Una metodología inductiva para la adecuación terminográfica de glosarios explicativos en internet[1]

Ana Fernández Pampillón Cesteros (apampi@filol.ucm.es)
María Matesanz del Barrio (mmatesanz@filol.ucm.es)
Olimpia Pérez Broncano (olimpia-perez@hotmail.com)
Universidad Complutense de Madrid

1.- INTRODUCCIÓN

La creciente disponibilidad de herramientas de construcción y difusión de materiales en internet está propiciando la aparición de un tipo nuevo de obras alfabetizadas, de carácter enciclopédico o lexicográfico, creadas de forma abierta -libre y colaborativa- para la organización y descripción del conocimiento, como es, por ejemplo, Wikipedia. En el ámbito académico, este fenómeno se ve favorecido por la reciente incorporación de las plataformas *e-learning*, que integran, en un mismo entorno informático, las herramientas necesarias para la creación, uso y difusión de materiales educativos.

Los glosarios académicos son un ejemplo de este tipo de obras creadas por los profesores, investigadores o estudiantes en los entornos *e-learning*, con el propósito específico de facilitar el acceso y la comprensión de los contenidos de aprendizaje de las materias académicas, así como el aprendizaje de la terminología de las disciplinas de estudio (Soergel, 2002). De los glosarios que se explotan en entornos virtuales hemos centrado esta investigación en lo que hemos denominado *glosarios explicativos*, glosarios académicos construidos para registrar los términos, conceptos e información de dominios específicos, en algunos casos nuevos, poco establecidos y, normalmente, interdisciplinares. Constituyen una nueva forma de vocabulario que amplía el modelo de contenido tradicional de los glosarios. Formalmente, definimos el glosario explicativo como un *conjunto de términos que recogen y explican el conocimiento sobre una o varias disciplinas con un modelo de contenido multidimensional y en niveles*.

Estos productos académicos, claramente empíricos y pensados para una explotación prioritariamente docente, presentan unas características comunes fácilmente reconocibles: (i) están creados en los entornos virtuales académicos; (ii) se desarrollan con pocos recursos, (básicamente las herramientas que incluyen las plataformas *e-learning*: editores HTML, wikis, bases de datos, almacenamiento y publicación de archivos, etc.); (iii) tienen como usuarios finales a los autores –profesores, investigadores y estudiantes-; (iv) son abiertos –de creación libre y colaborativa-; (v) son dinámicos; y (vi) están orientados al uso académico. De las ventajas más importantes que presentan, frente a los glosarios tradicionales de especialidad,

[1] Este trabajo se ha desarrollado en el marco del Proyecto de Innovación y Mejora de la Calidad de la Docencia (PIMCD-66/2007-2008) financiado por la Universidad Complutense de Madrid.

destacamos las siguientes: (i) requieren, relativamente, poco esfuerzo de construcción; (ii) son precisos y contienen un amplio conocimiento del ámbito concreto del vocabulario; (iii) están adaptados a las necesidades académicas; y (iv) son fácilmente comprensibles, porque están redactados con el lenguaje propio del dominio de conocimiento y de la comunidad científico-académica que lo utiliza.

Sin embargo, presentan también una serie de inconvenientes que dificultan seriamente su efectividad y que derivan de la falta de sistematización del proceso de construcción. En estos procesos de creación, en general, no se aplican criterios lexicográficos que garanticen su uniformidad y coherencia porque quienes construyen estos glosarios no tienen conocimientos lingüísticos. De los principales problemas que han detectado los usuarios finales (fundamentalmente, profesores), señalamos los siguientes (Sierra, et. al, 2009): (i) limitación de las posibilidades de explotación; (ii) acortamiento del ciclo de vida; (iii) dificultad de reutilización de su contenido y de integración en otros sistemas lingüísticos o informáticos; y, finalmente, (iv) mantenimiento costoso, por lo que se corre el riesgo de que pierdan validez.

Dos son las posibles soluciones para resolver la falta de sistematicidad: (i) rehacer estas obras aplicando técnicas lexicográficas, o (ii) realizar sobre ellas lo que hemos dado en llamar una *adecuación terminográfica* (AT). En el primer caso, el coste es excesivo y no es, en absoluto, rentable. En último término, esto significaría en la práctica crear una obra nueva, y no hay que olvidar que se trata de glosarios que ya están siendo utilizados y, por tanto, aceptados como tales. Es más, en muchos casos estos glosarios cuentan con una versión previa publicada en formato papel. En el segundo caso, la adecuación terminográfica supondría recoger, en un modelo de contenido nuevo ajustado a criterios terminográficos, *la estructura implícita general más específica* de la obra ya creada, lo que permitiría unificar y homogeneizar todas las entradas. Para que sea realmente eficiente la adecuación terminográfica el proceso debe ser sistemático. Esta segunda vía, que respeta la concepción y contenido originales de la obra, es la que hemos desarrollado y presentamos en este artículo.

