ASIGNACIÓN AUTOMÁTICA DE DESCRIPTORES AL DOCUMENTO PARLAMENTARIO. EL SOFTWARE DEL CENTRO COMÚN DE INVESTIGACIÓN DE LA COMISIÓN EUROPEA ADAPTADO AL CONGRESO DE LOS DIPUTADOS

VICTORIA FERNÁNDEZ MERA (*)

(*) Archivera-bibliotecaria de las Cortes Generales.

Desde que en julio de 1986, coincidiendo con el comienzo de la III Legislatura, se pusiera en marcha Argo, sistema de información de la actividad parlamentaria del Congreso de los Diputados, unos cuantos años han pasado y mucho han avanzado las tecnologías de la información.

La gestión y la consulta de la actividad parlamentaria ha cambiado enormemente, si aquellos primeros terminales «tontos», cuya consulta, necesariamente en modo experto, nos exigía como gestores y usuarios de información un conocimiento profundo del lenguaje documental utilizado, la actual informática documental nos ofrece aplicaciones capaces de buscar, organizar y recuperar información y contenidos a través del navegador Web, sobre interfaces amigables y sencillas.

Una evidente consecuencia de todo ello ha sido que la información sobre la actividad de la Cámara sea accesible en su consulta a todo el personal de la misma. En ese sentido, el profesional de la información ha dejado de ser imprescindible para poder tener un conocimiento básico de la actividad parlamentaria, no así para la gestión y el tratamiento de la misma y para aquellas búsquedas documentales cuya complejidad siempre requiere un mayor nivel de conocimientos.

En el ámbito de la aplicación de las nuevas tecnologías a la gestión de la información parlamentaria, el Congreso de los Diputados en colaboración con el *Joint Research Centre* (Centro Común de Investigación de la Comisión Europea) implantó un software de asignación automática de descriptores al documento parlamentario. De su gestación, desarrollo e instalación vamos a ocuparnos en las siguientes páginas.

1. El Centro Común de Investigación de la Comisión Europea

El Centro Común de Investigación (CCI) (*http://ec.europa.eu/dgs/jrc/*) es una Dirección General de la Comisión Europea que proporciona asesoramiento científico y tecnológico a los responsables de elaborar la política europea. Esta Institución está compuesta por siete institutos, cuyas actividades abarcan diversas áreas, con sede en distintos países europeos, concretamente en Alemania, Bélgica, España, Holanda e Italia.

En Ispra, Italia, tiene su sede el Instituto para la Seguridad y Protección del Ciudadano, donde desarrolla su actividad la unidad denominada Grupo de Tecnología del Lenguaje, (*http://langtech.jrc.it/)*. Este Grupo, integrado por lingüistas e informáticos, tiene como misión fundamental desarrollar y poner en marcha proyectos que luchen contra el crecimiento exponencial de la información y hagan frente a la barrera idiomática que supone la existencia de numerosas lenguas en el seno de la Unión Europea, sobre todo, tras la última ampliación.

Los proyectos desarrollados por el Grupo de Tecnología del Lenguaje del CCI se dirigen al análisis, filtrado y visualización de la información. Sobre la base de técnicas de aprendizaje automático, han elaborado aplicaciones que proporcionan al usuario un acceso a la información de forma condensada y para necesidades específicas.

Con este fin, han utilizado el tesauro plurilingüe Eurovoc, instrumento de indización desarrollado por el Parlamento Europeo en colaboración con la Oficina de Publicaciones, que abarca todos los ámbitos de actividad de la Unión Europea. La primera edición del tesauro se publicó en 1984 en siete lenguas. Desde entonces, Eurovoc se ha ido adaptando a la evolución y a los cambios geopolíticos y lingüísticos que han tenido lugar en el seno de la Unión, siendo objeto de siete actualizaciones y de varias revisiones lingüísticas, que han integrado, sucesivamente, las nuevas lenguas: nueve en 1987, once en 2000 y veintiuna en el 2006.

El Parlamento Europeo, la Oficina de Publicaciones de las Comunidades Europeas, varios parlamentos nacionales y algunos regionales de Europa, instituciones públicas nacionales y determinadas organizaciones europeas lo utilizan actualmente para la indización de sus documentos.

