LA RECUPERACIÓN DE LA INFORMACIÓN EN BASES DE DATOS JURÍDICAS: EVALUACIÓN DE ARANZADI Y LA LEY

Autores: Mª Luisa Alvite Díez
Universidad de León. Área de Biblioteconomía y Documentación
dphlad@unileon.es

Resumen: Se pretende realizar una evaluación cualitativa de cuatro bases de datos jurídicas españolas implantadas en numerosas unidades de información y producidas por dos importantes empresas editoriales de reconocido prestigio. Analizaremos la recuperación de la información en cada una de ellas evaluando la calidad de la indización y las características del correspondiente software documental.

Palabras clave: Bases de datos jurídicas ; Evaluación de bases de datos ; Calidad de bases de datos; Recuperación de la información ; Indización

Abstract: We have the intention to make a qualitative evaluation of four Spanish legal databases incorporated in numerours information centres and produced by two important and prestigious publishing enterprises. We will analyze the information retrieval in each of them evaluating the indexation quality and the characteristics of the software.

Keywords: Legal databases ; Evaluation of databases ; Quality of databases ; Information retrieval ; Indexation

Introducción

Recopilar, clasificar y analizar documentos son actividades vinculadas estrechamente al mundo jurídico.

A partir de la Baja Edad Media comienza un proceso de unificación normativa con la consiguiente aparición de recopilaciones de cuerpos legales. La gran época de la codificación coincidirá con el paso del Antiguo Régimen al Estado moderno: recopilaciones estructuradas cuyos textos aparecen analizados y referenciados a través de índices. Por último, la entrada en vigor de la Constitución de 1978, la división territorial del Estado, la configuración y desarrollo de nuevos tribunales y la pertenencia a la Unión europea han determinado la explosión actual de la documentación jurídica en nuestro país.

El requisito de publicidad de la norma jurídica se ha llevado a cabo de diversas formas a lo largo del tiempo, dependiendo del desarrollo y evolución de los medios de comunicación social. Si la aparición de la imprenta supuso un punto de inflexión, la incorporación de las tecnologías de la información al mundo del Derecho ha supuesto cambios, tanto en el soporte como en la forma de acceso: bases de datos en línea, en CD-ROM, DVD, vía web, etc. Quizá la verdadera revolución se ha producido a partir de la implementación de herramientas de indexación y búsqueda adaptadas al CD-ROM.

Las bases de datos parten de un concepto muy semejante al del repertorio en papel, esto es, documentos ordenados cronológicamente a los que se accede a través de índices. Las posibilidades de los nuevos soportes y medios vienen a solucionar cuatro problemas tradicionales de la documentación jurídica: difusión del ordenamiento jurídico, acceso fácil a las fuentes del derecho, obtención de forma exhaustiva de cualquier texto legislativo o jurisprudencial y aumento de la seguridad jurídica por medio de interrelaciones legislativas, jurisprudenciales y doctrinales.

Atendiendo a la definición establecida por López-Muñiz Goñi (1984) se consideran bases de datos jurídicas a aquellos conjuntos de documentos jurídicos básicos (legislación, jurisprudencia, interpelaciones parlamentarias y doctrina), almacenados en soportes magnéticos o de cualquier otro material y susceptibles de ser tratados, recuperados y transmitidos de forma total o parcial mediante procedimientos y medios informáticos, que con la aplicación de técnicas derivadas de la utilización de la informática jurídica pretenden ser utilizados con una finalidad divulgadora pública y generalizada de su contenido.

Estas bases de datos presentan, como ha indicado Páez Mañá (1995), ciertas particularidades respecto al resto de bases de datos: son bases de datos a texto completo, se emplean distintas unidades documentales, contienen un enorme volumen de documentación almacenada con diferente grado de vigencia, precisan una permanente actualización, su carácter exhaustivo es el garante de la seguridad jurídica, su aplicabilidad está delimitada a un ámbito jurisdiccional concreto, existe interconexión de la documentación jurídica, las fuentes documentales están claramente delimitadas, tienen un coste elevado y cuentan con sistemas de tratamiento de la información adaptados a las necesidades de los profesionales del mundo jurídico.

