Fernanda Ribeiro e Maria Elisa Cerveira, org.

I Congresso ISKO Espanha e Portugal

XI Congreso ISKO España

7 a 9 de novembro de 2013

Informação e/ou Conhecimento: as duas faces de Jano

Atas

Faculdade de Letras da Universidade do Porto
CETAC.MEDIA
ISKO

Fernanda Ribeiro e Maria Elisa Cerveira, org.

I Congresso ISKO Espanha e Portugal

XI Congreso ISKO España

7 a 9 de novembro

Informação e/ou Conhecimento: as duas faces de Jano

Atas

Porto
Faculdade de Letras da Universidade do Porto
CETAC.MEDIA
2013

REPRESENTACIÓN DEL VOCABULARIO DE INDIZACIÓN EN HUMANIDADES CON UN GESTOR DE TESAUROS

La experiencia del léxico de las bases de datos ISOC en *TemaTres*

JOSÉ IGNACIO VIDAL LIY
Centro de Ciencias Humanas y Sociales - Unidad de Análisis y Producción de Bases de Datos

LUIS RODRÍGUEZ YUNTA
Centro de Ciencias Humanas y Sociales - Unidad de Análisis y Producción de Bases de Datos

ROSARIO DE ANDRÉS VERDÚ
Centro de Ciencias Humanas y Sociales - Unidad de Análisis y Producción de Bases de Datos

Resumen La siguiente ponencia presenta *Vocindario*, léxico elaborado por el área de Humanidades de las bases de datos ISOC. Se muestran sus utilidades actuales y su potencial para el futuro en cuanto que otorga más precisión y pertinencia a los términos de indización y búsqueda, absorbe con facilidad nuevos términos y conceptos, planteándose además como un léxico multidisciplinar común a varias áreas de la base de datos.

Palabras-clave Indización. Recuperación de la información. Lenguajes controlados. Tesauros. Bases de datos bibliográficas. Intercambio de datos. Multidisciplinariedad.

Abstract The following paper presents *Vocindario*, lexicon elaborated by the area of Humanities of the databases ISOC. His current usefulness and his potential appear for the future in all that grants more precision and relevancy to the terms of indexation and search, it absorbs with facility new terms and concepts, appearing in addition as a multidisciplinary common lexicon to several areas of the database.

Keywords Indexation. Information recovery. Controlled languages. Thesaurus. Referential databases. Exchange of data. Multidisciplinarity.

Introducción

Los intentos de adaptar el lenguaje natural a las búsquedas en bases de datos bibliográficas son una preocupación inherente a los sistemas de información. Esta problemática no ha terminado de resolverse ni siquiera con la emergencia de la web semántica y las ontologías, de tal manera que los problemas de precisión en el lenguaje (sinonimias, polisemias y ambigüedades) siguen afectando a menudo al carácter disperso y parcial de cualquier búsqueda. (Moreiro, 2013).

En la concepción tradicional de la recuperación booleana estas imprecisiones han podido minimizarse en las bases de datos bibliográficas merced a los necesarios lenguajes controlados que facilitan la indización y homogenizan la recuperación de la información a partir de la normalización y sistematización de la fase de introducción de datos. Los tesauros nacieron para cubrir las necesidades de información de los usuarios mediante un sistema de descriptores que a la vez sirve para indizar o representar el contenido de los documentos. Un tesauro tradicional se basa esencialmente en un conjunto de términos preferidos –descriptores- y no preferidos –sinónimos y cuasi-sinónimos- utilizados para representar un campo del conocimiento y/o representar el contenido de los documentos de un sistema de información. Los términos se regulan a través de relaciones de equivalencia, y se enriquecen con relaciones jerárquicas o asociativas que permiten expandir las opciones de recuperación (Currás, 2005; Lancaster, 2002; Lancaster, 1996).

A pesar del incremento de la complejidad de los sistemas de información y de los lenguajes documentales desde hace unos años con la irrupción de la web semántica, en las bases de datos bibliográficas o referenciales sigue siendo recomendable el empleo de lenguajes controlados tipo tesauros tradicionales o léxicos de indización. En primer lugar porque trabajan con datos de referencia bibliográfica y suelen carecer de documentos a texto completo. Y en segundo lugar, porque incluso ofreciendo la posibilidad de indización a texto completo, la utilización de un lenguaje controlado favorece la búsqueda al ayudar al usuario a explorar el fondo documental (Codina & Pedraza Jiménez, 2011).

En la base de datos ISOC se utilizan lenguajes controlados en la indización por materias asignadas a cada registro, dentro de su ya larga línea de trabajo sobre terminología y lenguajes documentales. Resultado de esta trayectoria ha sido la creación de tesauros y léxicos en diferentes disciplinas: Economía, Urbanismo, Psicología, Biblioteconomía, Historia Contemporánea y el tesauro de topónimos; mientras otras áreas trabajaban con léxicos de indización. La base de datos ISOC no ha contado nunca con un lenguaje controlado unificado para varias disciplinas, lo cual ha dificultado su aplicación en la interfaz de recuperación, ya que las decisiones tomadas en cada área disciplinar eran a menudo contradictorias. Como el producto final si era multidisciplinar y muchos documentos recibían clasificaciones de diferentes ámbitos, en la práctica las contradicciones en los criterios de indización se podían percibir en cada disciplina concreta. El objetivo de unificar los tesauros y lenguajes de indización utilizados, dio lugar a diferentes proyectos, pero no llegó nunca a concluir por falta de herramientas comunes de trabajo.

Pero, al margen de los lenguajes controlados, los propios registros de la base de datos también constituyen un recurso de interés para el análisis de la terminología empleada en Ciencias Sociales y Humanidades, ya que incorporan campos específicos para el análisis de contenido: resumen, clasificación, descriptores, identificadores, topónimos, legislación, jurisprudencia y periodo histórico.

Aprovechando la producción misma de la base de datos junto a la elaboración de lenguajes documentales y las posibilidades técnicas que ofrecen los actuales gestores electrónicos de tesauros y ontologías, el área de Humanidades de la Unidad de Análisis y Producción de Bases de Datos ISOC (Centro de Ciencias Humanas y Sociales - CSIC) ha elaborado *Vocindario*, un léxico controlado cuyos listados han sido extraídos de los índices empleados en la base de datos. Es resultado de una labor de sistematización y control de los descriptores utilizados en la base de datos (descriptores, identificadores y topónimos), si bien todavía se encuentra en fase de construcción. Las áreas implicadas hasta el momento son Antropología Cultural y Social, Arqueología, Bellas Artes e Historia, esperándose en un futuro próximo la inclusión de otras disciplinas. El vocabulario está accesible de forma gratuita en la sede web del Centro de Ciencias Humanas y Sociales del CSIC[1].