2. La indización de la actividad parlamentaria en el sistema de información Argo

Argo recoge toda la actividad del Congreso de los Diputados, siendo su fuente principal el Registro General de la Cámara. Se trata de una base de datos referencial, donde la información se organiza en unidades documentales que contienen todos los datos sobre la tramitación de una iniciativa.

El sistema Argo está estructurado en bases de datos independientes y cerradas para cada legislatura. Así, como resultado de la gestión informatizada de la actividad de la Cámara tenemos: III Leg. (1986-1989), IV Leg. (1989-1993), V Leg. (1993-1996), VI Leg. (1996-2000), VII Leg. (2000-2004) y VIII Leg. (2004-2008), y una base de datos abierta, la correspondiente a la Legislatura en curso.

El Archivo de la Cámara recuperó las legislaturas precedentes: Legislatura Constituyente (1977-1979), I Leg. (1979-1982) y II Leg. (1982-1986).

El documento que inicia el paso de la iniciativa por la Cámara lo denominamos documento tipo expediente, a él se van sumando todos los documentos que genera su tramitación. Es decir, en una pregunta al gobierno para respuesta escrita, el documento origen es la pregunta que formula el diputado, este escrito constituye el documento tipo expediente, abre una carpeta donde se incluyen la respuesta del Gobierno y otras incidencias que, en su caso, puedan ocurrir, como la retirada de la pregunta, solicitud de aclaración, prorroga de plazo para la contestación, etc. A este documento tipo expediente es al que asignamos los descriptores.

El Congreso de los Diputados, siguiendo la línea de actuación de aquellos parlamentos que informatizaron su actividad a mediados de los años ochenta –Cámara de Representantes de Bélgica o Asamblea de la República de Portugal– decidió adoptar Eurovoc para la indización de la actividad de la Cámara.

En 1987 se publicaba la segunda edición de Eurovoc, añadiéndose a las siete lenguas de origen, las versiones en español y en portugués.

Desde entonces, Eurovoc ha sido el lenguaje utilizado para la indización de los expedientes en Argo, siempre respetando su versión oficial, es decir, sin introducir ningún cambio en su estructura y contenidos.

No obstante, con el fin de poder asignar a las iniciativas parlamentarias descriptores geográficos, fue necesario desarrollar un tesauro complementario con los nombres de las provincias y municipios españoles. Este tesauro se elaboró a partir de la Relación de Municipios y Códigos por provincias que periódicamente publica el Instituto Nacional de Estadística (1).

Al comienzo de cada legislatura, la edición oficial de Eurovoc junto con el tesauro geográfico se cargan en la nueva base de datos. Si durante el transcurso de la misma se produce algún mantenimiento del tesauro que implique cambios en su contenido, nosotros no introducimos esos cambios hasta que no comienza la legislatura siguiente. La razón estriba en el gran número de documentos con los que tratamos, ya que cualquier cambio de descriptores nos obligaría a reindizar los documentos anteriores; además, el hecho de que trabajemos con bases de datos cerradas por legislatura, nos permite adoptar este criterio.

De este modo, las primeras cinco legislaturas fueron indizadas con la edición 2 de Eurovoc, las legislaturas V, VI y VII con la edición 3 y 3.1, la VIII con la edición 4.1 y en la Legislatura en curso, la X, estamos utilizando la edición 4.2. La indización de las Legislaturas Constituyente, I y II se realizó retrospectivamente en el Archivo de la Cámara.

La indización en Argo se realiza sobre una aplicación cliente-servidor con software tanto de desarrollo como operacional ORACLE. Esta aplicación, si bien permite la indización en línea al poder consultar y seleccionar directamente los términos del tesauro, tiene, sin embargo, como principal inconveniente la necesidad de teclear tanto el código numérico del expediente que se va a indizar como los descriptores asignados, cuando éstos no son seleccionados a través de la consulta en línea del tesauro. Además, no permite visualizar al mismo tiempo el

(1) *Relación de municipios y códigos por provincias a 1 de Enero de 2003*. Madrid, Instituto Nacional de Estadística, 2003.

texto del expediente que se está indizando y los descriptores que se van asignando al mismo.