Consideramos que la evaluación de estos recursos informativos constituye una clara necesidad para los profesionales de la información y para los operadores jurídicos. Ha de contribuir a valorar con rigor la eficacia y eficiencia de estos productos, a determinar las diferencias entre las diversas bases de datos existentes, numerosas en el sector jurídico, que, a priori, cubren los mismos contenidos y, por último, ha de ayudar a justificar el coste de productos y la preferencia por modalidades concretas de suscripción.

La evaluación de las bases de datos

La selección de las bases de datos objeto de la evaluación se realizó partiendo de una doble premisa, analizar bases de datos jurídicas a texto completo - la mayoría de los estudios evaluativos se asientan en el análisis de bases de datos bibliográficas - y, por otro lado, permitir interrelacionar la legislación y la jurisprudencia - característica esencial de la documentación jurídica - Se excluyeron los productos sectoriales, dedicados a ámbitos jurídicos especializados.

Tras examinar las suscripciones mantenidas por la Universidad de León, las bases de datos que se ajustaban a estos criterios fueron: *Base de datos Aranzadi Legislación*, *Base de datos Aranzadi Jurisprudencia*, *Base de datos Repertorio de Legislación La Ley* y *Base de datos Jurisprudencia La Ley*. Todas ellas cumplen las condiciones establecidas, incluir la normativa, la documentación jurisprudencial y las referencias cruzadas entre los diferentes documentos jurídicos sustentados en un mismo precepto legal o jurisprudencial.

La evaluación se realiza, por tanto, empleando la versión en red de estos productos instalada en la intranet del campus, en estos momentos en soporte CD-ROM.

Atendiendo a Palma Villalón (1995), consideramos que la indización y la recuperación de la información son procesos que han de estudiarse conjuntamente por lo que hemos decidido como primera aproximación a la evaluación de las bases de datos señaladas estudiar la indización y el correspondiente software gestor.

Estamos de acuerdo, sin embargo, con Purificación Moscoso (1997) al considerar que para la evaluación y el control de calidad de las bases de datos se han de tener en cuenta tres categorías de análisis: evaluación descriptiva (contenido de la base de datos, contenido de los registros e indización de los documentos), evaluación del software de recuperación y evaluación de la interfaz de usuario. En esta misma línea, para Olvera Lobo (1999) cualquier estudio evaluativo ha de combinar el sistema, el usuario y la información.

Emplearemos criterios de evaluación y control de calidad sustentados en estudios sobre evaluación de bases de datos y, en concreto, sobre calidad de la indización y análisis del gestor documental.

Atendiendo, entre otros, a Sievert y Andrews (1991),Salvador Oliván, Angós Ullate y Fernández Ruiz (1999), Extremeño (1999), etc. podemos concluir que la evaluación de la calidad de la indización se asienta en tres pilares: consistencia, relevancia y exhaustividad. Para determinar su calidad se han elegido muestras representativas de documentos.

Por último, el análisis se realiza sobre las bases de datos legislativas y sobre las bases de datos de jurisprudencia de cada uno de los productores, con el fin de evaluar los aspectos coincidentes y diferenciadores, así como las interrelaciones entre ambas, considerando que son productos diseñados para utilizarse, normalmente, de forma complementaria y cuya valoración más objetiva requiere el análisis conjunto.