2.- METODOLOGÍA PROPUESTA PARA LA ADECUACIÓN TERMINOGRÁFICA

Entendemos por adecuación terminográfica el proceso de organización y descripción explícita del contenido de vocabularios de especialidad semiestructurados mediante la aplicación de criterios lexicográficos, respetando la estructura implícita y la riqueza de conocimiento originales. La AT permite definir una macroestructura, homogeneizar la microestructura, y explotar y mantener automáticamente los repertorios terminológicos, en particular los glosarios explicativos, creados con finalidad docente en los campus virtuales universitarios. No obstante, ese proceso solamente es rentable en la medida en que es sistemático y automatizable, lo cual, desde nuestro punto de vista, se puede conseguir

mediante la extracción inductiva de un esquema de la estructura común más específica subyacente a todos los artículos de una obra de estas características. La metodología inductiva propuesta consta de dos etapas: (i) búsqueda de la estructura global y (ii) normalización terminográfica. La primera etapa consiste en buscar esa estructura global más específica subyacente del glosario, la cual, en la segunda, se ajusta o normaliza manualmente utilizando criterios terminográficos (Adelstein y Cabré, 2003; Wrigth y Budin, 1997; Sager, 1990).

El resultado de la AT es un *modelo de contenido* formal basado en la estructura implícita de organización de información generada de forma libre y colaborativa por los autores del glosario, que permite, además, su almacenamiento y manipulación. Este modelo optimiza la gestión y explotación de dicha información, tanto si es automática como inteligente.

Una vez obtenido el modelo, se aplica al contenido original del glosario para obtener una versión mejorada o, lo que es lo mismo, adecuada terminográficamente. Esta aplicación la hemos llevado a cabo con el soporte de un sistema de gestión terminográfica (SGT) que, además, facilita el mantenimiento permanente del glosario. La estrategia seguida, inductiva e iterativa, se produce en seis fases consecutivas, como puede verse en el siguiente esquema (figura 1).

