Las herramientas terminológicas en los sistemas de información jurídica

María Luisa Alvite Díez
Universidad de León (España)

0.1. Resumen

Se analizan los instrumentos terminológicos integrados en las interfaces de las bases de datos jurídicas Aranzadi, *Iberlex* y La Ley, seleccionadas por su alta implantación en unidades de información. Se observa un abanico de soluciones que abarca desde las simples listas de palabras clave hasta los tesauros de estructura navegacional. Se aprecian diferencias sustanciales entre las bases de datos legislativas y las jurisprudenciales. Aunque constituyen instrumentos potentes para la recuperación de información, estos entornos automatizados adolecen de una escasa normalización de la terminología jurídica y de usos tradicionales parcialmente obsoletos.

Palabras clave: Sistemas de información. Información jurídica. Lenguajes controlados. Indización. Terminología. Recuperación de la información.

0.2. Abstract

The different terminological tools integrated in the most popular Spanish legal databases interfaces (Aranzadi, *Iberlex* and La Ley) are analyzed. A wide scope of them is provided, ranging from simple keywords to thesauri with navigational structure. Substantial differences are also observed between normative and juridical databases. Although they are powerful instruments for information retrieval, their terminological standardization should be enhanced and rigorous scientific criteria should be preferred to procedures based in former legal uses.

Keywords: Legal information. Controlled indexing languages. Thesauri. Indexing. Terminology. Information retrieval.

1. Introducción

La naturaleza compleja de la documentación jurídica determina las características peculiares de los sistemas de recuperación de información legal: gestión de un enorme volumen de información, unidades documentales recogidas mayoritariamente a texto completo —dado que la información parlamentaria, legisla-

tiva y jurisprudencial ha de ser almacenada de forma íntegra para responder a las necesidades reales de los usuarios—, exigencia de exhaustividad en los corpus legales almacenados como aval de garantía jurídica, resolución de las complejas relaciones entre documentos —relaciones normativas entre documentos anteriores y posteriores que complementan, modifican, amplían, derogan o interpretan una disposición—, interconexión entre la documentación normativa, judicial y bibliográfica, y actualización permanente. Las bases de datos jurídicas en nuestro país responden, fundamentalmente, al modelo de recuperación tradicional asentado en el uso de un fichero lineal, un fichero inverso y un lenguaje de recuperación basado en el álgebra de Boole; si bien La Ley y Aranzadi han comenzado tímidamente a introducir la presentación de resultados ordenados por relevancia. Por tanto, la interacción entre usuario y sistema —elemento básico en la recuperación de la información— se efectúa mediante la realización a la base de datos de preguntas cuyos términos han de coincidir exactamente con los que constituyen los puntos de acceso de sus registros.

En el caso de las bases de datos *legislativas*, los sistemas de recuperación han de adaptarse adecuadamente a las peculiaridades y complejidad de los documentos normativos. Ello requiere la estructuración de los registros en una serie de campos, y en ocasiones, subcampos, que se acomoden a la tipología concreta tratada. La unidad documental es, generalmente, la norma. Además del texto completo de la disposición, un tratamiento documental previo asegura la recuperación de la información y garantiza la seguridad jurídica, mediante la normalización conceptual y la resolución de las interconexiones entre las normas.

En los sistemas *jurisprudenciales* la unidad documental es, habitualmente, la resolución. Además del texto completo, con las particularidades adoptadas por cada sistema, los registros incorporan el análisis documental, dirigido fundamentalmente a la asignación de descriptores y al tratamiento de las conexiones con la normativa y la jurisprudencia tratada en la correspondiente sentencia. Las bases de datos habrán de adaptarse a las características de la documentación judicial, articulando los registros en una serie de campos y subcampos ajustados a la correspondiente tipología del documento.

Los sistemas tienen en cuenta la naturaleza de la información procesada y tratan de adaptarse a las necesidades del usuario. Si bien la información jurídica forma parte de la vida cotidiana o profesional de cualquier ciudadano, las bases de datos evidencian un tratamiento y diseño orientado en gran medida a los profesionales del Derecho. En este sentido, tal vez *Iberlex* sea la más alejada del estándar jurídico. El estudio más relevante sobre el comportamiento de los usuarios especialistas durante el proceso de búsqueda del que se dispone es el realizado por Kuhlthau y Tama (2001). Estos autores señalan una clara tendencia de los usuarios a concentrarse en aprender lo mínimamente imprescindible sobre el sistema

para obtener resultados, y, seguidamente, a concentrar su esfuerzo en su lectura y valoración, resultados que coinciden con otros trabajos de evaluación de interfaces. Constatan, asimismo, la estrecha relación que existe entre los sistemas y las necesidades, habilidades y métodos de trabajo de los profesionales del mundo del Derecho. Las interfaces de las bases de datos jurídicas confirman lo señalado. En la recuperación de información legislativa se produce una combinación de información conceptual e información factual, que los sistemas resuelven con la introducción de campos de búsqueda suficientes para dar respuesta adecuada a consultas por fechas, rango, número o fecha de boletín, título, texto, voces o índice, etc. Los sistemas jurisprudenciales pretenden la mayor integración posible con los hábitos de trabajo y la práctica jurídica cotidiana de los profesionales del Derecho y atienden a la naturaleza de la información tratada, sea ésta factual o conceptual. Existe una gran coincidencia en los campos de consulta proporcionados por los sistemas: resolución, jurisdicción, ponente, disposiciones estudiadas, texto, voces, etc.