Evaluación de la indización

Análisis de la consistencia

El análisis de la consistencia, se realiza mediante la selección de 8 racimos temáticos, recurriendo al campo de texto libre. En las bases de datos legislativas la mitad de racimos abarcan temas generales (*Urbanismo; Medio*

ambiente; Administración local ; Abastecimiento de aguas) y la otra mitad temas específicos (*Propiedad horizontal; Aditivos alimentarios; Convenios colectivos de artes gráficas; Productos tóxicos)*. En el análisis de las bases de datos de jurisprudencia además de respetar este principio, se han establecido *clusters* que representen las cuatro jurisdicciones fundamentales (Civil: *Compraventa mercantil; Derecho de autor.* Penal: *Falsedad documental; Tráfico de influencias.* Social: *Despido; Prestaciones por maternidad.* Administrativo: *Responsabilidad de la Administración ; Impuesto sobre construcciones). S*e han anotado los descriptores utilizados en 6 documentos de cada grupo. Se atienen al principio de consistencia aquellos descriptores repetidos en tres o más de los registros.

El nivel de consistencia de *Aranzadi Legislación* alcanza un porcentaje del 75%, sin haber incluido en este análisis los descriptores secundarios que presentan muchos de los documentos de esta base de datos. La consistencia del *Repertorio de legislación La Ley* no supera el 37,5%.

El nivel de consistencia alcanzado en las dos bases de datos jurisprudenciales es el mismo, situándose en el 62,5%.

Se confirma un elevado control terminológico, - no es baladí el hecho de que los productos analizados partan de publicaciones impresas con una gran experiencia en la terminología jurídica y en la indización -. La mayor coherencia en la asignación de descriptores se aprecia, particularmente, en los racimos temáticos más específicos. Resulta igualmente reseñable el hecho de que muchos de los descriptores empleados que no alcanzan el nivel de consistencia mantienen una clara afinidad con la temática del cluster.

En las bases de datos estudiadas los documentos se analizan con el propósito de servir al profesional jurídico, es una indización orientada a las necesidades de un usuario concreto, que precisa conocer no sólo los conceptos que recogen la semántica del documento sino también aspectos temporales, procedimentales, etc.

Análisis de la relevancia

Se ha establecido partiendo de los valores de discriminación establecidos por Chu y Ajiferuque (1989) dividiendo el número de registros asociados a un descriptor por el número total de registros de la base de datos. Hemos elegido 20 descriptores al azar. En la evaluación de las bases de datos de jurisprudencia los descriptores se han distribuido por jurisdicciones, correspondiendo cinco a cada una de ellas.

El grado de relevancia en *Aranzadi Legislación* alcanza un promedio de 0,011, lo que supone un porcentaje de recuperación del 1%. La relevancia de la indización de *La Ley Legislación* se aleja al 0,006. El porcentaje medio de relevancia en *Aranzadi Jurisprudencia* es del 0,003 y en *Jurisprudencia La Ley* del 0,002. Ninguna de los dos bases de datos se acerca a ese porcentaje óptimo que se ha establecido entre el 0,02 y el 0,05, para lograr recuperar entre el 2% y el 5% de los documentos.

Las bases de datos cuentan con descriptores muy específicos, temáticos, onomásticos y geográficos y, sin embargo, para que la recuperación alcance el nivel de precisión deseado, será necesario emplear, en muchos casos, búsquedas mixtas, en las que se combine el descriptor con términos del texto libre, resumen etc. Consideramos que el indizador de documentación jurídica trabaja con unidades documentales de una gran diversidad y extensión y sólo un aspecto muy concreto – un artículo, uno de los fundamentos de Derecho, etc. – serán relevantes en un momento dado para un usuario.

Análisis de la exhaustividad

La evaluación de la exhaustividad está ligada al número de términos que describen los diferentes conceptos del documento. Hemos analizado 50 documentos de cada base de datos y se ha calculado la media dividiendo el total de descriptores de la muestra entre el número de registros de ésta.

La exhaustividad media en *Aranzadi Legislación* es de 8,64 lo que implica una exhaustividad alta, dado que la recomendación más extendida es la de utilizar un número de descriptores entre 8 y 12 por documento. Sin embargo, llama la atención la gran heterogeneidad que se aprecia en el número de descriptores establecidos de forma individual. La media en el *Repertorio de legislación La Ley* es de 2,98, lo que supone una exhaustividad no muy alta. Como aspecto positivo podemos destacar el grado de uniformidad en el número de descriptores asignados por documento.