Si bien la aplicación de indización que ofrecía Argo resultó eficaz y suficiente durante las primeras legislaturas, a medida que éstas se sucedían y se incrementaba la actividad parlamentaria, aquélla comenzó a resultarnos insuficiente y lenta.

El volumen de expedientes indizados ha ido aumentando con el paso de las legislaturas, si en un principio se indizaron una media de 30.000 expedientes por legislatura, este número aumentó con la VI Legislatura, en la cual ya hubo 58.112 expedientes, alcanzando casi el doble en la VII Legislatura con un total de 109.431 expedientes. En esta última se produjo un incremento del 88% en el número de expedientes indizados con respecto a la anterior, siendo este incremento del 130% en el número de preguntas para respuesta escrita.

Estos datos son bastante significativos del aumento progresivo que se estaba produciendo en la actividad parlamentaria y de su repercusión sobre la labor de indización de la misma, lo que nos hacia plantearnos la necesidad de buscar algún sistema de indización alternativo que, dada la carga de trabajo, facilitara y agilizara su realización, teniendo además en cuenta que el número de expedientes durante la VIII Legislatura iba a superar a la anterior, como de hecho así fue.

En esta situación nos encontrábamos cuando, en el 2003, conocimos el trabajo que había realizado el Grupo de Tecnología del Lenguaje del Centro Común de Investigación.

3. El software automático de indización del Centro Común de Investigación y su adaptación al Congreso de los Diputados

El software desarrollado por el Grupo de Tecnología del Lenguaje del Centro Común de Investigación se puede definir como un sistema automático de indización por asignación.

La indización por asignación se basa en la existencia de un corpus de textos manualmente indizados y en la selección de palabras clave

dentro de un lenguaje controlado, que describen los conceptos más importantes tratados en esos textos.

Es necesario que este corpus sea significativo en cuanto a su número. El software desarrollado por el Centro Común de Investigación se basó en una colección de textos integrada por unos 60.000 documentos en inglés, procedentes del Parlamento Europeo y de la Oficina de Publicaciones Oficiales de la Unión Europea, todos ellos indizados manualmente con Eurovoc.

Los textos que integran la colección que sirve como base para el aprendizaje del sistema son procesados con métodos estadísticos que descartan los términos vacíos, limitándose a aquellos con mayor peso según su frecuencia de aparición y su relación con los descriptores de Eurovoc. Estos términos, una vez normalizados, forman listas de términos asociados estadísticamente a cada uno de los descriptores que han sido previamente asignados en la indización manual a esos textos.

A partir de este momento, el software está preparado para asignar un descriptor de forma automática a un nuevo texto. Así, durante la fase de indización, el sistema analiza el nuevo texto y produce un lista de términos que compara con las listas ya existentes de términos asociados, si los términos encontrados existen como términos asociados dirigirán directamente a un descriptor dado, que el sistema asigna al nuevo texto. Si los términos extraídos del nuevo texto no son encontrados en las listas de términos asociados, el sistema no va a asignar los descriptores pertinentes a ese nuevo texto, por lo que será necesario introducirlos manualmente para que el software aprenda de ello.

El sistema, por lo tanto, aprende continuamente de la indización que se va realizando, por ello, los descriptores que el sistema selecciona automáticamente, han debido ser previamente asignados de forma manual por el indizador a un número determinado de textos, el sistema lo aprende y entonces es capaz de hacer su asignación automática.

Según el mayor o menor número de relaciones que se establezcan entre términos asociados y descriptores, éstos tendrán un mayor o menor nivel de pertinencia. En la presentación del interfaz de validación

del sistema, los descriptores asignados aparecen clasificados en un ranking, según el valor del peso entre el descriptor y los términos asociados que dirigen a él (2).

Este software fue presentado en Bruselas, en marzo de 2003, por Ralf Steinberger, director del Grupo de Tecnología del Lenguaje del Centro Común de Investigación, durante la reunión que, de forma anual, venía organizando el Parlamento Europeo sobre el tesauro Eurovoc (3).