Los usuarios a los que se dirigen estos sistemas cifran sus expectativas —creemos— en lograr la máxima precisión en las respuestas, y ésta se alcanza cuando la interrogación se realiza sobre campos factuales: fechas de emisión, de publicación, número de la disposición, etc. Se trata de búsquedas extraordinariamente efectivas sobre las que se pueden emitir juicios de relevancia *a priori*, que no suelen provocar problemas de recuperación y que están disponibles a través de muchos sistemas gratuitos dispuestos en la web, si bien es cierto que con una menor cobertura temporal. Sin minusvalorar la eficacia de estas consultas, creemos que la capacidad real de sistemas de la envergadura de los aquí estudiados se determina, más bien, por su capacidad de recuperar la información temática o conceptual tratada. Los puntos de acceso por materias son fundamentales y su estudio creemos que es central en el diseño de sistemas de información; si bien es cierto que no podemos constreñir la recuperación temática a la proporcionada por los puntos de acceso de materias desdeñando otros puntos de acceso descriptivos —título, resumen, texto, disposiciones estudiadas—. En el caso de las bases de datos jurisprudenciales, estos puntos de acceso suponen una ayuda inestimable para conocer el articulado estudiado y la línea jurisprudencial seguida en una resolución. Tampoco se debe soslayar la trascendencia de los enlaces hipertextuales en la recuperación de información, en cuyo uso, por otra parte, fueron pioneros estos sistemas.

2. La recuperación temática: lenguaje natural v. lenguaje controlado

Recuperación de información y organización del conocimiento son conceptos imbricados de modo indisoluble. El propio Mooers, a quien se atribuye la introducción del término "Recuperación de Información" acuñó el de "Lenguajes de

Recuperación de Información" para referirse de modo genérico a aquellos sistemas artificiales de signos normalizados —clasificaciones, palabras clave, puntos de acceso temáticos, etc.— que hacen posible la representación del contenido de los documentos facilitando la recuperación de información (Salton, 1987). La recuperación del texto completo de los documentos marca según Hjørland y Nielsen (2001) la lucha y el paso final en el desarrollo de los puntos de acceso por materias. En este contexto, al que pertenecen los sistemas jurídicos, cada palabra o combinación de palabras se convierten en potenciales puntos de acceso temáticos. Tenopir y Ro (1990, p. 197-217) abogan por la importancia de incluir algún tipo de análisis intelectual en campos que denominan de valor añadido y que complementan las búsquedas efectuadas sobre el texto completo. Insisten en señalar como principal inconveniente en las búsquedas en lenguaje natural sobre bases de datos a texto completo la inconsistencia en el uso de las palabras y sus formas propia del lenguaje y, a la vez, reclaman más investigaciones que demuestren la eficacia de la indización en lenguaje controlado —para confirmar si han de continuar formando parte de los sistemas a texto completo— y también para determinar cuál es el tipo de lenguaje más apropiado para cada disciplina. Rowley (1994) establece cuatro eras en el debate sobre la indización en lenguaje natural o controlado, concluyendo que de las investigaciones se infiere que el lenguaje natural y controlado han de usarse conjuntamente. El propio Hjørland defiende la necesidad de profesionales de la información argumentando que los documentos en sí mismos pueden no incluir implícitamente la materia de la que tratan o bien hacerlo de modo erróneo. El trabajo señalado de Tenopir y Ro sobre sistemas a texto completo adelanta líneas de investigación que han resultado fructíferas más de una década después: el desarrollo del procesamiento en lenguaje natural, la indización automática, el uso del hipertexto, las evaluaciones sobre el rendimiento de sistemas —cuyo hito fundamental está sin duda representado por los encuentros Text Retrieval Conference (TREC)—, y una tendencia que creemos afecta de modo específico a la documentación jurídica, el valor de los términos que aparecen en el texto según la zona del mismo en la que aparecen. Éste es precisamente uno de los aspectos explotados en los proyectos en los que se aplica Standard Generalized Markup Language (SGML) y eXtensible Markup Language (XML) al tratamiento y recuperación semántica de la información legal.