La media de descriptores empleados por documento en *Aranzadi Jurisprudencia* es de 5 y *Jurisprudencia La Ley* emplea una media de 2 descriptores por registro. La baja exhaustividad en las bases de datos jurisprudenciales podría justificarse atendiendo a lo establecido en la norma ISO 50-121-91: "Si un lenguaje de indización incorpora un tesauro, el número de términos asignados al documento puede reducirse sin pérdidas, ya que los términos generales y otras relaciones pueden establecerse en el tesauro".

Evaluación del software de recuperación

Los softwares objeto de evaluación son gestores de bases de datos documentales, basados en el modelo booleano.

Software de recuperación en las bases de datos legislativas

	Aranzadi	**La Ley**
Software	Knosys Windows	Cicerón
Versión	Entrega oct. 2000	2.01 (entrega oct. 2000)
Arquitectura de acceso	Monopuesto Red	Monopuesto Red
Operadores booleanos	Y (Intersección) O (Unión) NO (Exclusión)	Y (Intersección), (+) O (Unión), (,) SIN (Exclusión)

Operadores de adyacencia y proximidad	P (Párrafo)	CERCAn JUNTOn
Truncamiento	* (Abierto) ? (Cerrado)	* (Abierto) ? (Cerrado)
Operadores relacionales	No	>(Mayor) < (Menor)
Navegación por índices	Disposición Órgano emisor Publicaciones Voces	Voz Título Texto Ámbito Rango Organismo
Uso de hipertexto	Sí	Sí
Búsqueda por conceptos	Diccionario de términos	Indice
Búsqueda dentro del texto	Sí	Sí
Almacenamiento de estrategias de búsqueda	Sí	Sí
Visualización de resultados	Línea Completo Línea/Completo	Sólo Título Sólo documentos Títulos/Documentos Presentación horizontal Presentación vertical
Formatos de resultados	Selección de texto Selección de documentos Impresión Exportar (ASCII, RTF)	Selección de texto Selección de documentos Impresión Exportar (ASCII)

Ambos softwares cuentan con aspectos coincidentes, algunas de las prestaciones son superiores en Knosys (Micronet) y otras lo son en el sistema Cicerón-La Ley (tecnología Dataware). Valoramos especialmente la rapidez de respuesta, los operadores de proximidad y la posibilidad de consultar los ficheros inversos de *La Ley.* Destacamos del software de *Aranzadi* el uso de paréntesis para efectuar búsquedas sofisticadas, la mayor profusión en el uso de elementos hipertextuales y el enorme grado de profundidad de su tesauro.

Los operadores de adyacencia de *La Ley Legislación* son completos, permiten pedir el número de caracteres de proximidad entre los términos y el orden de los mismos (curiosamente no están activos en la base de datos de jurisprudencia). El operador de párrafo de *Aranzadi* es, sin embargo, muy limitado, no es posible controlar la proximidad y produce demasiado *ruido* en la recuperación.

En las bases de datos legislativas, las búsquedas por conceptos jurídicos se realizan en el campo "Voz" o "Voces" (descriptor). En las dos bases

de datos legislativas se nos presenta un listado normalizado de descriptores sin mostrarnos sus relaciones jerárquicas, semánticas y asociativas. *Aranzadi* señala si el término es voz principal o no, lo que ayuda al usuario a realizar una consulta más específica o más genérica. Asociado al índice de "Voces", *La Ley* presenta el número de documentos de la base de datos que contienen ese descriptor, sin embargo, *Aranzadi*, obliga al usuario a trasladar el descriptor a la pantalla de consulta para saber el número de documentos indizados con el mismo.

En las búsquedas por texto libre el motor de búsqueda de *Aranzadi* discrimina las palabras vacías, no así el de *La Ley.* Los dos son insensibles al uso de mayúsculas, minúsculas o acentos.

El software de Aranzadi nos permite elaborar formatos de salida propios y realizar una ordenación personalizada.