En la siguiente ponencia se pretende describir la experiencia del desarrollo de este léxico a partir de un gestor de tesauros, con la finalidad de mostrar la necesidad de una herramienta dinámica para el control del vocabulario y con ello llegar a una valoración más pertinente de las aplicaciones del vocabulario.

1 La base de datos ISOC: lenguajes documentales, características y problemas de la recuperación

La base de datos ISOC es el principal sistema analítico de información científica en Ciencias Humanas y Sociales en España como atestigua el millón anual de sesiones de consulta abiertas por los usuarios suscritos a las bases de datos del CSIC[2]. Forma parte de las bases de datos bibliográficas

[1] URL de consulta: http://archivos.cchs.csic.es/vocabularioisoc/vocab/index.php

[2] El uso de las bases de datos se mide por el número de sesiones de consulta abiertas por los usuarios suscritos a las bases de datos del CSIC. Datos referidos al año 2012 extraídos de http://www.investigacion.cchs.csic.es/isoc/

del CSIC (Rodríguez-Yunta, 2009) y se crea a comienzos de la década de los ochenta a partir de dos repertorios bibliográficos impresos: *Índice Español de Ciencias Sociales* (IECS) y el *Índice Español de Humanidades* (IEH). Desde entonces recoge toda la producción científica española desde 1975 en el área de Humanidades y Ciencias Sociales. Actualmente recoge casi 700.000 registros, y cumple con los principales requisitos de calidad referentes a la cobertura temática y la selección de documentos, a la par que ofrece un conjunto de valores añadidos que van desde el análisis de contenido hasta las posibilidades de recuperación de su interfaz de consulta[3].

Desde el año 2006 aplica un riguroso sistema de evaluación de revistas (utilizando como base de su sistema los criterios de evaluación *Latindex*), aunque lo que otorga un carácter distintivo es el enriquecimiento de su estructura de campos de contenido y que es su apuesta por una recuperación eficaz de la información ajustada a la especialidad científica, garantizando un equilibrio entre exhaustividad y pertinencia (Abejón Peña, Maldonado Martínez, Rodríguez Yunta, & Rubio Liniers, 2009).

La base de datos ISOC está formada por 16 subconjuntos correspondientes a las áreas de Bellas Artes, Biblioteconomía y Documentación, Educación, Antropología, Arqueología y Prehistoria, Filosofía, Geografía, Urbanismo, Historia, Derecho, Lengua, Literatura, Economía, Psicología, Sociología, Ciencias Políticas y América Latina. Cada documento incorporado se analiza y describe su contenido por indizadores especializados en la materia que emplean conceptos recogidos en lenguajes controlados de elaboración propia (léxicos, vocabularios de indización, tesauros) basados en la significación y el contexto. Los campos para la indización de contenidos son:

- Descriptores o conceptos representativos del tema de trabajo. Se están metiendo también en este campo los hechos históricos.
- Identificadores: nombres propios, instituciones con sede, títulos de obras. Se incluyen igualmente aquí los yacimientos arqueológicos.
- Topónimos: nombres geográficos de lugares, tanto físicos como divisiones administrativas y denominaciones históricas.
- Período histórico, décadas y siglos: datos numéricos que se utilizan para analizar y recuperar los artículos de tema histórico.
- Legislación y jurisprudencia: leyes y sentencias objetos de estudio en los trabajos jurídicos.
- Palabras clave de autor (sólo a partir de 2012 en un campo de uso interno hasta el momento).
- Aunque no es un campo de análisis la recuperación de la información también se puede efectuar a través de los resúmenes de autor incorporados en la ficha de los documentos.

De cara a los lenguajes de indización, esta división en campos ha determinado que los tesauros disciplinares construidos contemplasen exclusivamente los términos utilizados como descriptores. También se publicaron varias ediciones de un tesauro específico para los topónimos, pero limitado a las entidades político-administrativas (países, estados, provincias, municipios), mientras que en el mismo campo se han utilizado otras entradas para la localización geográfica (comarcas, cordilleras, ríos e incluso denominaciones históricas). Así pues, los tesauros disponibles en la base ISOC sólo afectaban a determinadas disciplinas e incluso en estos no recogían la

[3] Las bases de datos del CSIC disponen de dos accesos web, uno básico y gratuito en la dirección http://bddoc.csic.es:8080/, otro para suscriptores que incluye la interrogación de todos sus campos de análisis de contenido en http://bddoc.csic.es:8085/

estructura completa de los campos para la indización. Además, la presencia de campos numéricos para siglos y periodos, diseñados para facilitar las búsquedas por rango, ha justificado la exclusión de entradas de este tipo en el léxico, imprescindibles para el análisis de contenido en los artículos de Humanidades.

La recuperación de la información se puede efectuar o bien a cada una de las sub-bases mencionadas o bien a todo el conjunto, obteniendo de esta forma un resultado multidisciplinar. En cualquiera de los dos casos, existen las mismas tres modalidades de búsqueda:

La búsqueda simple, que es la más sencilla, es también la más imprecisa, pues en caso de que se haga la consulta al conjunto de la base, la respuesta producirá resultados de todas las sub-bases.

a) Una segunda forma es la búsqueda por campos, que permite combinar diferentes criterios y cruzarlos con una clasificación temática y/o con un año o intervalo temporal.
b) Un tercer y último modo de búsqueda es a través de índices (Autor, Revista, Descriptores, Identificadores, Topónimos, etc.) y que visibiliza en forma de listado alfabético el contenido de casi la totalidad de los campos que componen el registro. Es una opción muy útil para conocer el vocabulario utilizado en el análisis de los documentos y mejorar la precisión en la obtención de resultados.
c) La búsqueda experta mediante comandos y etiquetas de campo.

Gracias al hipertexto (campos de autor, descriptores, identificadores y topónimos) se pueden realizar nuevas búsquedas a partir de la visualización de un registro. Finalmente, cualquiera que haya sido la búsqueda, una vez que aparece en pantalla el listado de los documentos recuperados, se puede activar un filtro con el fin de visualizar las referencias con enlace al texto completo.