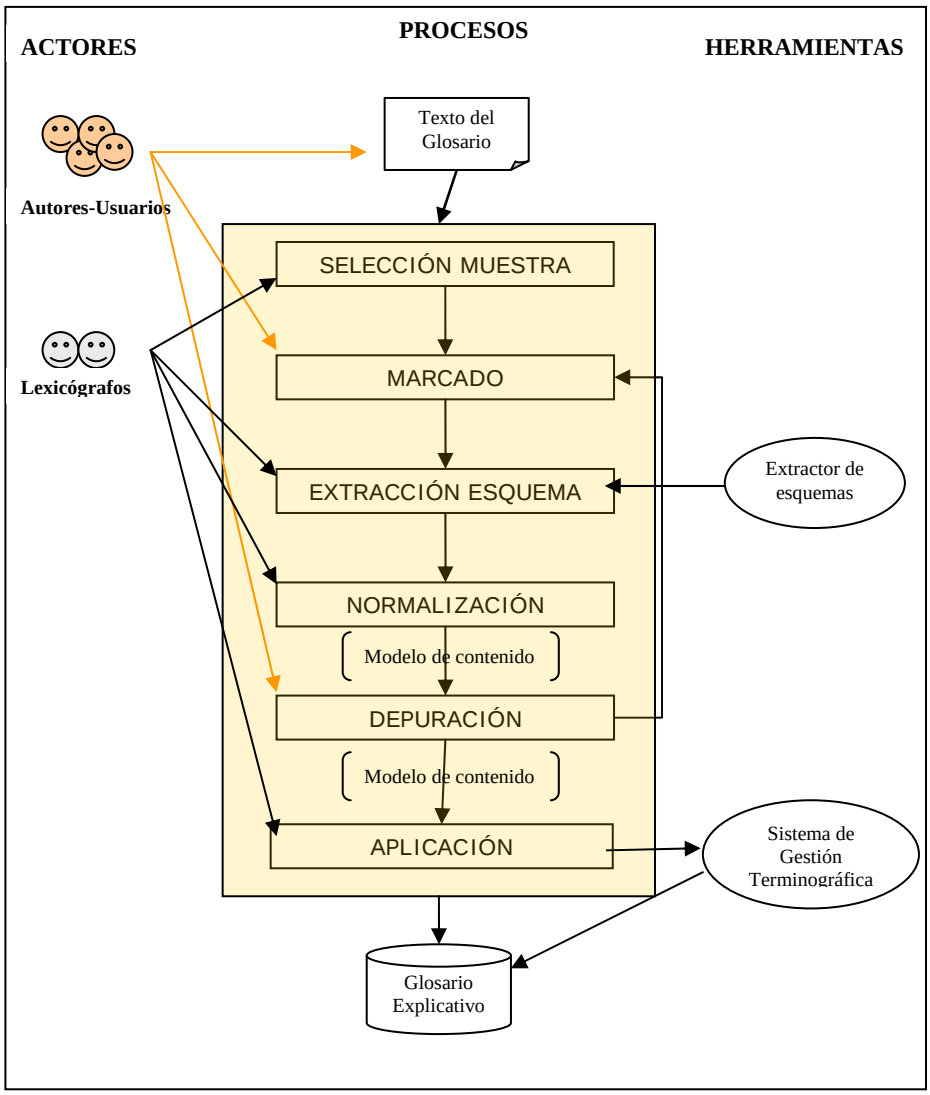

Figura 1. Método de Adecuación Terminográfica: esquema de actores, procesos y herramientas

El desarrollo de las fases se produce como se detalla a continuación.

Fase 1. ***Selección de la muestra.*** En esta fase el equipo de lexicógrafos selecciona una colección representativa de entradas del glosario que sirva de fuente del proceso de AT. La muestra se divide en varias partes: la muestra inicial y una o más muestras de depuración. La muestra inicial se utiliza en la primera ejecución de las fases 2, 3 y 4, mientras que las muestras de depuración se utilizan en las siguientes iteraciones de las fases 2, 3, 4 y 5 que van refinando el esquema obtenido. La muestra debe ser lo más representativa posible, puesto que servirá para extraer y depurar el modelo de contenido. El criterio de selección que hemos seguido es la elección de entradas de una misma letra, consiguiendo una muestra total de aproximadamente el 30%.

Fase 2. ***Marcado.*** Esta fase consiste en marcar en la muestra los elementos de contenido de las entradas. Las marcas las introducen los propios autores del glosario, con la ayuda de un lingüista, cuando no están familiarizados con el lenguaje de marcado, dado que

son los expertos en el dominio quienes reconocen sin problemas cada una de las partes o estructuras de las entradas si bien no hay un esquema declarado en la obra original. El marcado se realiza con el metalenguaje XML (Bray; Paoli; Sperberg-McQueen) porque presenta, frente a otros lenguajes de marcado, claras ventajas como, por ejemplo, (i) su fácil manejo, (ii) su adecuación para el marcado textual; (iii) su carácter estándar como código en la Web para el intercambio y manipulación automática de documentos estructurados; y (iv) el amplio abanico de herramientas software de las que se dispone para su procesamiento. Estas características facilitan, por un lado, el procesamiento automático de la muestra para la extracción del esquema y, por otro, la implementación y explotación del glosario explicativo final.

Fase 3. ***Extracción del esquema.*** En esta fase se extrae la estructura más específica común a todas las entradas marcadas de la muestra inicial. El objetivo es obtener los elementos estructurales y las relaciones que están presentes en todas las entradas. Esta fase se realiza con el apoyo de una herramienta de extracción automática de esquemas de desarrollo propio[2]. El esquema obtenido está formado por entidades, atributos y relaciones entre entidades, descritos en forma de reglas, las cuales se pueden representar gráficamente y, en consecuencia, ese esquema resulta fácil de interpretar.

Fase 4. ***Normalización terminográfica.*** La normalización terminográfica del esquema obtenido es el objetivo de esta fase. La finalidad es lograr la máxima coherencia y uniformidad en el glosario electrónico, sin perder la organización básica y el contenido del glosario original. Esta normalización consiste en: (i) regularizar los lemas de la nomenclatura; (ii) identificar y organizar los elementos de las entradas originales para ajustarlos al esquema terminográfico; (iii) explicitar las relaciones léxicas y de uso que existen entre las entradas; (iv) crear un tesauro a partir de las relaciones establecidas previamente. El resultado es un modelo de contenido que describe y organiza la información original del glosario y que permite su explotación y mantenimiento automático.