En aquel momento, el señor Steinberger pidió la colaboración voluntaria de documentalistas formados en Eurovoc con el fin de evaluar los resultados de su proyecto. Sin dudarlo, el Departamento de Edición Oficial del Congreso de los Diputados, allí representado, se ofreció para revisar la asignación automática de descriptores sobre documentos del Parlamento Europeo y de la Oficina de Publicaciones Oficiales.

De este modo comenzó la colaboración entre el Congreso de los Diputados y el Centro Común de Investigación. El primer paso fue evaluar unos cien documentos, comprobando si los descriptores asignados automáticamente por el sistema eran pertinentes o no. Esta primera aproximación al software de asignación automática ya nos permitió apreciar que los resultados eran bastante aceptables.

---

(2) Para un conocimiento más técnico y detallado del sistema automático de indización desarrollado por el Grupo de Tecnología del Lenguaje del Centro Común de Investigación se pueden consultar las siguientes publicaciones:

– Bruno Pouliquen, Ralf Steinberger, Camelia Ignat (2003). *Automatic annotation of multilingual text collections with a conceptual thesaurus*. Workshop on Ontologies and Information Extraction. Held at EUROLAN. Romanian Academy of Sciences, Bucharest, Romania. 29 July 2003.

– Ralf Steinberger, Bruno Pouliquen (2003). *Cross-lingual Indexing*. Final Report for the IPSC Exploratory Research Project (4/2001-3/2003). JRC Internal Note. 30 pages. October 2003.

– Bruno Pouliquen (2004). *Automatic Eurovoc indexing: approach, evaluation and results*. JRC Workshop Addressing the Language Barrier Problem in the Enlarged EU - Automating Eurovoc Descriptor Assignment, JRC-Ispra, Italy, 16-17 September 2004.

(3) Eurovoc-2003 Conference (European Parliament): *Automating the assignment of EUROVOC descriptors to text*. European Parliament, Brussels, Belgium. 7 March 2003.

Un segundo paso fue procesar nuestros propios documentos con el software desarrollado por el Centro Común de Investigación. Sin embargo, en este caso, los resultados no fueron satisfactorios. La razón del relativo fracaso estaba en lo que habían apuntado tanto Steinberger como Pouliquen en algunos de sus trabajos, al señalar que el comportamiento del software empeoraba significativamente cuando el texto indizado por el sistema era diferente del corpus que había servido para entrenar el software (4).

Y la experiencia demostró que así era. Los textos del Parlamento Europeo y de la Oficina de Publicaciones Oficiales que habían sido la base de su sistema eran documentos muy extensos, extraídos de bases de datos en texto completo, a los cuales se había asignado una media de 5,65 descriptores. Por el contrario, los textos extraídos de Argo eran, en su mayoría, bastantes cortos, tres o cuatro líneas a lo sumo, y con un media de 2,5 descriptores por documento.

Por lo tanto, para que el software de asignación automática de descriptores funcionara correctamente debía ser entrenado sobre el mismo tipo de textos, manualmente indizados, que luego indizaría automáticamente, y así lo hicimos. El sistema desarrollado por el Centro Común de Investigación fue reconfigurado sobre la base de más de 80.000 iniciativas (5) extraídas de Argo, todas ellas indizadas con la edición 3 y 3.1 de Eurovoc.

Esta vez el resultado obtenido fue bastante bueno. La comparación entre la indización automática y la manual mostraba que en el 86% de los casos, los descriptores que habían sido asignados manualmente se encontraban entre los diez primeros descriptores que el software asignaba automáticamente.

Una nueva comprobación se hizo al validar la indización automática de más de 1000 preguntas escritas que no habían sido indizadas

---

(4) Ralf Steinberger, Bruno Pouliquen (2003), *Cross-lingual Indexing*, pp. 20-21.

(5) Se hizo una selección de aquellas iniciativas más relevantes cara a la indización: proyectos de ley, proposiciones no de ley, interpelaciones, preguntas orales y escritas y acuerdos internacionales.

previamente. Esta vez se trataba de una prueba de lo que sería el trabajo de indización en tiempo real, obteniendo, de nuevo, unos resultados con un alto nivel de acierto en la asignación de descriptores por el sistema.