Es importante resaltar que, salvo en la búsqueda simple, en todas las demás modalidades de consulta juega un papel imprescindible la gestión de índices y el control del vocabulario. La presencia de índices por frase garantiza que es posible recuperar una entrada única por su sentido exacto sin que se entremezcle con otras entradas en las que un término está presente. Esto afecta a los campos de análisis de contenido y también a otros elementos como Autores o Títulos de revista, que es necesario presentar en forma de índice por frase para garantizar la precisión en la búsqueda. Así, por ejemplo sólo pueden recuperarse exactamente los registros de la revista "Historia social" sin mezclarse con otros títulos que contienen estas dos palabras ("Historia social y de la educación" o "Historia de la Comunicación Social"), si se ofrece al usuario un índice por frase que solo contiene entradas diferenciadas exactas.

La división sectorial de la base en conjuntos especializados más reducidos, supuso que en cada área se empleaba cada término desde el significado pertinente a cada disciplina (Abejón Peña, 1997). Consecuencia de ello fue el desarrollo de los problemas clásicos derivados de las ambigüedades, sinonimias y polisemias, y que afecta especialmente a la consulta general de la base. Así por ejemplo, si se hace una búsqueda general en texto libre por "Restauración", obtendrá simultáneamente resultados que tratan de restauración artística, otros sobre el período de la Restauración borbónica, referidos a la restauración eclesiástica o sobre restauración de edificios. De la misma manera, si se desea saber publicaciones en torno a la "Ilustración", lo más seguro es que encuentre mezclados resultados de la Ilustración como movimiento cultural e intelectual europeo, pero también relacionados con estampas, dibujos o grabados. Si en una disciplina como Historia se utiliza en la indización una entrada como "Restauración" o "Ilustración" con un significado único y restrictivo, es necesario coordinar criterios y asegurar que el mismo criterio se respeta y utiliza con el mismo sentido igualmente en otras disciplinas como Bellas Artes o Literatura.

El empleo de lenguajes controlados de elaboración propia en la base de datos ISOC no ha terminado de solventar estos inconvenientes, puesto que hasta el momento no se ha elaborado un

léxico controlado de carácter global y los que se disponen actualmente se ciñen al vocabulario de cada especialidad. Éste es uno de los objetivos de *Vocindario*, el léxico que aquí se presenta.

2 *Vocindario*: Una propuesta para la gestión de un vocabulario controlado general para las bases de datos ISOC

Más de treinta años de análisis documental permiten que la base de datos ISOC disponga de un vocabulario de indización de las áreas de Ciencias Sociales y Humanidades que ha sido imprescindible a la hora de elaborar herramientas y recursos documentales propios a los que ya se ha hecho referencia. De esta manera, este vocabulario es el que ha servido de base para la confección del léxico a partir del gestor de tesauros *TemaTres*, con el que se pretende crear por primera vez una herramienta para el análisis de contenido de manera global para varias disciplinas.

Son ya muy conocidas las aportaciones del entorno digital a la gestión y control electrónicos del vocabulario (Arano, 2005): enriquecimiento de la funcionalidad de la estructura de los tesauros a partir de la hipertextualidad, esto es, el establecimiento de hipervínculos entre los términos y las partes del vocabulario; reducción de costes de actualización y mantenimiento mediante software libre y licencias Creative Commons; posiblidad de integrar al usuario en el proceso de creación, gestión y optimización a través de testes, técnicas de modelado de usuario. El cuarto y último es quizá la aportación más interesante al permitir la posibilidad de aplicar medidas de reutilización e interoperabilidad en la planificación y construcción de tesauros, léxicos o vocabularios. Con ello se posibilita el aprovechamiento y enriquecimiento de la información conceptual y lingüística ya generada para otros recursos.

Precisamente éste último elemento ha sido el que se subyace al espíritu de *Vocindario*, pues uno de los fines que se pretende con el mismo es relacionar los índices con las diferentes clasificaciones del área de Humanidades de la base de datos. La aplicación que se escogió para ello, dadas sus múltiples ventajas y funcionalidades, fue *TemaTres*, software libre para la gestión de lenguajes documentales (Gonzáles Aguilar, Ramírez Posada, & Ferreyra, 2012). Con esta finalidad, se perseguía enriquecer u orientar la búsqueda temática en la base de datos ISOC, para las personas con acceso a la versión completa del producto por pertenecer a una entidad suscriptora de las Bases de datos documentales del CSIC. Además, *TemaTres* permite descargar, compartir o reutilizar datos de este vocabulario en condiciones de licencia Creative Commons Reconocimiento – No Comercial (by-nc).

Vocindario recoge los términos de indización utilizados en la base ISOC a partir de ciertos márgenes de frecuencia. No es un tesauro propiamente dicho aunque se presenta con una estructura jerárquica similar a este tipo de lenguajes. Así, las relaciones de equivalencia (USE/UP) identifican los términos no utilizados o palabras clave de autor que pueden traspasarse a un término preferido sí utilizado, por ser sinónimo, cuasi-sinónimo o variantes formales; las jerárquicas (TGP/TEP) refieren a la relación genérico-específico entre clasificación y términos de indización utilizados en el campo de Descriptores, mientras que las instancias (TGI/TEI) remiten a la relación genérico-específico entre clasificación y términos de indización utilizados en el campo de Identificadores. Se utiliza una estructura multijerárquica para presentar cada entrada dentro de todas las diferentes agrupaciones de la clasificación en las cuáles ha sido utilizado con una frecuencia suficiente (10 registros para los descriptores, 5 para los identificadores).

La página de inicio del *Vocindario* ofrece tres opciones de navegación (fig. 1) para los términos genéricos (que se corresponden con las clasificaciones utilizadas en la base de datos ISOC):

- por campos del análisis de contenido: clasificación, descriptores, identificadores, palabras clave de autor, periodos históricos y topónimos;
- por disciplinas, extraídas de la clasificación utilizada en la base de datos y

- por términos que incluyen aclaraciones útiles para la búsqueda, una agrupación que pretende destacar aquellos registros de términos que cuentan con notas de alcance de interés para los usuarios.