En el ámbito internacional, el campo jurídico ha sido pionero en la recuperación en línea del texto completo. El primer sistema implementado para la recuperación de una gran colección legislativa fue presentado en el año 1960 por el Health Law Center de la Universidad de Pittsburg y utilizaba ya una especie de tesauro concebido como ayuda en el proceso de búsqueda. Lancaster (1995, p. 184) se refiere a este tesauro como una simple compilación de palabras con significados similares, con una estructura semejante a la del *Roget´s Thesaurus* que,

sin embargo, suponía una estimable ayuda de búsqueda, pues evitaba al usuario tener que pensar en todas las palabras que pudieran expresar una idea concreta. Se trataba, por tanto, de un tesauro simple que en palabras de Lancaster *carecía de un significativo grado de "estructura"*. Desde entonces han sido numerosos los experimentos evaluativos a los que han sido sometidos los diversos sistemas de recuperación legal, hallándose conclusiones dispares sobre su rendimiento en términos de exhaustividad y precisión. Se constata, sin embargo, de modo unánime, el esfuerzo que el usuario necesita hacer en la búsqueda en texto libre para predecir las palabras y frases necesarias en orden a recuperar documentos relevantes y evitar los irrelevantes (Alvite Díez, 2003).

Para intuir la complejidad que supone la predicción de los términos de consulta, se ha de tener presente que en el lenguaje jurídico, atendiendo a la procedencia terminológica, podemos encontrar tanto vocabulario estrictamente jurídico, como lenguaje técnico o especializado de otras ramas del saber, así como, por supuesto, lenguaje común. El lenguaje controlado elaborado para la indización y recuperación de la documentación jurídica se enfrenta a la dificultad de atender a las tres parcelas. Como señala López-Muñiz Goñi (1984, p. 74-77) no cabe la expresión de una ley o una sentencia en puro lenguaje de técnica jurídica. En nuestro país las bases de datos jurídicas, tal vez por herencia de los repertorios impresos en los que algunas de ellas tienen su origen, cuentan, generalmente, con instrumentos de control terminológico, en muchos casos asociados a un campo que recibe la denominación de "Voz" o "Voces". Ya los repertorios decimonónicos recurrían a índices de "voces" para organizar el complejo lenguaje tratado en la documentación jurídica, esto es, listas autorizadas de términos preferentes que hacen referencia a un marginal que conduce al documento. Creemos que los sistemas actuales intentan precisamente, por un lado, adaptarse al modo de trabajo de los profesionales jurídicos apegados a las listas autorizadas de entradas; y, por el otro, evitar las dificultades de la búsqueda en lenguaje natural sobre documentos de la extensión que alcanzan muchos de los registros normativos o jurisprudenciales almacenados en este tipo de bases de datos. El usuario ha de esmerarse para predecir inequívocamente los términos de búsqueda y para formular una ecuación de consulta efectiva que evite que el sistema devuelva un elevado número de documentos no relevantes.

Hjørland y Nielsen (2001) se refieren a diferentes tipos de puntos de acceso temáticos en cuanto que describen la materia de un documento dado de modos diversos: con mayor o menor exhaustividad, de forma más o menos general o específica, de modo abierto o cerrado, pensando en el valor futuro dentro de la base de datos, etc. Convenimos con estos autores en conceder el máximo valor a aquellos puntos de acceso temático que hacen posible al usuario identificar los documentos de superior relevancia, esto es, los que en la base de datos hacen más visibles a los

documentos más relevantes en detrimento de los menos relevantes. Los sistemas jurídicos que nos ocupan, como se ha señalado, disponen de varios campos para acceder a información temática y, sin embargo, todos ellos establecen en su interfaz un campo distintivo que sirve de punto de acceso al contenido mediante el uso de vocabulario controlado. Cuando se pretende recuperar información cognitiva, evitar el ruido, simplificar las ecuaciones de búsqueda y contextualizar los términos, estas bases de datos precisan la utilización de lenguaje controlado.

Nos sumamos al amplio consenso de la literatura científica en la que se considera clara la complementariedad entre el lenguaje libre y el controlado. López Alonso (2001) o Lopes (2002) presentan una recapitulación de esta controversia sobre la que Gil Urdiciain (1998a), tras una comparación del rendimiento de diferentes tesauros españoles frente al lenguaje natural en la recuperación de información, concluye señalando que el control del vocabulario es un factor determinante del éxito en el proceso de recuperación, dado que el lenguaje controlado neutraliza las deficiencias del lenguaje libre y viceversa.

