar apropiadas para la interrogación en sistemas de información estructurada.

En estos momentos se halla en curso un proyecto del Istituto per la Documentazione Giuridica: «Metodologie di categorizzazione, descrizione strutturale e analisi semantica di documenti giuridici per l'accesso all'informazione in rete», que persigue el empleo de XML y protocolos como XSL, XLink o XPointer para la definición de la estructura formal y funcional de los documentos jurídicos legislativos, jurisprudenciales y doctrinales.

7 Reflexiones

Las particularidades de los documentos jurídicos conforman sistemas de recuperación con unas características singulares. Sin embargo, y a pesar de constituir un sector de gran relevancia en la industria de los contenidos, podemos calificar de muy escasas las investigaciones dirigidas al estudio de la recuperación de información jurídica en nuestro país, con la excepción de los recientes trabajos centrados en los lenguajes de marcas aplicados a la documentación legislativa y jurisprudencial.

Fuera de nuestras fronteras se evidencia, sin embargo, un gran interés por la recuperación de información jurídica desde la perspectiva de la ciencia de la Documentación, inclinación corroborada por la existencia de instituciones y publicaciones científicas especializadas en este área.

El campo jurídico ha sido pionero en la recuperación en línea del texto completo de los documentos, de ahí que las investigaciones evaluativas sobre los SRI legal se remonten a los años sesenta, prevaleciendo en esta línea, hasta el momento, una tendencia adscrita al modelo tradicional de evaluación centrado en el rendimiento de los sistemas en términos de exhaustividad y precisión. Parece recomendable, no obstante, acometer investigaciones empíricas en la línea cognitiva, atendiendo al comportamiento de los usuarios en el proceso de búsqueda y recuperación, aplicables al diseño de SRI jurídica.

Las diversas tentativas llevadas a cabo en el terreno de los sistemas legales expertos, sistemas basados en conocimiento legal, así como técnicas basadas en PLN muestran las limitaciones de este tipo de prototipos o modelos ideales contraidas por el escaso volumen de información tratada, así como las evidentes restricciones de escasa proporcionalidad uso/coste. Sin embargo, creemos que los esfuerzos han conducido a una mayor madurez en la aplicación de técnicas derivadas de la IA y a un análisis profundo de la terminología jurídica desde el punto de vista semántico y estructural.

Por último, sin obviar los altos costos de la codificación de textos y el lento proceso en la estandarización de las distintas tecnologías de la familia XML, parece probable que se imponga entre los distintos lenguajes de etiquetado disponibles y resulta, igualmente palmario, que la documentación jurídica, por su carácter fuertemente estructurado y por las complejas relaciones que precisa expresar entre los distintos documentos, se adapta de modo óptimo a esta tecnología.

8 Bibliografía

1. BING, J. Legal text retrieval and information services. En Kent, A. y Lancour, H. (eds.). *Encyclopaedia of Library and Information Science*. New York: Marcel Deker, 1991, vol. 48, supl. 11, p. 219-254.
2. MORENO DE LA FUENTE, I. Documentación jurídica extranjera. En Maciá, M. (ed.) *Manual de documentación jurídica.* Madrid: Síntesis, 1998, p. 367-446.
3. MACIÁ, M. *La documentación de la Unión Europea.* Madrid: Síntesis, 1996, p. 203-213.
4. TENOPIR, C. y RO, J .S. *Full text databases*. New York: Greenwood Press, 1990, p. 75-78.
5. BLAIR, D. C. STAIRS redux: thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science,* 1996, vol. 41, n. 1, p. 4-22.
6. BLAIR, D. C. y MARON, M. E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 1985, vol. 28, n. 3, p. 289-299.
7. BLAIR, D. C. y MARON, M. E. 1990. Full text information retrieval: further analysis and clarification. *Information Processing and Management*, 1990, vol. 26, n. 3, p. 437-447.
8. BERRING, R. C. Full-text databases and legal research: backing into the future. *High Technology Law Journal*, 1986, vol. 27, n. 1, p. 27-60.
9. DABNEY, D. P. The curse of Thames: an analysis of full-text legal document retrieval. *Law Library Journal*, 1986, vol. 78, n. 5, p. 5-40.
10. GERSON, K. Evaluating legal information retrieval systems: how do the ranked-retrieval methods of WESTLAW and LEXIS measure up?. *Legal Reference Services Quaterly*, 1999, vol. 17, n. 4, p. 53-67.
11. BURSON, S. F. A reconstruction of Thamus: comments on the evaluation of legal information retrieval systems. *Law Library Journal*, 1987, vol. 79, n. 1, p. 133-143.
12. NEWMAN, M. S. Evaluation criteria and quality control for legal knowledge systems on the Internet: a case study. *Law-library-Journal*, 1999, vol. 91, n.1, p. 9-27.
13. BARMAKIAN, D. Better search engines for law. *Law Library Journal*, 2000, vol. 92, n. 4, p. 399-438.
14. DEMPSEY, B. L.; VREELAND, R. C.; SUMNER, R. J. y YANG, K. Design and empirical evaluation of search software for legal professionals on the WWW. *Information Processing and Management*, 2000, vol. 36, p. 253-273.
15. HANFT, J. K. A model for legal research in the electronic age. *Legal Reference Services Quaterly*, 1999, vol. 17, n. 3, p. 77-83.
16. WILSON, T. D. Models in information behaviour research. *Journal of Documentation*, 1999, vol. 35, n. 3, p. 249- 270.
17. LECKIE, G. J.; PETTIGREW, K. E. y SYLVAIN, C. Modeling the information seeking of professionals: a general model derived from research on engineers, health care professionals and lawyers. *Library Quarterly*, 1996, vol. 66, n. 2, p. 161-193.
18. KUHLTHAU, C. C. y TAMA, S. L. Information search process of lawyers: a call for «just for me» information services. *Journal of Documentation*, 2001, vol. 57, n. 1, p. 25-43.
19. WILKINSON, M. A. Information sources used by lawyers in problem-solving: an empirical exploration. *Library and Information Science Research*, 2001, vol. 23, p. 257-276.
20. ERDELEZ, S. y O'HARE, S. Legal informatics: application of information technology in law. En *Annual Review of Information Science and Technology (ARIST).* Medford, N. J: American Society for Information Science, 1997, p. 367-402.
21. CURRAN, K. y HIGGINS, L. A legal retrieval information system. [En línea]. *Journal of Information, Law and Technology (JILT)*, 2000, n. 3. <http://elj.warwick.ac.uk/jilt/00-3/curran.html>. [Consultado: 31/03/2002].
22. BENCH-CAPON, T. J. M. Knowledge-based systems in the legal domain. En Kent, A. y Lancour, H. (eds.). *Encyclopaedia of Library and Information Science*. New York: Marcel Deker, 1996, vol. 54, supl. 20, p. 269-291.