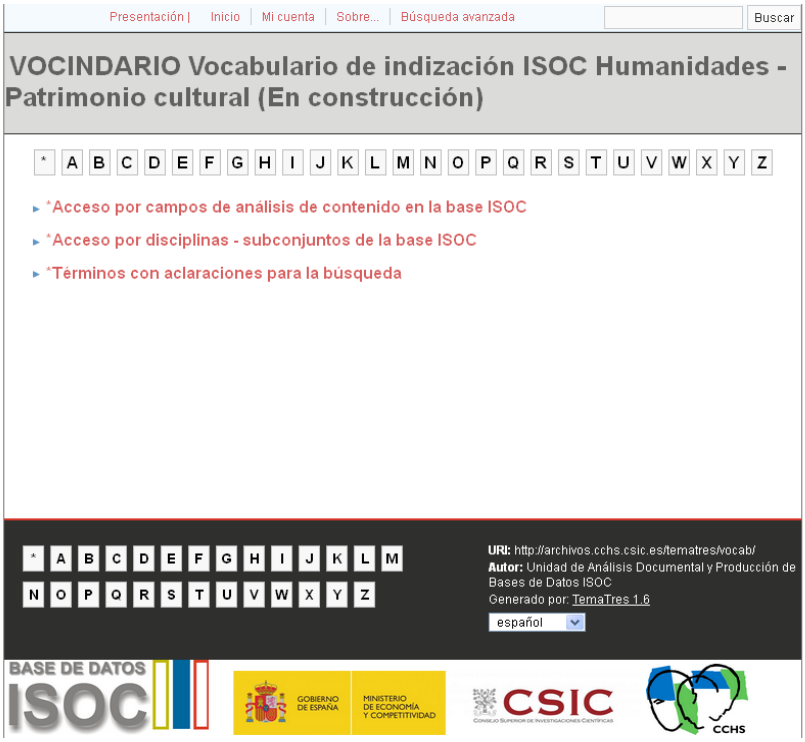

Fig.1: Página de inicio de la interfaz de consulta de *Vocindario* con las tres opciones de navegación

El vocabulario refleja todas las opciones de análisis documental de contenido empleadas en la base ISOC. Para los descriptores e identificadores, se parte de la estructura de las clasificaciones ISOC, agrupando algunos de sus epígrafes para obtener conjuntos con la mayor coherencia posible y que se correspondan con un mínimo de 100 documentos. En cada una de estas entradas se indica que se trata de clasificaciones y en nota de alcance se hace constar su correspondencia dentro de la base de datos (fig. 2).

Se insertan como términos específicos aquellos descriptores utilizados al menos en 10 registros en cada conjunto definido a partir de la clasificación. Para distinguir este tipo de entradas se utiliza la etiqueta de TEP "término específico partitivo" de la que dispone el programa *TemaTres*. En el caso de los identificadores se ha establecido como límite la presencia en al menos 5 registros en cada agrupación de la clasificación, y se distingue con la marca "TEI (Término Específico Instancia)". Para los topónimos se ha contemplado la frecuencia de 10 registros, pero se analiza su uso a nivel global para el conjunto de las disciplinas que abarca este vocabulario.

Entre los descriptores se incluye una familia especial para reflejar entradas de uso general que permitían recoger de forma diferenciada dos tipos de problemas:

- Términos de uso poco frecuente, aquellos que no superaban los 10 registros dentro de ninguna familia, pero sí en el conjunto de las disciplinas agrupadas en el vocabulario. A 15 de mayo de 2013 esta agrupación reunía 3134 términos.

- Términos modificadores de uso frecuente, aquellos términos de indización que pueden figurar indistintamente en cualquiera de las clasificaciones (están presentes en 10 o más de las agrupaciones contempladas en este vocabulario) y suelen emplearse con un significado secundario, similar a un modificador de otro término, por ejemplo "Origen", "Localización", "Datos biográficos" o "Descripción". A 15 de mayo de 2013 se había asignado esta categoría a 28 entradas.

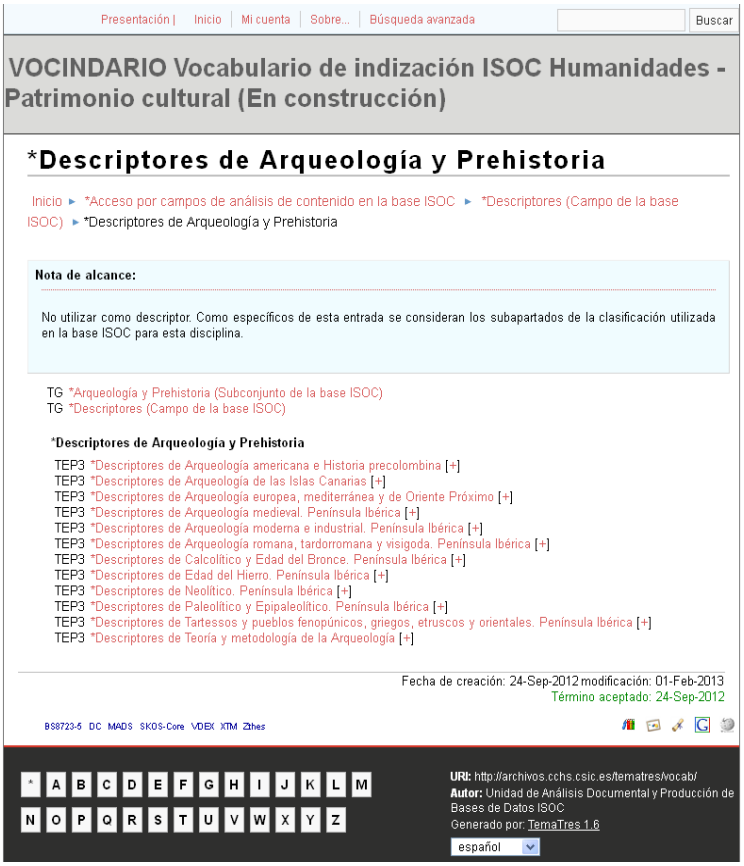

Fig. 2: Ejemplo de presentación de las agrupaciones de la clasificación de Arqueología en la interfaz de consulta de *Vocindario*

Respecto a las palabras clave de autor se han introducido en la base de datos ISOC a nivel interno a partir de las publicaciones editadas en 2012. Por ello aún es pronto para tener un acopio suficiente de términos, de modo que en *Vocindario* inicialmente sólo se han incorporado los términos utilizados por los autores en al menos 5 registros en el área de Historia. Este apartado podrá enriquecerse con nuevas entradas en próximas actualizaciones del vocabulario, según vayan recogiéndose más registros con esta información.

Finalmente, de forma ocasional se han introducido como términos candidatos (sin asignación de familia) algunos descriptores utilizados menos de 10 veces en la base de datos, especialmente para recoger en su nota de alcance las posibles variantes que pueden crear duda en la forma de la entrada.