3. Análisis y resultados

A continuación, se presenta un análisis de los lenguajes documentales empleados en cinco bases de datos que hemos considerado representativas del área temática que nos ocupa. De ellas tres son legislativas —*Iberlex* (Boletín Oficial del Estado), *Legislación Aranzadi/Westlaw* (Thomson) y *Repertorio de legislación la Ley* (Wolters Kluwer)— y dos de jurisprudencia —*Jurisprudencia Aranzadi/Westlaw* (Thomson) y *Jurisprudencia La Ley* (Wolters Kluwer)—. Los formatos utilizados son los suscritos por la Biblioteca de la Universidad de León en el año 2003: CD-ROM para *Iberlex*, versión web para los productos de Aranzadi y DVD para los de La Ley. Todos los productos estudiados incorporan en sus interfaces herramientas de control terminológico que auxilian al usuario en la realización de consultas temáticas y pretenden paliar el ruido del lenguaje natural favoreciendo la eficacia de la recuperación.

Abordamos el estudio con la intención de escudriñar los instrumentos terminológicos que proporcionan los sistemas señalados en orden a efectuar un análisis sobre su tipología, tamaño, nivel de precoordinación, morfología, relaciones jerárquicas, de equivalencia y asociativas, y notas de aplicación empleadas. Se atiende, en esta selección de criterios valorativos, a los reconocidos en la literatura al respecto y, particularmente, en los trabajos de Lancaster (1995), Gil Leiva y Rodríguez Muñoz (1997), Gil Urdiciain (1998b) y López Alonso (1998).

3.1. Tipología

Las interfaces de los sistemas presentan una diversidad de posibilidades de acceso, en algunos casos combinadas. Se trata fundamentalmente de las listas

de palabras clave, las listas autorizadas de descriptores y los tesauros. La *lista de palabras clave* se emplea en la base de datos *Jurisprudencia La Ley*. Se trata de una lista mayoritariamente de unitérminos no estructurados que se presentan ordenados alfabéticamente. Se compone de palabras clave conceptuales; no se incluyen entradas onomásticas ni topónimos. Su principal ventaja se cifra en la especificidad de los términos que la integran frente al grado de generalidad que presentan los descriptores del tesauro de la correspondiente base de datos.

Las *listas autorizadas de descriptores* ordenados alfabéticamente se utilizan en las tres bases de datos legislativas analizadas. Solamente el *Repertorio de legislación la Ley* permite al usuario, sin necesidad de ejecutar la consulta, conocer el número de documentos asociados a un determinado descriptor. El campo de acceso a estas listas recibe la denominación en Aranzadi y La Ley de "Voz"; *Iberlex*, por su parte, usa "Indice". La base de datos jurisprudencial de La Ley dispone de dos listas de descriptores preferentes: "Voces" y "Voces secundarias" que complementan, como veremos, el tesauro.

Respecto a los *tesauros*, a pesar de que las dos bases de datos de jurisprudencia estudiadas emplean este término, lo usan más allá del sentido estrictamente documental en que lo definen ISO y AFNOR. Si bien es cierto que, desde el punto de vista de su estructura, los tesauros de estos sistemas son vocabularios controlados y dinámicos de términos jurídicos que mantienen entre ellos relaciones jerárquicas, semánticas y asociativas; sin embargo, su cuerpo léxico, además de incluir descriptores o términos preferentes y no descriptores, incluyen toda una serie de términos simples y sintagmáticos que constituyen una auténtica red semántica en la que se refieren multitud de aspectos procedimentales, doctrinales, resolutivos, referentes al fallo de la sentencia, subjetivos, etc. En fin, se trata de una verdadera terminología especializada, tomada de la práctica jurídica habitual, con una carga informativa enorme. El usuario interactúa con ella para ampliar o restringir la búsqueda navegando por la estructura de conceptos jurídicos que le proporciona el tesauro.

Jurisprudencia Aranzadi organiza el tesauro en campos semánticos que coinciden con los órdenes jurisdiccionales: Civil, Penal, Conflictos de jurisdicción, Constitucional, Contencioso-Administrativo, Militar, Social y Sala Especial. El paso inicial consiste en señalar la "materia" para que aparezcan las familias conceptuales de ese campo ordenadas alfabéticamente. Las relaciones jerárquicas descienden hasta incluso diez niveles y se visualizan a través de las típicas carpetas y subcarpetas con las que cualquier usuario de Windows está familiarizado. El sistema indica el número de documentos indizados con el descriptor correspondiente y el número disponible en la base de datos consultada y en el conjunto del sistema. En esta versión web el listado inicial no presenta distinción tipográfica o icónica entre descriptor y no descriptor. Una vez que el usuario ha

seleccionado el término, el sistema propone la entrada preferente precedida de una flecha indicativa del reenvío.