Fase 5. ***Depuración.*** Esta fase tiene por objeto probar el modelo de contenido obtenido en la fase anterior. Para ello se pide a los autores del glosario original que lo apliquen a la muestra de depuración. Si el modelo no responde adecuadamente al contenido de las nuevas entradas que se han marcado, los autores podrán utilizar etiquetas ya definidas (con nuevas dependencias estructurales) o bien crear otras nuevas que recojan los elementos identificados en esta fase. En este punto se repite el proceso desde la fase 3. La iteración que se lleva a

[2] Aunque existen otras propuestas de extracción automática de estructuras de datos XML (Garofalakis, et. al, 2000; Jung et al., 2002; Chen et al., 2003; Eki et al., 2008) nosotros hemos implementado un algoritmo propio basado en el modelo Entidad Relación de bases de datos. Esta solución proporciona esquemas sencillos de interpretar pero tiene el inconveniente de que éstos no son concisos, es decir, no se obtiene necesariamente el esquema más simple. Este inconveniente no es un problema en el método AT porque el esquema resultado se someterá a un proceso de normalización posterior.

cabo permite ir refinando el modelo de contenido hasta que se logre describir con precisión la estructura más específica común a las entradas de la muestra.

Fase 6. ***Aplicación.*** En esta última fase el modelo de contenido obtenido se aplica a todo el glosario, creando una versión electrónica, preferentemente, con ayuda de un SGT, aunque si no se dispone de un software específico es posible utilizar herramientas de edición de documentos XML. El empleo de un SGT facilita, además, la publicación y mantenimiento dinámico de la información. La versión electrónica resultante del glosario es independiente de la plataforma y, por ello, puede ser exportada y reutilizada en otros entornos. Esta versión incrementa las posibilidades de acceso y consulta automática de las entradas a partir de cada uno de los elementos definidos en el modelo de contenido. El mantenimiento, como hemos dicho, se simplifica enormemente puesto que se puede ampliar o modificar el contenido del glosario de forma coherente conforme al modelo.

3. APLICACIÓN DEL MÉTODO PARA LA ADECUACIÓN TERMINOGRÁFICA DEL *Glosario explicativo e-Derecho*

La viabilidad del método de AT fue probada en el marco de un Proyecto de Innovación y Mejora de la Calidad de la Docencia de la Universidad Complutense de Madrid -PIMCD-66/2007-2008, con resultados satisfactorios. Con este proyecto se buscaba la adaptación del glosario Seguridad y propiedad intelectual en internet (Flores y Navarro, 2008) sobre la creación intelectual, orientada a los campus virtuales universitarios. En esta obra convergen los conocimientos de los campos jurídicos de propiedad intelectual y Derecho de internet, actualmente fragmentados y disconexos en distintos trabajos sectoriales.

El glosario original contiene 345 términos en español e inglés organizados alfabéticamente. Las entradas recogen, además del significado del término, no siempre en forma de definición, una gran cantidad de información del ámbito jurídico y tecnológico relacionada con el concepto o conceptos asociados al término, principalmente explicaciones didácticas, ejemplos, voces relacionadas y legislación (figura 2). La redacción de las entradas es libre y en ellas suelen participar varios autores, los cuales firman cada una de sus aportaciones. Al margen de su aportación científica, este glosario, así concebido, resulta poco sistemático, difícil de consultar, de mantener y explotar académicamente. El resultado de la AT de esta obra es el *Glosario explicativo e-Derecho*, actualmente de libre acceso en la dirección http://www.ucm.es/info/contratos y accesible también desde el campus virtual UCM.

Autoridad de certificación (Certification-service-provider)

La autoridad de certificación o prestador de servicios de certificación es el mediador de confianza en la seguridad de Internet, porque acredita la *autenticidad** de los sujetos que operan en Internet, de los actos por ellos suscritos con *firma electrónica reconocida** y de los documentos asociados a esta última modalidad de *firma electrónica avanzada**. Mediante la emisión de un *certificado**, el tercero acredita la identidad de las personas, así como la integridad de los actos asociados al certificado. El mediador de la confianza en la seguridad del *campus virtual* UCM es Verisign (w.w.w.verisign.com/CPS Incorp.by.Ref.LiaBILITY LTD.(c) 97 VeriSign). Así se acredita en el certificado emitido a favor de www.campusvirtual.ucm que puede visualizarse desde su página de inicio (https://www.campusvirtual.ucm.es/); vigente en la fecha en que se escriben estas páginas.