Por último, hay términos que tienen la marca del asterisco (*) el cual distingue las entradas añadidas para las agrupaciones de términos en la navegación por campos y disciplinas. Todas las entradas que comienzan por asterisco son no-descriptores. Esta estrategia se inspira en la aplicada en el Tesauro de Patrimonio Histórico Andaluz (López Hernández, 2000), que incluye términos estructurales para denominaciones arbitrarias de familias, niveladores e indicadores de faceta.

En la relación de palabras clave de autor se aplica exclusivamente a las que no se corresponden con términos empleados en ISOC. También se asigna el asterisco para los apartados destinados a período histórico, décadas y siglos. En este vocabulario se han introducido entradas como "Siglo XVIII" exclusivamente para poder aclarar que en la base de datos ISOC este aspecto se refleja de forma numérica en un campo específico y por tanto no debe emplearse en los descriptores.

A 15 de mayo de 2013 el *Vocindario* contiene:

- 15120 entradas, de las que 190 son entradas con asterisco, no usadas en la indización (véase explicación en el párrafo anterior) y 20 son términos candidatos.
- 4000 entradas que cuentan con más de un término genérico.
- 3564 relaciones entre términos.
- 933 términos equivalentes.
- 612 notas de alcance.

3 Aplicaciones de *Vocindario*: usos internos y externos

El proyecto permite dar al vocabulario utilizado en la base de datos ISOC un doble uso, como herramienta interna para la gestión y para mejorar la recuperación en este producto, así como para usos externos, a través de la exportación de datos que pueda facilitarse para que otros proyectos los analicen o integren con diferentes fuentes en otros contextos. En este sentido, describimos cuatro posibles ámbitos de aplicación de *Vocindario*:

a. Herramienta de gestión interna para los analistas o indizadores de la base de datos.

En la política de calidad y mejora continua del mantenimiento de una base de datos bibliográfica, es una preocupación constante la normalización en los términos utilizados en la indización durante el análisis de contenido. Este proceso pretende conseguir dos objetivos:

- Reducir el número de términos de los índices de materia, para facilitar la recuperación. Se pretende que un mismo tema de búsqueda que puede ser expresado con diferentes formas por los autores de los documentos, se concentre en entradas únicas o en una combinación razonable de términos, en la medida en que sea posible hacerlo sin pérdida de significado sustancial.
- Evitar la ambigüedad de los términos que pueden tener varias interpretaciones, asignándoles un único significado y utilizando de forma sistemática entradas más precisas cuando sea necesario. Con ello se reduce el riesgo de ruido en la recuperación, al menos cuando esta se realice a través de los índices de materias.

Por ejemplo, para referirse al movimiento cultural del siglo XVIII se utiliza como término preferente la entrada "Ilustración" en lugar de "Siglo de las Luces" (considerado como término

equivalente y por consiguiente no-descriptor, eliminado de las opciones de indización). Pero este concepto conlleva problemas de ambigüedad en la recuperación, ya que los autores pueden utilizarlo con otro sentido en el ámbito del diseño gráfico. Para evitar la polisemia se admite una segunda entrada diferenciada más precisa para este segundo sentido: "Ilustración gráfica" (fig. 3). La diferenciación entre ambos conceptos se produce en la recuperación solamente si se realiza a través de los índices de materias. Su aplicación en la búsqueda a texto libre no puede garantizar la eficacia puesto que la forma léxica utilizada por los autores está sujeta a muchas variaciones.

Fig. 3: Ejemplo de búsqueda de entradas con el término "Ilustración" en la interfaz de consulta de *Vocindario*

La experiencia de los indizadores y su conocimiento de la materia son fundamentales para el mantenimiento de una política de indización en aquellos aspectos que precisan una toma de decisiones. No obstante, no es eficaz confiar su aplicación a la memoria, ya que pueden producirse olvidos y errores, y también el modo de trabajo debe adaptarse fácilmente a las sustituciones temporales o definitivas en el personal dedicado a estas tareas. En la Unidad de bases de datos ISOC se trabaja con un Manual de indización para uso interno, que marca criterios de procedimiento para una amplia gama de casos. Pero este manual no puede reflejar cada ejemplo y las dudas y dobles usos surgen con frecuencia. Por ello, es muy recomendable disponer de una herramienta donde reflejar las decisiones relativas a la indización, especialmente en aquellos temas, hechos, instituciones o personajes más tratados por la bibliografía. Un gestor de vocabularios es eficaz para ello en cuanto que permite establecer relaciones de equivalencia y explicitar el alcance de un término a través de las notas.

La puesta en marcha del proyecto ha servido para poner de manifiesto algunas deficiencias en el tratamiento de algunos temas que precisaban una depuración de variantes, desde los problemas aparentemente más simples (uso de singular y plural) hasta los problemas más complejos que precisan la consulta de diccionarios y obras de referencia para determinar la entrada más adecuada.

Igualmente, se ha contemplado la normalización con otros lenguajes documentales. Aunque en la base de datos ISOC no se toma como norma el fichero de autoridades de la Biblioteca Nacional de España, si se consulta habitualmente como fuente referencia igual que las autoridades del catálogo colectivo de las bibliotecas del CSIC. En este sentido se ha optado por la inclusión en *Vocindario* en las notas de alcance de los códigos VIAF cuando se han localizado entradas que efectivamente cuentan con una referencia normalizada en este sistema.

La inclusión de palabras clave de autor es una medida reciente en la base de datos ISOC. En un futuro próximo, esta herramienta puede utilizarse también para el establecimiento de puentes entre ambos modelos de indización, así como para la detección de los principales problemas de ambigüedad en la recuperación que pueden surgir a partir del uso de una indización sin control del vocabulario.

3.1 Herramienta abierta a los usuarios de la base de datos ISOC para mejorar la recuperación de información bibliográfica

La facilidad del gestor *TemaTres* para editar el vocabulario en la web, permite que los posibles usuarios de la base de datos utilicen esta herramienta para la preparación de una estrategia de búsqueda, de forma previa a su realización. De igual manera puede emplearse en cursos de formación o en presentaciones prácticas del producto.

A priori, puede parecer que para este uso sería preferible que la herramienta estuviera integrada dentro de la interfaz de interrogación de las bases de datos del CSIC. El hecho de que el acceso sea independiente, tiene sin embargo una ventaja: la presentación filtrada del vocabulario utilizado a partir de un umbral mínimo de frecuencia de uso y limitada a unas disciplinas concretas. La base de datos ISOC es un gran fichero multidisciplinar, lo cual sin duda es uno de sus principales valores, pero también un inconveniente a la hora de presentar un vocabulario. Por otra parte, los índices de materia tienden a crecer de forma constante, se pueblan de entradas que a menudo no se puede asegurar que sean eficaces para la recuperación. Por lo general, y salvo que se trate de un término emergente en el ámbito de la investigación en Humanidades, hay que tener precaución porque cuando una entrada está presente solamente en un número muy reducido de registros, o bien puede ser una errata, o bien una variante que puede estar expresada por otras alternativas en la propia base o bien no define una utilidad clara para la recuperación de información. La gestión del vocabulario en un programa externo a la interfaz permite una presentación más depurada de aquellas entradas que realmente sí permiten realizar una búsqueda con éxito en la base de datos.