23. GALINDO AYUDA, F. y LASALA CALLEJA, P. Metodología para el desarrollo de sistemas jurídicos de inteligencia artificial: el prototipo ARPO-2 como ejemplo. *Scire: Representación y Organización del conocimiento*, 1995, vol. 1, n. 2, p. 73-103.
24. MATTHIJSSEN, L. An architecture for legal information retrieval using task models. *Information & Communications Technology Law*, 1997, vol. 6, n. 3, p. 229-248.
25. MATTHIJSSEN, L. A task-based interface to legal databases. *Artificial Intelligence and Law,* 1998, vol. 8, p. 81-103.
26. MATTHIJSSEN, L. *Interfacing between lawyers and computers: an architecture for knowledge-based interfaces to legal databases.* The Hage: Kluwer Law International, 1999.
27. PÉREZ-CARBALLO, J. y STRZALKOWSKI, T. Natural language information retrieval: progress report. *Information Processing and Management*, 2000, vol. 36, p. 155-178.
28. SPARCK JONES, K. What is the role of NLP in text retrieval?. En Strzalkowski, T. (ed.). *Natural Language information retrieval*. Dordrecht: Kluwer Academic Publishers, 1999, p. 1-24.
29. ARAMPATZIS, A.; VAN DER WEIDE, T. P.; VAN BOMMEL, P. y KOSTER, C. H. A. Linguistically motivated information retrieval. En Kent, A. y Lancour, H. (eds.). *Encyclopaedia of Library and Information Science*. New York: Marcel Deker, 2001, vol. 69, supl. 32, p. 201-222.
30. SCHWEIGHOFER, E. The revolution in legal information retrieval or the Empire strikes back. [En línea]. *Journal of Information, Law and Technology (JILT)*, 1999, n. 1. <http://elj.warwick.ac.uk/jilt/99-1/schweigh.html>. [Consultado:26/02/2002]
31. MERKL, D.; SCHWEIGHOFER, E. y WINIWARTER, W. Exploratory analysis of concept and document spaces with connectionist networks. *Artificial Intelligence and Law*, 1999, vol. 7, p. 185-209.
32. PIETROSANTI, E. y GRAZIADIO, B. Advanced techniques for legal document processing and retrieval. *Artificial Intelligence and Law*, 1999, vol. 7, p. 341-361.
33. YEAP, W. K. Computing rich semantic models of text in legal domains. *Information & Communications Technology Law*, 1998, vol. 7, n. 2, pp. 135-145.
34. NOGALES FLORES, J. T. y ARELLANO, M. C. La organización hipertextual de textos legislativos con HTML y XML: una necesidad y las soluciones de presente y futuro. En *VII Jornadas españolas de Documentación*. Bilbao: Universidad del País Vasco, 2000, p. 179-188.
35. NOGALES FLORES, J. T. *et al*. La difusión de los textos legislativos en Internet haciendo uso de los lenguajes de marcado HTML y XML. [CD-ROM]. En *III Conferencia Internacional de Derecho e Informática de la Habana. Info 2000*. La Habana: Ministerio de la Informática y las Comunicaciones, 2000.
36. SJÖBERG, C. M. *Critical factors in legal document management: a study of standardised markup languages.* Stockholm: Jure AB, 1998.
37. NOGALES FLORES, J. T. *et al.* Un repertorio legislativo hipertextual mediante marcado de texto: las Disposiciones Generales del Boletín Oficial de la Comunidad de Madrid. [En línea]. En I Jornadas españolas de bibliotecas digitales. Valladolid: Departamento de Informática, Universidad de Valladolid, 2000. <http://gaia.dcs.fi.uva.es/~jbidi2000/> [Consultado: 30/04/2002]
38. MARTÍNEZ, M. M. *et al.* Explotación dinámica de relaciones en las bibliotecas digitales: aplicación a una biblioteca jurídica. [En línea]. En II *Jornadas españolas de bibliotecas digitales,* 2001. <http://gaia.dcs.fi.uva.es/~jbidi2001/comunicaciones/07_jbidi01.pdf>. [Consultado: 30/04/2002]