María de la Sierra Flores Doña

Figura 2. Ejemplo de entrada de la obra original[3]

Fase 1. La muestra seleccionada constituyó un 29,5 % (102 de 345) del total de entradas del glosario y se distribuyó de la forma siguiente:

- Muestra inicial: el 10,1% (35 entradas) pertenecientes a la letra C.
- Muestras de depuración: el 19,4% (67 entradas) restante eran entradas para la depuración separadas en 1) primera depuración (letras A y B); y 2) segunda depuración letra D.

Fase 2. Se pidió a los creadores del glosario, profesores de Derecho, que marcasen, con ayuda de un lingüista, el tipo de información que contenían las entradas de la muestra (figura 3). Para ello, utilizaron libremente su léxico habitual pero teniendo en cuenta dos restricciones intrínsecas al lenguaje de marcado elegido: i) el vocabulario (conjunto de marcas) debía ser lo más controlado posible[4]; y (ii) las marcas debían ajustarse al estándar XML[5]. Ambas restricciones eran imprescindibles para el procesamiento automático de las muestras: la exploración y la extracción automática del esquema subyacente común. Entender y aplicar estas restricciones no fue muy problemático para los autores del glosario, a pesar de no conocer nada de terminografía ni de XML y, además, contaron siempre para los casos conflictivos con el apoyo externo que les proporcionaba el lingüista.

[3] Por razones de espacio se ha sustituido parte del texto por puntos suspensivos.

[4] Por ejemplo, si se utilizaba la marca “regulación” los autores no debían usar términos equivalentes o variantes, como “leyes” o “regulaciones”.

[5] Esto significa que a) debían utilizar siempre una etiqueta de apertura y otra de cierre para delimitar cada parte o elemento estructural (siendo las etiquetas de la forma <elemento>Contenido</elemento>), y b) que el anidamiento de etiquetas debía ser coherente (es decir, si dentro de un elemento A se marca un elemento B, primero debe cerrarse B y luego A: <A><B>texto</B></A>.

```
<categoría letra="B">
<entrada>
<termino>base de datos</termino><ingles>database</ingles>
<concepto>
 
<definicion>Una base de datos o banco de datos es un conjunto de
su uso posterior.</definicion> <explicacion> Debido al gran desar
realizadas en formato electrónico, almacenándose y accediendo a e
.....</explicacion>
 <autor>José A. López Orozco</autor>
</concepto>
<regulacion>
<norma><articulo>Arts. 12.2</articulo>LPI<articulo>art. 29</artic
97</articulo><articulo>art. 98</articulo><articulo>art. 99</artic
101</articulo><articulo>art. 102</articulo><articulo>art. 103</ar
</regulacion>
<concepto>
<definicion> 
Definición.- Según la <norma>LPI</norma>, una base de datos es la
manera sistemática o metódica  accesibles individualmente por med
desde la base de datos digital que es accesible en internet, para
Madrid, hasta la "lista" con las marcas y goles de cada jugador o
de datos.</explicacion>
</concepto>
<regimen_juridico>
Esa colección puede reunir originalidad suficiente en la estructu
<voces_relacionadas>"obra*"</voces_relacionadas>. Si no es así, o
<voces_relacionadas>autor*</voces_relacionadas>, porque no hay <v
```

Figura 3. Ejemplo de etiquetado

Fase 3. El proceso de análisis y síntesis que se realiza en esta fase, con el apoyo de esquemas gráficos o con gramáticas independientes del contexto, debe representar todas las posibles estructuras de entrada de la muestra etiquetada en la fase anterior y, de ellas, extraer la más específica común a todas. Esta tarea fue realizada por los autores de este trabajo inicialmente a mano, y los resultados sirvieron de base para la definición de un algoritmo de extracción automática de esquemas y la construcción de una aplicación que permitiera obtener estos esquemas automáticamente a partir de las muestras XML (figura 4).