La recuperación de información bibliográfica resulta muy sencilla solamente cuando un único término usado de forma sistemática resuelve la extracción de los registros pertinentes a una búsqueda. Pero a menudo pueden producirse necesidades que precisan hilar más fino o para las cuáles resulta conveniente conocer las características concretas del sistema de indización. Así por ejemplo:

- Temas que están presentes tanto en la clasificación como en los descriptores de materia. Los resultados pueden no ser idénticos y tampoco puede establecerse una recomendación general de si es preferible buscar mediante clasificación o mediante descriptores, o combinar ambas opciones. En ocasiones se le otorga un matiz distinto según el campo. Por ejemplo, el descriptor “Historia medieval” está usado en referencia a la subdisciplina en el contexto de la historiografía, mientras que para buscar estudios medievalistas debe utilizarse la clasificación. En otros casos, como “etnomusicología” se mantiene el mismo sentido entre descriptor y epígrafe de clasificación, pero en este último se asegura que tiene un sentido central en el tema del documento, mientras que en descriptores podría haberse reflejado como tema secundario.

- Periodos históricos que pueden buscarse a través de formas textuales o bien a través de fechas o siglos. En Prehistoria, Arqueología e Historia Antigua los periodos se expresan preferentemente a través de formas textuales (Neolítico, Alto Imperio,...) mientras que en Historia Medieval, Moderna y Contemporánea pueden utilizarse tanto formas textuales (Trienio constitucional, Primera República) como frecuentemente solo siglos y fechas. En este caso se deben utilizar los conceptos cuando se buscan estudios que aborden de forma pertinente el periodo histórico de que se trate, mientras que las búsquedas por fechas aportan mayor exhaustividad en cuanto al contexto histórico (personajes de la época, sociedad, cultura,...).
- Conceptos genéricos que aparecen expresados con frecuencia a través de otros más específicos. Así, por ejemplo, para realizar una búsqueda sobre Andalucía es demasiado restrictivo limitarse a esta entrada, que debe combinarse con las diferentes provincias. Por el contrario, no hace falta en general la interrogación por entradas más específicas (otras poblaciones andaluzas) ya que en la indización de la base ISOC se reflejan de forma sistemática las provincias tratadas en cada documento.
- Conceptos que se expresan a menudo por los autores con una construcción simple aunque para una mayor precisión es recomendable una construcción más compleja. Por ejemplo es frecuente que en título, resúmenes o palabras clave de autor se utilice la expresión "guerra civil" para referirse a la contienda del periodo 1936-1939 en España. Sin embargo, en descriptores se utiliza la fórmula más precisa "guerra civil española", indispensable para asegurar la pertinencia y eliminar la ambigüedad de la formulación simple.
- Palabra: Página de inicio de la interfaz de consulta de *Vocindario* con las tres opciones de navegación.

El vocabulario refleja todas las opciones de análisis documental de contenido empleadas en la base ISOC. Para los descriptores e identificadores, se parte de la estructura de las clasificaciones ISOC, agrupando algunos de sus epígrafes para obtener conjuntos con la mayor coherencia posible y que se correspondan con un mínimo de 100 documentos. En cada una de estas entradas se indica que se trata de clasificaciones y en nota de alcance se hace constar su correspondencia dentro de la base de datos (fig. 2).

Se insertan como términos específicos aquellos descriptores utilizados al menos en 10 registros en cada conjunto definido a partir de la clasificación. Para distinguir este tipo de entradas se utiliza la etiqueta de TEP "término específico partitivo" de la que dispone el programa *TemaTres*. En el caso de los identificadores se ha establecido como límite la presencia en al menos 5 registros en cada agrupación de la clasificación, y se distingue con la marca "TEI (Término Específico Instancia)". Para los topónimos se ha contemplado la frecuencia de 10 registros, pero se analiza su uso a nivel global para el conjunto de las disciplinas que abarca este vocabulario.

Entre los descriptores se incluye una familia especial para reflejar entradas de uso general que permitían recoger de forma diferenciada dos tipos de problemas:

- Términos de uso poco frecuente, aquellos que no superaban los 10 registros dentro de ninguna familia, pero sí en el conjunto de las disciplinas agrupadas en el vocabulario. A 15 de mayo de 2013 esta agrupación reunía 3134 términos.
- Términos modificadores de uso frecuente, aquellos términos de indización que pueden figurar indistintamente en cualquiera de las clasificaciones (están presentes en 10 o más de las agrupaciones contempladas en este vocabulario) y suelen emplearse con un significado secundario, similar a un modificador de otro término,

por ejemplo “Origen”, “Localización”, “Datos biográficos” o “Descripción”. A 15 de mayo de 2013 se había asignado esta categoría a 28 entradas.