Jurisprudencia La Ley no organiza el tesauro en campos semánticos y refiere al usuario previamente a una búsqueda por palabras clave. Desde allí, el sistema le propone las distintas familias conceptuales en las que puede localizar la información solicitada. La visualización se basa en carpetas y subcarpetas sucesivas, por lo que muestra la profundidad jerárquica de modo semejante al tesauro de Aranzadi. El tesauro de La Ley adolece de cierta rigidez, pues no permite seleccionar varios descriptores a la vez y combinarlos con el operador booleano deseado. Ofrece la posibilidad de navegar por la estructura, pero, sin embargo, para iniciar una nueva consulta obliga a abrir el menú desplegable de navegación y seleccionar la opción "colapsar todo". Es preciso aclarar que si bien Aranzadi presenta una interfaz en la que el tesauro es el instrumento ideado para las búsquedas temáticas, en el sistema la Ley el acceso al tesauro sólo está previsto desde la sub-base *Sumarios*, lo que parece sugerir que el método propuesto por el sistema para iniciar las búsquedas conceptuales es acudir a ésta. Así se entiende el mantenimiento de los campos "Voces" y "Voces secundarias" desde la sub-base *Textos* para que el usuario emplee las listas alfabéticas de descriptores, en este caso agrupadas en dos niveles: conceptos primarios y secundarios.

3.2. Tamaño

Solamente Aranzadi Legislación permite al usuario conocer el número total de términos de indización empleados: 9377. Tal vez sea más importante para el usuario final saber cuántos documentos se pueden recuperar con una determinada entrada, ayuda inestimable para conocer el grado de generalidad o especificidad del término desde el propio índice terminológico sin necesidad de ejecutar la consulta. Esta posibilidad no la ofrecen las listas de descriptores de las bases de datos legislativas Aranzadi e *Iberlex*.

3.3. Nivel de precoordinación

El análisis se ha efectuado a partir de una muestra aleatoria de cien términos seleccionados en cada base de datos de los que entran por la letra "C" en las correspondientes listas de descriptores libres y en los dos tesauros jurisprudenciales. Los resultados referentes al nivel de precoordinación se ofrecen en la tabla I. En las quinientas entradas que componen el total de la muestra se observa un predominio de descriptores bitérminos y unitérminos en las cinco bases de datos. El mayor nivel de precoordinación se observa en *Legislación Aranzadi/Westlaw,* en la que un 29% de los descriptores cuenta con cuatro términos significativos o más. El uso más moderado de descriptores sintagmáticos se corresponde con las dos bases de datos de La Ley, en las que se observa una mayor tendencia a la generalidad en el vocabulario seleccionado.

	Unitérminos	Bitérminos	Tritérminos	Cuatritérm.	Superior
Iberlex	33	38	14	6	9
Legislación Aranzadi	33	21	17	15	14
La Ley Legislación	48	31	8	6	7
Jurisprudencia Aranzadi	30	53	14	3	---
La Ley Jurisprudencia	33	55	9	3	---
Total	177	198	62	33	30

Tabla I. Nivel de precoordinación

Son notorias, igualmente, las diferencias entre las bases de datos legislativas y jurisprudenciales en cuanto al uso de descriptores sintagmáticos, cuya pretensión es conseguir una mayor precisión en la respuesta. En las bases de datos de jurisprudencia analizadas no aparece ningún descriptor con más de cuatro términos significativos y es que la propia estructura del tesauro permite contextualizar los términos. Se aprecia una excesiva precoordinación que puede ejemplificarse con alguna entrada como "Canon de compensación de precios de papel prensa" que se recoge en *Iberlex*. Estimamos excesivo el número de descriptores sintagmáticos, los denominados por López-Muñiz Goñi (1984, p. 82-82) "descriptores gruesos", que creemos se han heredado de los índices de los repertorios impresos, en los que ésta era la única fórmula prevista para la recuperación conceptual.

3.4. Morfología

La terminología recogida en las herramientas analizadas incluye entradas onomásticas, geográficas y temáticas o de materias. No se han encontrado descriptores cronológicos, si bien en ciertos casos las acotaciones temporales se resuelven con calificadores. La distribución se presenta en la tabla II.

	Sustantivos	Entidades	Topónimos
Iberlex	74	18	8
Legislación Aranzadi	53	38	9
La Ley Legislación	74	16	10
Jurisprudencia Aranzadi	98	1	1
La Ley Jurisprudencia	99	---	1
Total	398	73	29

Tabla II. Morfología

Las entradas temáticas son las de mayor importancia en la recuperación e, igualmente, las que suponen un considerable nivel de dificultad en su redacción y control. Se manifiesta un predominio de los sustantivos y, para la redacción de

términos compuestos, de los sintagmas nominales, tanto adjetivos como preposicionales. Es notorio, asimismo, el predominio de descriptores temáticos en las bases de datos de jurisprudencia debido a la propia naturaleza de la documentación que tratan. En cuanto al uso de singulares y plurales se observa el respeto al carácter cuantificable o no del término; no obstante, se advierte cierta confusión. El mayor nivel de inconsistencia e indeterminación se aprecia en el *Repertorio de legislación la Ley* donde aparecen descriptores admitidos en singular y plural como "Cajas de ahorro" y "Cajas de ahorros" o "Canon" y "Cánones".