La gran diferencia entre las bases de datos de legislación y las de jurisprudencia es que en estas últimas, para garantizar la recuperación de la información y evitar el ruido derivado del excesivo volumen de documentos, el software introduce la posibilidad de localizar la información empleando un tesauro. Las dos bases de datos evaluadas incluyen un tesauro compuesto por descriptores simples y sintagmáticos que constituyen una auténtica red semántica en la que se incluyen multitud de aspectos procedimentales, doctrinales, aspectos subjetivos, resolutivos o referentes al fallo de la sentencia, , etc.

Software de recuperación en las bases de datos de jurisprudencia (aspectos no coincidentes con las bases de datos legislativas)

	Aranzadi	**La Ley**
Operadores de proximidad y adyacencia	P (Párrafo)	No
Navegación por índices	Disposiciones estudiadas Ponentes Voces	Tesauro Voces secundarias Disposiciones legales Texto sumario Fecha Tribunal Sala Sección Ponente
Búsqueda por conceptos	Tesauro	Tesauro Palabras clave Voces secundarias

A pesar de que el término empleado por las bases de datos es Tesauro, no podemos entender éste en el sentido estrictamente documental del término. Sí es cierto que desde el punto de vista de su estructura los tesauros de estas dos bases de datos son vocabularios controlados y dinámicos de términos jurídicos que mantienen entre ellos relaciones jerárquicas, semánticas y asociativas, sin embargo, el cuerpo léxico de estos tesauros además de incluir descriptores o términos preferentes y no descriptores, basan su riqueza y utilidad en la inclusión de una serie de términos compuestos que definen los conceptos o entidades jurídicas. Se trata de una terminología muy rica y especializada en la que se utiliza una sintaxis propia, tomada de la práctica jurídica habitual, con una carga informativa enorme.

Aranzadi organiza el tesauro en cuatro campos semánticos que coinciden con las cuatro jurisdicciones fundamentales (civil, penal, laboral y administrativa), el paso inicial es 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. Al lado de cada carpeta, el sistema indica el número de documentos indizados con el descriptor correspondiente.

La Ley no organiza su tesauro en campos semánticos y se ve en la obligación de enviar 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, con un simple clic de ratón direcciona la búsqueda al tesauro. La visualización se basa en carpetas y subcarpetas.

Ninguno de los dos tesauros cuenta con notas de alcance en el sentido tradicional, si bien existen multitud de especificaciones temporales que sirven para aclarar un concepto especificando la normativa bajo la que se instruyó la causa. Las relaciones asociativas son escasas en los dos.

El tesauro de la Ley adolece de cierta rigidez, no permite seleccionar varios descriptores a la vez y combinarlos con el operador booleano deseado. Permite navegar en la estructura, pero, sin embargo, para iniciar una nueva consulta obliga al usuario a abrir el menú desplegable de navegación y seleccionar la opción "colapsar todo".

Conclusiones

- Consideramos que la consistencia de la indización tiene una mayor incidencia en la recuperación de información normativa y un valor más relativo en la recuperación de jurisprudencia, donde los productores de base de datos jurídicas emplean además de un vocabulario controlado herramientas más eficaces como el tesauro.

- Por lo que se refiere a la profundidad de la indización, en las bases de datos jurisprudenciales el cómputo del número de descriptores puede ocultar la

profundidad real de la indización asentada en el uso del tesauro, en el cual la carga informativa va más allá de los descriptores.

- La evaluación nos permite constatar la relación directa entre lenguajes controlados de indización y calidad del sistema de información. Las bases de datos a texto completo no garantizan una recuperación eficaz dado que difícilmente pueden analizar los conceptos en función de su contexto. Los lenguajes de indización permiten búsquedas más precisas, y su utilización y calidad no debe relegarse únicamente a las bases de datos bibliográficas.

- El gran problema de las bases de datos a texto completo continúa siendo el ruido. Para lograr la precisión exigida por el usuario a la calidad de la indización debe conjugarse la búsqueda por otros campos, la navegación por las relaciones hipertextuales, el empleo de herramientas sofisticadas de control terminológico, etc.