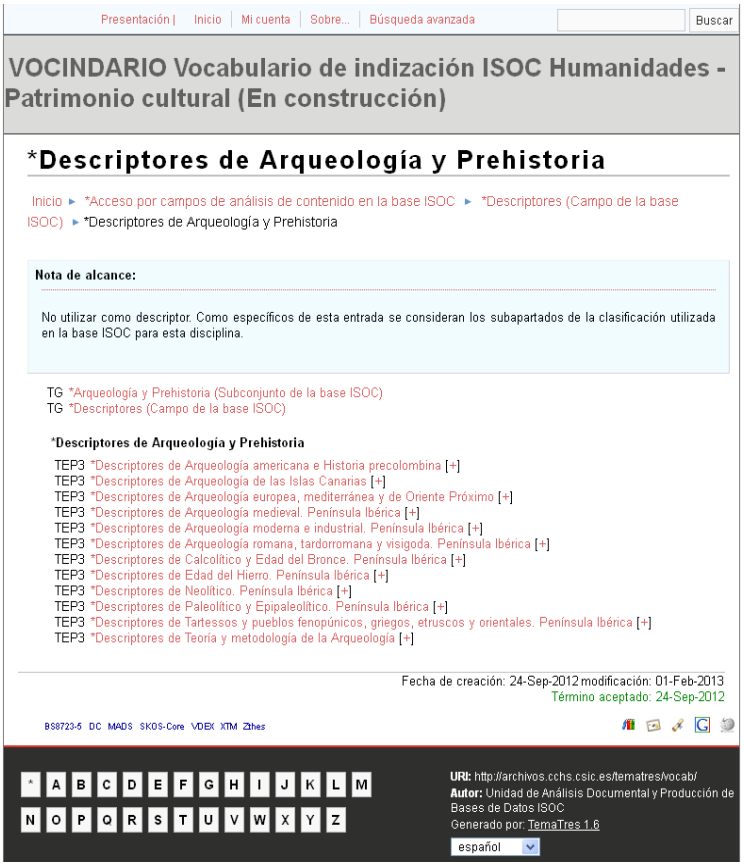

Fig. 4: Ejemplo de presentación de las agrupaciones de la clasificación de Arqueología en la interfaz de consulta de *Vocindario*

Respecto a las palabras clave de autor se han introducido en la base de datos ISOC a nivel interno a partir de las publicaciones editadas en 2012. Por ello aún es pronto para tener un acopio suficiente de términos, de modo que en *Vocindario* inicialmente sólo se han incorporado los términos utilizados por los autores en al menos 5 registros en el área de Historia. Este apartado podrá enriquecerse con nuevas entradas en próximas actualizaciones del vocabulario, según vayan recogiéndose más registros con esta información.

Finalmente, de forma ocasional se han introducido como términos candidatos (sin asignación de familia) algunos descriptores utilizados menos de 10 veces en la base de datos, especialmente para recoger en su nota de alcance las posibles variantes que pueden crear duda en la forma de la entrada.

Por último, hay términos que tienen la marca del asterisco (*) el cual distingue las entradas añadidas para las agrupaciones de términos en la navegación por campos y disciplinas. Todas las entradas que comienzan por asterisco son no-descriptores. Esta estrategia se inspira en la aplicada en el Tesauro de Patrimonio Histórico Andaluz (López Hernández, 2000), que incluye términos estructurales para denominaciones arbitrarias de familias, niveladores e indicadores de faceta.

En la relación de palabras clave de autor se aplica exclusivamente a las que no se corresponden con términos empleados en ISOC. También se asigna el asterisco para los apartados destinados a período histórico, décadas y siglos. En este vocabulario se han introducido entradas

como “Siglo XVIII” exclusivamente para poder aclarar que en la base de datos ISOC este aspecto se refleja de forma numérica en un campo específico y por tanto no debe emplearse en los descriptores.

A 15 de mayo de 2013 el *Vocindario* contiene:

- 15120 entradas, de las que 190 son entradas con asterisco, no usadas en la indización (véase explicación en el párrafo anterior) y 20 son términos candidatos.
- 4000 entradas que cuentan con más de un término genérico.
- 3564 relaciones entre términos.
- 933 términos equivalentes.
- 612 notas de alcance.

4 Aplicaciones de *Vocindario*: usos internos y externos

El proyecto permite dar al vocabulario utilizado en la base de datos ISOC un doble uso, como herramienta interna para la gestión y para mejorar la recuperación en este producto, así como para usos externos, a través de la exportación de datos que pueda facilitarse para que otros proyectos los analicen o integren con diferentes fuentes en otros contextos. En este sentido, describimos cuatro posibles ámbitos de aplicación de *Vocindario*:

4.1 Herramienta de gestión interna para los analistas o indizadores de la base de datos

En la política de calidad y mejora continua del mantenimiento de una base de datos bibliográfica, es una preocupación constante la normalización en los términos utilizados en la indización durante el análisis de contenido. Este proceso pretende conseguir dos objetivos:

- Reducir el número de términos de los índices de materia, para facilitar la recuperación. Se pretende que un mismo tema de búsqueda que puede ser expresado con diferentes formas por los autores de los documentos, se concentre en entradas únicas o en una combinación razonable de términos, en la medida en que sea posible hacerlo sin pérdida de significado sustancial.
- Evitar la ambigüedad de los términos que pueden tener varias interpretaciones, asignándoles un único significado y utilizando de forma sistemática entradas más precisas cuando sea necesario. Con ello se reduce el riesgo de ruido en la recuperación, al menos cuando esta se realice a través de los índices de materias.

Por ejemplo, para referirse al movimiento cultural del siglo XVIII se utiliza como término preferente la entrada “Ilustración” en lugar de “Siglo de las Luces” (considerado como término equivalente y por consiguiente no-descriptor, eliminado de las opciones de indización). Pero este concepto conlleva problemas de ambigüedad en la recuperación, ya que los autores pueden utilizarlo con otro sentido en el ámbito del diseño gráfico. Para evitar la polisemia se admite una segunda entrada diferenciada más precisa para este segundo sentido: “Ilustración gráfica” (fig. 3). La diferenciación entre ambos conceptos se produce en la recuperación solamente si se realiza a través de los índices de materias. Su aplicación en la búsqueda a texto libre no puede garantizar la

eficacia puesto que la forma léxica utilizada por los autores está sujeta a muchas variaciones.

Fig. 5: Ejemplo de búsqueda de entradas con el término "Ilustración" en la interfaz de consulta de *Vocindario*

La experiencia de los indizadores y su conocimiento de la materia son fundamentales para el mantenimiento de una política de indización en aquellos aspectos que precisan una toma de decisiones. No obstante, no es eficaz confiar su aplicación a la memoria, ya que pueden producirse olvidos y errores, y también el modo de trabajo debe adaptarse fácilmente a las sustituciones temporales o definitivas en el personal dedicado a estas tareas. En la Unidad de bases de datos ISOC se trabaja con un Manual de indización para uso interno, que marca criterios de procedimiento para una amplia gama de casos. Pero este manual no puede reflejar cada ejemplo y las dudas y dobles usos surgen con frecuencia. Por ello, es muy recomendable disponer de una herramienta donde reflejar las decisiones relativas a la indización, especialmente en aquellos temas, hechos, instituciones o personajes más tratados por la bibliografía. Un gestor de vocabularios es eficaz para ello en cuanto que permite establecer relaciones de equivalencia y explicitar el alcance de un término a través de las notas.

La puesta en marcha del proyecto ha servido para poner de manifiesto algunas deficiencias en el tratamiento de algunos temas que precisaban una depuración de variantes, desde los problemas aparentemente más simples (uso de singular y plural) hasta los problemas más complejos que precisan la consulta de diccionarios y obras de referencia para determinar la entrada más adecuada.