Se observa la necesidad de introducir ciertos calificadores que ayuden al usuario a contextualizar determinados términos polisémicos, pero solo Aranzadi hace uso de ellos. En las bases de datos de jurisprudencia introduce entre paréntesis la norma bajo la que se instruyó la causa, de evidente valor aclaratorio para el usuario, como por ejemplo "Proceso Civil (LECiv/2000)". Sin embargo, algunos calificadores emplean la inversión de términos, sin respetar el orden natural que exigen las normas: "Candelas (Indemnización de)", "Competencia (Cuestiones de)", etc. En las bases de datos de legislación son abundantes los términos que, aun cumpliendo las condiciones sintagmáticas, tienen escaso poder de discriminación o no proporcionan ningún tipo de información temática, como sucede con entradas como "Caídos", "Campamentos" "Cancelación" o "Cambios".

Los lenguajes controlados empleados en las bases de datos legislativas requieren la introducción de términos onomásticos y geográficos. Es Aranzadi quien muestra un porcentaje de uso más elevado de estas entradas en la indización, circunstancia que puede obedecer al carácter meramente referencial de esta base de datos en el periodo comprendido entre los años 1930 y 1977. Parece urgente la inclusión en estos sistemas de la posibilidad de búsqueda por frase exacta que ayudaría a resolver, en alguna medida, la carencia de estas entradas.

En lo que se refiere a la formación de términos compuestos, se localizan muy diferentes estructuras sintagmáticas: sustantivo + adjetivo —"Compraventa mercantil"—; sustantivos + de + sustantivo —"Comités de empresa"—; sustantivo + Sigla —"Cláusula CIF"—; sustantivo + de + la + sustantivo —"Caducidad de la acción"—; sustantivo + de + sustantivo + adjetivo —"Cierre de fincas rústicas"—; sustantivo + de + verbo —"Causa de pedir"—; sustantivo + de + sustantivo + o + sustantivo —"Cese de negocio o industria"—; sustantivo + Voz latina —"Cláusula rebus sic stantibus"—; sustantivo + de + Nombre propio —"Canalización del Manzanares"—; sustantivo + y + sustantivo —"Canales y pantanos"—; etc. Como muestra la tabla I, son numerosos los términos compuestos, especialmente en las bases de datos legislativas. Es cierto que la indización no sólo ha de tener en cuenta los temas de un documento y la naturaleza del mismo sino también las necesidades de información del usuario potencial de la base de datos. Sin embargo, la redacción de algunos descriptores atenta contra cualquier principio normalizador,

por ejemplo: "Residencia-Residencias", "Hostelería, Café, Bares y similares", "Industrias en general", "Rioja (La)", "Suelo y ordenación urbana-urbanismo", etc. Destacamos de modo específico la abundancia de construcciones coordinadas en las cinco bases de datos. Así, en *Iberlex* se hallan descriptores como "Cadenas de prensa y radio del Movimiento", "Carburantes y combustibles" o "Carrocerías y carreterías", "Caldos y sopas" "Campamentos, albergues, colonias y marchas juveniles" —estos dos últimos recogidos igualmente en el *Repertorio de legislación la Ley*—; en Aranzadi Legislación, "Canales y pantanos"; en *Jurisprudencia Aranzadi*, "Caso fortuito y fuerza mayor" o "Constructores y contratistas"; en *Jurisprudencia La Ley*, "Concursos y oposiciones" o "Concurrencia, preferencia y prelación de créditos" —en este último caso el carácter enumerativo de la entrada obliga a introducir incluso una coma—. El empleo de la coma en la redacción de entradas para expresar la inversión en lugar de emplear la entrada directa se ha detectado en las cinco bases de datos. Sirva de ejemplo el uso de "Carnes, exportaciones" en *Iberlex*. A este respecto señala la Norma UNE 50-106-90: "la forma invertida de una frase preposicional puede introducirse como término no preferente, acompañada de un reenvío al término preferente en su orden natural."

Finalmente —y de modo palmario en las bases de datos de jurisprudencia—, prima en la redacción de los descriptores una orientación a usuarios concretos sobre los principios normalizadores. Además de los descriptores que representan conceptos jurídicos los sistemas reflejan también casos jurídicos, de ahí que se introduzcan descriptores procedimentales que contextualizan el concepto; así, por ejemplo, en Aranzadi se emplean descriptores como "No resolución de todos los puntos objeto de acusación y defensa" o "Congruencia-incongruencia".