Quedaba claro y demostrado, que el éxito del software automático de asignación de descriptores dependía de que éste aprendiera sobre la misma base documental que luego analizaría automáticamente. Esto es así, pero también creemos que otra condición importante es el grado de coherencia que se haya conseguido en la indización manual del corpus base del sistema; si al entrenar el software de asignación automática sobre las iniciativas extraídas de Argo, el resultado dio un porcentaje tan alto de aciertos en la asignación, pensamos que ello se debió, en gran medida, a que esa colección de textos base había sido indizada por una sola persona, por lo que había sido más fácil mantener un mayor grado de coherencia en la indización.

En el software de asignación automática se trabaja, fundamentalmente, con dos pantallas: un interfaz de indización del documento y otro de validación del mismo.

En el interfaz de indización se seleccionan los documentos que se van a indizar. La indización siempre está referida al tipo de iniciativa, de tal modo que el indizador, en esta primera pantalla, puede bien seleccionar el expediente de una iniciativa dada o, lo que es más corriente, un intervalo de expedientes o todos los expedientes que sobre un determinado tipo de iniciativa no estén indizados. El programa muestra entonces la lista de documentos seleccionados con los siguientes datos: número de expediente (184 / 74901/ 0000), objeto o texto (*Actuaciones del Ministerio de Cultura en relación a las intervenciones arqueológicas* ...), fecha de presentación de la iniciativa en el Registro General del Congreso de los Diputados y un botón final con el literal *index it*; al hacer clic sobre él, se indiza el documento en cuestión y se abre el interfaz de validación. En esta fase de indización, la aplicación, dependiendo de la extensión del texto, tarda aproximadamente un minuto.

El sistema también permite seleccionar los expedientes que sobre una determinada iniciativa ya estén inidizados, bien con el fin de

consultarlos o actualizarlos. En este caso, el botón final muestra el literal *see/update*, debajo un número entre paréntesis informa del número de descriptores con los que el texto ha sido indizado.

En el interfaz de validación encontramos la asignación automática de descriptores que ha realizado el sistema. Su presentación nos muestra de forma clara y amigable, a un lado y otro de la pantalla, el texto que se ha indizado y los descriptores a él asignados, haciendo clic sobre los que se consideran más apropiados, el indizador valida la previa asignación.

Desde esta pantalla podemos consultar el tesauro y seleccionar aquellos descriptores que el sistema no haya asignado de forma automática.

Un aspecto importante de este software es que, además de la indización automática de un texto, permite la preindización de un determinado número de documentos. La idea es que el programa «trabaje» durante la noche indizando aquellos documentos seleccionados, de tal manera que, a la mañana siguiente, el indizador se encuentre ya con los documentos *ready to index,* es así como el literal del botón en el interfaz de indización nos indica que ese documento ya ha sido preindizado; haciendo clic sobre él, el interfaz de validación se abre inmediatamente.

Si el número de documentos que se quiere preindizar no es muy grande, entre unos 30 o 40, esta preindización se puede realizar durante el mismo día de trabajo. El programa es rápido, pues en diez o quince minutos los documentos están preindizados.

Otra de las adaptaciones que se hicieron del software del Centro Común de Investigación con el fin de aplicarlo en el Congreso de los Diputados, fue la inclusión de un tesauro geográfico con el nombre de las provincias y municipios de España (6).

---

(6) *Relación de municipios y códigos por provincias a 1 de enero de 2003*. Madrid, Instituto Nacional de Estadística, 2003.

En este caso, el interfaz de validación no muestra por defecto estos descriptores geográficos, ya que no forman parte de Eurovoc, es necesario, en esta misma pantalla, hacer clic en el botón *show INE*, para que el interfaz que incluye la asignación de los nombres de los municipios se despliegue y seleccionemos los apropiados.

Al margen de la asignación automática de descriptores, el sistema ofrece información sobre la situación de la indización: documentos indizados, documentos no indizados y número de documentos indizados por usuario.

4. Integración del software automático de asignación de descriptores con el sistema de información Argo

Los trabajos de validación y entrenamiento del software sobre nuestra colección de documentos llevaron, aproximadamente, un año y medio; no fue hasta finales de 2004 cuando oficialmente se decidió instalar esta aplicación.