- El uso de técnicas hipertextuales garantiza no sólo la combinación de textos normativos, jurisprudenciales y doctrinales sino también la seguridad jurídica.

- La capacidad heurística del tesauro hace de éste la herramienta más potente para la recuperación de información relevante es el tesauro. Se constata, curiosamente, la inexistencia de tesauros conceptuales en bases de datos legislativas españolas lo que quizá pueda explicarse atendiendo a las particularidades de la documentación normativa, en cuya recuperación prima la búsqueda por el área de título, rango, número oficial y fecha de aprobación y, obviamente, por la enorme dificultad que conlleva la indización de documentos legislativos (piénsese en una Ley presupuestaria, en un código legislativo, etc.).

- Los softwares analizados pertenecen a los denominados sistemas tradicionales de recuperación de la información, los más extendidos, a pesar de sus limitaciones. Parece adecuado pensar que es preciso introducir y combinar nuevos modelos, cuyas prestaciones ya están suficientemente constatadas y en funcionamiento en algunos de los motores de recuperación de la web y en otros softwares documentales: aplicación de técnicas de clustering que permiten la utilización de criterios de relevancia, stemming, mapping, ponderación de conceptos, relaciones de similaridad, reconocimiento de sinónimos y cuasisinónimos, etc.

Como corolario, hacemos hincapié en la necesidad de analizar el contenido y la interfaz de las bases de datos para realizar una evaluación global de la calidad de los productos.

Por último, queremos señalar, como claro indicador de la importancia económica y estratégica de estas bases de datos en nuestro país, el hecho de que los dos productores analizados en esta comunicación han pasado en los últimos años a manos de dos multinacionales de la información: Thomson Corporation (propietaria de Dialog, Derwent, ISI, Gale Group, etc.) en el caso

de Aranzadi y Wolters Kluwer (el mayor propietario de bases de datos jurídicas en España) en el caso de La Ley.

Bibliografía

CHU, C. M. ; AJIFERUQUE, I. “Quality of indexing library and information science databases”.*Online Review,* 1989, vol. 13, nº 1., p. 11-35.

EXTREMEÑO PLACER, A. “La calidad en la indización de documentos: elemento indispensable para optimizar la recuperación de información”. En *Congreso ISKO- España EOCONSID 99 (4. 1999. Granada).* Granada, 1999, p. 313-320.

LÓPEZ-MUÑIZ GOÑI, M. *Informática jurídica documental.* Madrid : Díaz de Santos, 1984.

MOSCOSO, P. “Pautas para evaluar bases de datos en CD-ROM”. *Revista General de Información y Documentación,* 1997, vol. 7, nº 1, p. 187-204.

OLVERA LOBO, M.D. “Evaluación de sistemas de recuperación de la información: aproximaciones y nuevas tendencias”. *El profesional de la información,* 1999, vol. 8, nº 11, p. 4-14.

PÁEZ MAÑÁ, J. “Comentarios sobre algunas particularidades de las bases de datos jurídicas”. *Actualidad informática Aranzadi,* 1995, nº 16, p. 4-10.

PALMA VILLALÓN, M.V. “Técnicas y métodos par mejorar la calidad de la indización y su recuperación en bases de datos documentales de ciencias sociales y humanidades”. En *Jornades catalanes de Documentació (5.1995.Barcelona)*. Barcelona: COBDC ; SOCADI, 1995, p. 223-239

SALVADOR OLIVÁN, J.A. ; ANGÓS ULLATE, J.A. ; FERNÁNDEZ RUIZ, M.J. "Comparación y evaluación de las bases de datos ERIC, LISA e ISA sobre el tema "Recuperación de la información". *Revista Española de Documentación Científica,* 1999, vol. 22, nº 1, p. 51-63.

SIEVERT, M.C. ; ANDREWS, M.J. "Indexing consistency in Information Science Abstracts". *Journal of the American Society of Information Science,* 1991, vol 42, p.1-6