3.5. Relaciones

Las bases de datos legislativas prescinden de establecer relaciones semánticas o asociativas en sus instrumentos terminológicos y prefieren presentar únicamente listas de términos preferentes. La Ley muestra los conceptos jurídicos ordenados alfabéticamente sin establecer relación alguna entre ellos. Como en el resto de sus índices, cada voz presenta al lado el número de documentos indizados por ese descriptor. Sorprende que versiones anteriores de Aranzadi contemplaran referencias del tipo "Véase" que se omiten en la actual versión en línea de Westlaw.es. Las relaciones de equivalencia solamente se muestran en los tesauros de las bases de datos jurisprudenciales. Sin embargo, no es posible conocer el número de descriptores y no descriptores que conforman los mismos. Atendiendo a la muestra estudiada, se aprecian diferencias notables; en el caso de Aranzadi se contabilizan 58, de las cuales 9 remiten a relaciones dobles y un caso a una triple, frente a tan solo dos relaciones propuestas por el sistema de La Ley. Las relaciones asociativas son escasas en ambos tesauros. En la muestra empleada

en el análisis sólo Aranzadi establece una relación de este tipo: "Constitución española" con "Derechos fundamentales" y "Protección jurisdiccional". Los dos tesauros empleados en las bases de datos de jurisprudencia carecen de presentación sistemática, lo que impide estudiar la estructura arborescente completa de los mismos para escrutar su profundidad. Sin embargo, a partir de la presentación alfabética se aprecia un mayor número de niveles jerárquicos en Aranzadi, lo que, sin embargo, no viene a implicar equilibrio o consistencia de los mismos.

3.6. Notas de aplicación

La propia concepción de las listas de descriptores preferentes empleadas por los sistemas legislativos parece alejarse de la naturaleza que explica el uso de las notas de aplicación. Ninguno de los dos tesauros cuenta con notas de alcance en el sentido tradicional; aunque existen numerosas especificaciones, principalmente temporales, que permiten aclarar un concepto especificando su tratamiento bajo una determinada norma o la legislación bajo la que se instruyó la causa.

4. Conclusiones

La búsqueda en lenguaje natural en sistemas de recuperación de información tradicional, como los aquí analizados, se basa, fundamentalmente, en el patrón del índice simple que proporciona "sólo palabras" sin tener en cuenta los conceptos o las relaciones entre los mismos, obligando al usuario a optar por introducir un mayor número de términos de búsqueda con el riesgo consiguiente del silencio en la respuesta, o bien, al esfuerzo de una tediosa selección. Para lograr una mayor eficacia en la recuperación temática, los sistemas estudiados en este trabajo incorporan en sus interfaces distintos modelos de lenguaje controlado. Aun reconociendo que constituyen instrumentos potentes para la recuperación de información, hemos de llamar la atención sobre la escasa normalización de la terminología jurídica en dichos entornos automatizados así como la notoria priorización de criterios apegados a la práctica jurídica frente al empleo de herramientas científicas rigurosas.

Se observan diferencias sustanciales en las herramientas empleadas en las bases de datos legislativas y en las jurisprudenciales. El examen de las listas de descriptores preferentes recogidas en los sistemas normativos evidencia políticas de indización diversas. El caso más notorio lo constituye *Legislación Aranzadi,* en la que se hallan tanto descriptores muy poco específicos —susceptibles, por tanto, de producir mucho ruido en la recuperación— y descriptores muy concretos, síntoma de cierto desequilibrio. Resulta palmario reconocer la enorme complejidad en el proceso intelectual requerido para tratar la documentación legislativa, tanto en el análisis jurídico riguroso como en el de contenido, dado que la documentación legislativa vertebra la vida cotidiana en su conjunto y no es posible restringir

los conceptos aparecidos en un único sublenguaje. En el caso de las bases de datos jurisprudenciales, por la propia naturaleza de los documentos tratados, los sistemas se decantan por el uso de tesauros asistemáticos, que tratan de reflejar el modelo mental de los usuarios a los que va dirigido el sistema. En estos tesauros, la terminología es mayoritariamente jurídica y las relaciones asociativas aparecen insuficientemente explotadas. Abogamos, en cualquier caso, por la compatibilidad de herramientas terminológicas estandarizadas sin deteriorar su adecuada adaptación al usuario. Podemos hablar de diferentes grados de homogeneidad en el vocabulario que aparece en las resoluciones judiciales respecto al que pueden presentar las diversas disposiciones normativas, diferencias que condicionan el tratamiento del lenguaje controlado. En las bases de datos legislativas quizá resultase de ayuda la posibilidad de discriminar por tipo de descriptor: temático, geográfico, onomástico e incluso primario o secundario.

Por último, consideramos que los instrumentos terminológicos que integran los sistemas estudiados muestran carencias que tal vez no puedan ser resueltas exclusivamente de la mano de la indización intelectual. Parece adecuado pensar que es preciso introducir y combinar nuevos modelos. Los desarrollos de software en el terreno de la recuperación semántica —que suponen un acercamiento entre el lenguaje natural y el controlado y que permiten la incorporación de una red semántica inteligente en la que los términos mantienen relaciones asociativas, jerárquicas y de sinonimia— o el empleo de la lógica difusa —utilizando algoritmos que permiten la ponderación de términos— podrían resolver aspectos hoy claramente deficientes.