3_entidades.txt - Bloc de notas

Archivo Edición Formato Ver Ayuda

```
<!ELEMENT entrada (((lema,equivalente)|(equivalente))*,autor)>
<!ELEMENT tr (entrada)>
<!ELEMENT quien_lo_paga (norma+,(norma))>
<!ELEMENT cuanto_se_paga (norma)>
<!ELEMENT jurisprudencia (fuente,(autor)*)>
<!ELEMENT lema (siglas)>
<!ELEMENT presupuesto_objetivo (norma)>
<!ELEMENT prescripción (norma,latinismo)>
<!ELEMENT fragmentos_u_obra_completa (norma)>
<!ELEMENT fines_docentes_o_de_investigación (ejemplo,norma)>
<!ELEMENT solo_en_la_medida_justificada_por_el_fin_de_esa_incorporación (norma)>
<!ELEMENT reseñas_de_prensa (norma)>
<!ELEMENT razon_de_ser (xr)>
<!ELEMENT obras_ya_divulgadas (xr+,(xr+))>
<!ELEMENT derecho_de_cita_y_campus_virtual (xr,ejemplo)>
<!ELEMENT ejemplo (xr)>
<!ELEMENT cuándo_hay_comunicación_pública (ejemplo+,(ejemplo+),norma)>
<!ELEMENT tipos_de_comunicación_pública (norma+,(norma),xr,norma)>
<!ELEMENT titulares_derivativos (latinismo)>
<!ELEMENT creaciones_proceso_producción_complejo (ejemplo+,(ejemplo),norma,ejemplo+)>
<!ELEMENT excepciones (ejemplo+,(ejemplo),xr,ejemplo,latinismo,ejemplo+,norma,xr)>
<!ELEMENT partes_del_contrato (latinismo+,(latinismo))>
<!ELEMENT objeto_del_contrato (latinismo+,(latinismo))>
<!ELEMENT copia_privada_digital (xr,norma)>
<!ELEMENT otras_disposiciones (norma,xr+)>
```

Figura 4. Extracción automática de la estructura común más específica de la muestra

Fase 4. El análisis de los datos obtenidos en las fases anteriores nos permitió ajustar el esquema atendiendo a nociones terminográficas básicas. El resultado fue la definición de un modelo de contenido basado en cuatro dimensiones de información -significado, enciclopedia, ámbito jurídica y tesauro- y en una jerarquía de cinco niveles de elementos de

estructura (Sierra et al. 2009). Las dimensiones proporcionan formas posibles de interpretación del contenido[6] (figura 6), mientras que la jerarquía muestra la microestructura del glosario (figura 5). A modo de ejemplo, podemos ver que los elementos del primer nivel de una *Entrada* son *Lema, Término Equivalente (TE), Acepción, Autor, Relación, Regulación, Régimen jurídico y Jurispruedencia*. Es posible encontrar elementos que tienen un contenido semi-estructurado y así, por ejemplo, el elemento *Texto*, tiene un contenido mixto formado por texto entremezclado con otros 9 tipos de elementos estructurales sin un esquema de organización fijo.

Las dimensiones y la jerarquía están relacionadas e integradas formando un único modelo, de modo que, por ejemplo, los elementos estructurales *Lema, TE, Acepción* y *Autor* junto con sus subelementos, excepto *Explicación*, pertenecen a la dimensión *Significado*. *Explicación* –y sus subniveles- conforman la dimensión *Enciclopedia*. Finalmente, la dimensión *Tesauro* se construye a partir del elemento *Relación*.

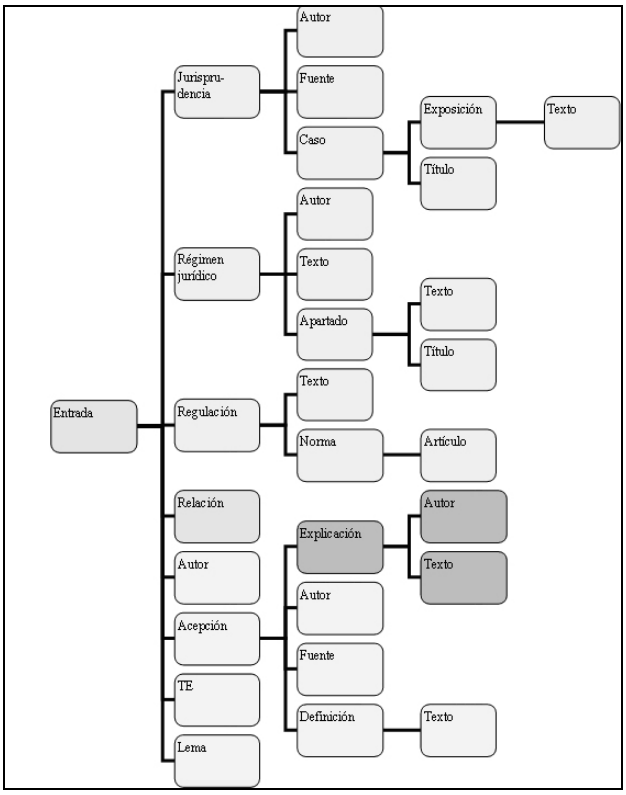

Figura 5. Niveles de la estructura jerárquica de las entradas del *Glosario explicativo e-Derecho*

[6] La dimensión Enciclopedia proporciona explicaciones y ejemplos que ayudan a entender el significado. La dimensión Ámbito Jurídico recoge toda la información de carácter jurídico que suplementa y complementa el significado del término: regulación, jurisprudencia, explicaciones, casos reales, etc. Y la dimensión Tesauro ofrece un mapa terminológico-conceptual del conjunto de términos del glosario para facilitar la localización y comprensión de la información que contiene el glosario.