El 5 de junio de 2005 se firmó entre el Congreso de los Diputados y la Comisión Europea el Contrato nº JRC.BXL.180103, de licencia gratuita para uso de la aplicación informática *Software for the automatic or interactive categorisation of documents according to the Eurovoc thesaurus*.

Durante los días 21 al 24 de noviembre de 2005, Bruno Pouliquen, experto informático del Centro Común de Investigación y «padre» del software, junto con analistas del Centro informático del Congreso instalaron la aplicación en la Cámara, integrándola con el sistema de información Argo. En junio de 2006, se resolvieron algunas cuestiones que habían quedado pendientes en el momento de la instalación.

La aplicación de asignación automática esta disponible desde cualquier PC con un navegador Web en el Congreso de los Diputados (*http://nogal.congreso.es*). Cada usuario autorizado entra en ella con un login y una contraseña asociada.

El software ha quedado integrado con el sistema Argo, de manera que puede tomar de éste los datos necesarios para la indización de los expedientes (número de expediente, objeto de la iniciativa, fecha de presentación en el Registro General de la Cámara y, en su caso, descriptores asignados), y una vez que el expediente ha sido indizado, puede volcar esa información sobre Argo.

5. Conclusión

Desde noviembre de 2005, este software se ha utilizado en el Servicio de Tratamiento Documental del Departamento de Edición Oficial del Congreso de los Diputados, convirtiéndose en herramienta básica para la indización de la actividad parlamentaria. En concreto, durante la VIII Legislatura, se indizaron con ella un total de 163.885 expedientes.

Su utilización no ha supuesto el abandono de la aplicación de indización manual que proporciona Argo, en ningún momento la instalación de aquél implicaba la desaparición de este último. En la actualidad son, de hecho, dos instrumentos complementarios de trabajo.

La experiencia durante estos años nos permite hablar muy positivamente de este sistema de indización. Sus ventajas principales son rapidez, seguridad y coherencia, y ello por las siguientes razones:

- Proporciona una lista de candidatos a descriptores, clasificados según su nivel de pertinencia, el indizador sólo necesita escoger entre ellos aquellos que considera más adecuados, sin necesidad de teclear el texto del descriptor o descriptores.

- Al visualizar los descriptores que el sistema ha asignado automáticamente, el indizador se evita, en la mayor parte de los casos, recordar como un determinado tema ha sido indizado previamente y tener que consultar textos anteriores.

- No es necesario teclear el número del expediente del documento que deseamos indizar, lo cual no sólo ahorra tiempo sino que

también evita errores que se pueden producir al equivocar el número del expediente.

- El interfaz es muy amigable, tanto por su estructura que nos permite visualizar al mismo tiempo el texto del documento a indizar y los descriptores asignados, como por los colores y la letra utilizados.

- La coherencia en la indización queda asegurada por el continuo aprendizaje sobre la colección de documentos y la indización que se hace de ellos, es un importante aspecto teniendo en cuenta que, en la actualidad, son varias las personas que indizan la actividad parlamentaria.

La indización es una labor muy intelectual, por lo que desarrollar un software automático fiable en más de un 80% de los casos no es tarea fácil. La mayoría de los programas que indizan automáticamente, lo que hacen es extraer palabras clave de un texto dado, esto es, se limitan a seleccionar las palabras más relevantes del mismo. Este software va más allá, da un paso más, ya que es capaz de relacionar esos términos relevantes de un texto con los términos de un léxico controlado no presentes en él.

Finalmente, quisiéramos mencionar a aquellas personas que con su trabajo contribuyeron a hacer realidad este proyecto: Ricardo Blanco, jefe del Departamento de Edición Oficial de la Cámara que lo apoyó y defendió en todo momento; Fátima Melo, Millán Gómez y Antonio Sepúlveda, analistas del Centro Informático del Congreso de los Diputados, que integraron la aplicación con el sistema informático de la Cámara. También quisiéramos agradecer a la dirección del Centro Común de Investigación su generosidad al cedernos la licencia de uso, y al Grupo de Tecnología del Lenguaje, en particular, a Ralf Steinberger y a Bruno Pouliquen, pues sin su esfuerzo y su interés este proyecto nunca hubiera sido posible.