5. Bibliografía

Alvite Díez, M. L. (2003). Tendencias en la investigación sobre recuperación de información jurídica. // Revista Española de Documentación Científica. 26:2 (2003) 191-212.

Espelt, C. (1998). Improving subject retrieval: user-friendly interfaces and effectiveness. [En línea]. // Bid. 1 (1998). <http://www.ub.es/biblio/bid/01espel1.htm>. Consultado: 21/09/2003.

Fugmann, R. (1993). Subject analysis and indexing: theoretical foundation and practical advice. Frankfurt: Indeks Verlag, 1993.

Gil Leiva, I.; Rodríguez Muñoz, J. V. (1997). Análisis de los descriptores de diferentes áreas del conocimiento en bases de datos del CSIC. Aplicación a la indización automática. // Revista Española de Documentación científica. 20:2 (1997) 150-160.

Gil Urdiciain, B. (1996). Manual de lenguajes documentales. Madrid: Noesis, 1996.

Gil Urdiciain, B. (1998a). Evaluación del rendimiento de tesauros españoles en sistemas de recuperación de información. // Revista Española de Documentación Científica. 21:3 (1998) 286-302.

Gil Urdiciain, B. (1998b). Evaluación semántica y estructural de tesauros. // Revista General de Información y Documentación. 8:2 (1998) 193-199.

Hearst, M. A. (1999). User interfaces and visualization. // Baeza-Yates, R. y Ribeiro-Neto, B. Modern information retrieval. New York: ACM Press, 1999. 257-323.

HjØrland, B. y Nielsen, L. K. (2001). Subject access points in electronic retrieval. // Williams, M. E. (ed.). Annual Review of Information Science and Technology (ARIST). Medford, NJ: Information Today. 35 (2001) 249-298.

Kuhlthau, C. C. y Tama, S. L. (2001). Information search process of lawyers: a call for "just for me" information services. Journal of Documentation. 57:1 (19) 25-43.

Lancaster, F. W. (1995). El control del vocabulario en la recuperación de información. Valencia: Universitat de València, 1995.

Lancaster, F. W. (2003). Do indexing and abstracting have a future?. Anales de Documentación. 6 (2003) 137-144.

Library of Congress. Global Legal Information Network. [En línea]. <http://www.loc.gov/law/glin/>. Consultado: 25/08/2003.

Lopes, I. L. (2002). Uso das linguagens controlada e natural em bases de dados: revisão da literatura. // Ciência da Informação. 31:1 (2002) 41-52.

López Alonso, M. A. (1998). Un tesauro conceptual para la recuperación de la información jurídica comercial. // Revista Española de Documentación Científica. 21:2 (1998) 164-173.

López Alonso, M. A. (2001). Integración de teorías para la representación y recuperación del conocimiento. [CD-ROM]. // V Congreso ISKO-España. Alcalá de Henares: Universidad de Alcalá de Henares, 2001.

López-Muñiz Goñi, M. (1984). Informática jurídica documental. Madrid: Díaz de Santos, 1984.

Maciá, M. (Ed.). (1998). Manual de documentación jurídica. Madrid: Síntesis, 1998.

Matthijssen, L. (1999). Interfacing between lawyers and computers: an architecture for knowledge-based interfaces to legal databases. The Hage: Kluwer Law International, 1999.

Osés, M. A. (2000). Tesauros relacionales y acceso a la información especializada: metodología para el desarrollo de un tesauro de terminología jurídica. Buenos Aires: Dunken, 2000.

Páez Mañá, J. (1994). Bases de datos jurídicos: características, contenido, desarrollo y marco legal. Madrid: CINDOC, 1994.

Páez Mañá, J. (1995). Comentarios sobre algunas particularidades de las bases de datos jurídicas. // Actualidad informática Aranzadi. 16 (1995) 4-10.

Páez Mañá, J.; et al. (Coord.). (1997). Tesauro de Derecho [En línea]. Madrid: CINDOC. <http://pci204.cindoc.csic.es/tesauros/Derecho/DerTes.htm>. Consultado: 06/10/2003.

Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. // Journal of Information Science. 20:2 (1994) 108-119.

Salton, G. (1987). Historical note: the past thirty years in information retrieval. // Journal of the American Society for Information Science. 38:5, (1987) 375-380.

Tenopir, C.; Ro, J .S. (1990). Full text databases. New York: Greenwood Press, 1990.

UNE 50-106-90. (1990). Directrices para el establecimiento y desarrollo de tesauros monolingües. Madrid: AENOR, 1990.