Fase 5. En esta fase, los autores del glosario original aplicaron sobre las muestras de depuración los sucesivos modelos de contenido generados en la fase anterior para poner a prueba su capacidad descriptiva. Se volvieron a etiquetar las muestras de depuración con nuevas etiquetas, cuando fue necesario, o corrigiendo el modelo en los casos de inadecuación.. Se realizaron dos depuraciones –dos iteraciones de las fases 3, 4 y 5- para conseguir el modelo final del glosario.

Fase 6. En esta fase se utilizó un SGT basado en XML -TshwaneLex (http://tshwanedje.com/)- ya que permite el control de la consistencia de la información, el control de versiones y la exportación del contenido en diferentes formatos. Con esta herramienta se implementó el modelo de contenido del glosario. Esta fase presentó dos dificultades: (i) la adaptación del modelo de contenido a la estructura interna de la herramienta, y (ii) la introducción del texto de la obra original con su propio esquema de contenido, que se ha mantenido intacto.

El resultado se exportó en archivos XML que eran visualizados mediante una interfaz web. Esta interfaz se construyó con el objetivo de reflejar y explotar la organización en dimensiones y niveles del modelo del glosario. Las dimensiones, que se visualizan por medio de pestañas: Significado, Enciclopedia, Ámbito jurídico y Tesauro (figura 6), permiten el acceso directo a la información -el usuario busca exactamente lo que necesita en la dimensión adecuada, sin tener que leer todo el texto-, pero también permiten un acceso progresivo, que ayuda al usuario no experto en Derecho a ir comprendiendo y profundizando en la materia. De este modo, cuando se accede a un término se puede consultar independientemente (i) el *Significado*, (ii) la dimensión *Enciclopedia*, que añade una explicación y, en ocasiones, ejemplos, (iii) la dimensión *Ámbito jurídico*, que ofrece información organizada de carácter jurídico y (iv) la dimensión *Tesauro*, que ofrece al usuario la posibilidad de seguir explorando en la materia a través de los términos relacionados. El Tesauro se ha construido inductivamente con las voces relacionadas de la obra original.

Figura 6. Interfaz de acceso al contenido del glosario, con la dimensión *Tesauro* desplegada

4. CONCLUSIONES Y TRABAJO FUTURO

El presente trabajo da prueba de la posibilidad de obtener un *modelo de contenido* válido para una explotación electrónica eficiente de glosarios especializados previamente editados (en formato papel o digital), pero creados sin el entramado previo de estructuras lexicográficas que aseguren su homogeneidad y coherencia interna. Se trata, por tanto, del desarrollo de una metodología inductiva de extracción de estructuras de contenido y su interpretación lexicogáfica con el fin de posibilitar una alta accesibilidad a la información semántico-conceptual que los especialistas de la materia han consignado. La cantidad de información que contienen muchos de estos vocabularios especializados semiestructurados no puede ser explotada por medios informáticos si no existe una adecuación previa. La adecuación terminográfica que proponemos se basa en la extracción del modelo de contenido subyacente y la marcación completa y adecuada del contenido.

Aunque lo ideal es abordar el proyecto de creación de cualquier obre de este tipo partiendo de unos criterios formales fijos para la obra, este trabajo demuestra que es posible realizar una adaptación lexicográfica básica a posteriori que permita: (i) su tratamiento informático: fruto de este estudio es la creación del sitio web *Glosario explicativo E-derecho*, en el que es posible navegar por cuatro dimensiones –Significado. Enciclopedia. Ámbito jurídico. Tesauro- y realizar búsquedas de términos a través de la jerarquía estructural.

(ii) mayor accesibilidad del contenido, ya que, de otro modo, una gran parte de la información no podría ser consultada por los usuarios. Se rentabiliza enormemente su densidad informativa y las relaciones internas de sus elementos, opacos y ocultos a todo usuario que no realice la lectura completa del texto, pero sin modificar, en lo esencial, lo ya existente. De esta forma se mantiene el interés científico de estas obras y se permite a un mayor número de usuarios acceder a ellas.

En cuanto al método AT, hemos comprobado su viabilidad ya que su aplicación permite, encontrar una estructura implícita común a todas las entradas de una obra alfabetizada creada de forma libre aplicando un procedimiento inductivo. La adecuación terminográfica es aplicable, con mínimo esfuerzo, a la edición electrónica de materiales académicos alfabetizados, respetando la variedad y riqueza de los documentos originales.

La necesidad de contar con SGT generales y flexibles que permitan definir modelos de contenido para cualquier tipo de glosario es imprescindible para lograr un desarrollo efectivo de la propuesta metodológica que presentamos en estas páginas.

Entre las cuestiones pendientes, que constituyen nuestro trabajo actual y futuro, destacamos la evaluación de la eficacia del modelo de contenido obtenido con el método AT respecto de otros modelos posibles de carácter terminográfico (Schneider, 2008). En este sentido, hemos puesto en marcha un estudio empírico de la usabilidad del *Glosario*

explicativo e-Derecho[7] que permitirá identificar los modelos preferidos por los usuarios para realizar búsquedas y acceder al contenido del glosario.

BIBLIOGRAFÍA

Adelstein, Andreína & Cabré, María Teresa (2003): "Representación lexicográfica y terminográfica de las unidades terminológicas". En *Terminologia e Industrias da Lingua. Actas do VII Simposio Ibero-Americano de Terminologia RITerm*. Lisboa ILTEC, Unió Latina, Fundaçao Calouste Gulbenkian: 103-116.

Bray, Tim; Paoli, Jean & Sperberg-McQueen, C. Michael (2008): Extensible markup language (XML) 1.0 W3C Recommendation. Disponible en http://www.w3.org/TR

Chen, Shyh-Kwei; Lo, Ling-Ming; Wu, Kun-Lung; Yih, Jih-Shyr & Viehrig, Collen (2006): "A practical approach to extracting DTD-conforming XML documents from heterogeneous data sources". Information Science 176: 820-844. Disponible en www.sciencedirect.com.

Eki, Masaya; Ozono, Tadachika y Shintani, Toramatsu (2008). "Extracting XML schema from multiple implicit XML documents based on inductive reasoning". En *Proceeding of the 17th international conference on World Wide Web*, Poster Session. Beijing, China.

Flores, María de la Sierra & Navarro, Ruth (dirs.) (2008): *Seguridad y propiedad intelectual en internet.* Madrid: Editorial Complutense

Flores, María de la Sierra, Fernández-Pampillón, Ana, López , José Antonio & Matesanz, María (2009): "El Glosario e-derecho: un modelo empírico de información jurídica para la transmisión y comprensión del Derecho de Propiedad Intelectual en los campus virtuales universitarios" en *Memorias de la Octava Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI 2009, 10 al 13 de Julio de 2009, Orlando, Florida ~ EE.UU*. (Premio al mejor artículo de Sesión: Educación / Ética, Informática y Cibernética).

Garofalakis, Minos; Gionis, Aristides; Rastogi, Rajeev; Seshadri, S. & Shim, Kyuseok (2000): "XTRACT: A System for Extracting Document Type Descriptors from XML Documents". ACM SIGMOD Record Volume 29 , Issue 2 (June 2000): 165 – 176.

Jung, Jong-Seok; Oh, Dong-Ik; Kong, Yong-Hae y Ahn, Jong-Keun (2002): "Extracting information from XML documents by reverse generating a DTD". En M.H. Shafazand y A. M. Tjoa (eds.): *Proceedings of the First EurAsian Conference on Information and Communication Technology (EurAsia-ICT 2002); Lecture Notes In Computer Science, vol. 2510*: 314-321. London: Springer-Verlag. Disponible en www.springerlink.com

Sager, Juan Carlos (1990): *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

[7] Con el apoyo del Proyecto PIMCD 164/2009 de la UCM

Schneider, Roman (2008): “E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database”. *Sprache und Datenverarbeitung 1/08*.

Soergel, Dagobert (2002): “Thesauri and ontologies in digital libraries. Tutorial - Part 1: Structure and use in knowledge-based assistance to users. Part 2: Design, evaluation, and development”. En *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*.

Wright, Sue Ellen & Budin, Gerhard (1997): *Handbook of terminology management.* Amsterdam/Philadelphia: John Benjamins.