Igualmente, se ha contemplado la normalización con otros lenguajes documentales. Aunque en la base de datos ISOC no se toma como norma el fichero de autoridades de la Biblioteca Nacional de España, si se consulta habitualmente como fuente referencia igual que las autoridades del catálogo colectivo de las bibliotecas del CSIC. En este sentido se ha optado por la inclusión en *Vocindario* en las notas de alcance de los códigos VIAF cuando se han localizado entradas que efectivamente cuentan con una referencia normalizada en este sistema.

La inclusión de palabras clave de autor es una medida reciente en la base de datos ISOC. En un futuro próximo, esta herramienta puede utilizarse también para el establecimiento de puentes entre ambos modelos de indización, así como para la detección de los principales problemas de ambigüedad en la recuperación que pueden surgir a partir del uso de una indización sin control del vocabulario.

4.2 Herramienta abierta a los usuarios de la base de datos ISOC para mejorar la recuperación de información bibliográfica.

La facilidad del gestor *TemaTres* para editar el vocabulario en la web, permite que los posibles usuarios de la base de datos utilicen esta herramienta para la preparación de una estrategia de búsqueda, de forma previa a su realización. De igual manera puede emplearse en cursos de formación o en presentaciones prácticas del producto.

A priori, puede parecer que para este uso sería preferible que la herramienta estuviera integrada dentro de la interfaz de interrogación de las bases de datos del CSIC. El hecho de que el acceso sea independiente, tiene sin embargo una ventaja: la presentación filtrada del vocabulario utilizado a partir de un umbral mínimo de frecuencia de uso y limitada a unas disciplinas concretas. La base de datos ISOC es un gran fichero multidisciplinar, lo cual sin duda es uno de sus principales valores, pero también un inconveniente a la hora de presentar un vocabulario. Por otra parte, los índices de materia tienden a crecer de forma constante, se pueblan de entradas que a menudo no se puede asegurar que sean eficaces para la recuperación. Por lo general, y salvo que se trate de un término emergente en el ámbito de la investigación en Humanidades, hay que tener precaución porque cuando una entrada está presente solamente en un número muy reducido de registros, o bien puede ser una errata, o bien una variante que puede estar expresada por otras alternativas en la propia base o bien no define una utilidad clara para la recuperación de información. La gestión del vocabulario en un programa externo a la interfaz permite una presentación más depurada de aquellas entradas que realmente sí permiten realizar una búsqueda con éxito en la base de datos.

La recuperación de información bibliográfica resulta muy sencilla solamente cuando un único término usado de forma sistemática resuelve la extracción de los registros pertinentes a una búsqueda. Pero a menudo pueden producirse necesidades que precisan hilar más fino o para las cuáles resulta conveniente conocer las características concretas del sistema de indización. Así por ejemplo:

- Temas que están presentes tanto en la clasificación como en los descriptores de materia. Los resultados pueden no ser idénticos y tampoco puede establecerse una recomendación general de si es preferible buscar mediante clasificación o mediante descriptores, o combinar ambas opciones. En ocasiones se le otorga un matiz distinto según el campo. Por ejemplo, el descriptor “Historia medieval” está usado en referencia a la subdisciplina en el contexto de la historiografía, mientras que para buscar estudios medievalistas debe utilizarse la clasificación. En otros casos, como “etnomusicología” se mantiene el mismo sentido entre descriptor y epígrafe de clasificación, pero en este último se asegura que tiene un sentido central en el tema del documento, mientras que en descriptores podría haberse reflejado como tema secundario.
- Periodos históricos que pueden buscarse a través de formas textuales o bien a través de fechas o siglos. En Prehistoria, Arqueología e Historia Antigua los periodos se expresan preferentemente a través de formas textuales (Neolítico, Alto Imperio,...) mientras que en Historia Medieval, Moderna y Contemporánea pueden utilizarse tanto formas textuales (Trienio constitucional, Primera República) como frecuentemente solo siglos y fechas. En este caso se deben utilizar los conceptos cuando se buscan estudios que aborden de forma pertinente el periodo histórico de

que se trate, mientras que las búsquedas por fechas aportan mayor exhaustividad en cuanto al contexto histórico (personajes de la época, sociedad, cultura,...).

- Conceptos genéricos que aparecen expresados con frecuencia a través de otros más específicos. Así, por ejemplo, para realizar una búsqueda sobre Andalucía es demasiado restrictivo limitarse a esta entrada, que debe combinarse con las diferentes provincias. Por el contrario, no hace falta en general la interrogación por entradas más específicas (otras poblaciones andaluzas) ya que en la indización de la base ISOC se reflejan de forma sistemática las provincias tratadas en cada documento.
- Conceptos que se expresan a menudo por los autores con una construcción simple aunque para una mayor precisión es recomendable una construcción más compleja. Por ejemplo es frecuente que en título, resúmenes o palabras clave de autor se utilice la expresión "guerra civil" para referirse a la contienda del periodo 1936-1939 en España. Sin embargo, en descriptores se utiliza la fórmula más precisa "guerra civil española", indispensable para asegurar la pertinencia y eliminar la ambigüedad de la formulación simple.
- Palabras que están presentes en diferentes términos, con diferente sentido. Así por ejemplo la entrada "Cuenca" se utiliza exclusivamente en el campo de topónimos para la ciudad manchega, pero la misma palabra puede aparecer en la búsqueda libre con el sentido de concepto geográfico (cuenca hidrográfica) formando parte de otras entradas como "Tajo (Cuenca)". En este tipo de ambigüedades solo la búsqueda a través de índices por frase puede garantizar la pertinencia en la recuperación.
- Conceptos relacionados que conviene emplear como alternativas en la recuperación si se busca la mayor exhaustividad posible. Por ejemplo si se emplea el descriptor "Pintura academicista" o "Pintura cubista" conviene buscar también por el movimiento en general "Academicismo" y "Cubismo" pues son términos relacionados, si bien no sinónimos exactos pero que puede ser necesarios para una recuperación más amplia sin una pérdida importante de pertinencia.

4.3 Recurso para los estudios terminológicos en Humanidades

Los estudios terminológicos en cualquier disciplina precisan utilizar corpus documentales que a menudo no resultan fáciles de reunir. Los lenguajes documentales permiten aportar un conjunto de entradas que definen con claridad expresiones formales de temas de investigación. Los términos de indización aportan listados de sintagmas nominales que garantizan la consistencia en la selección.

En esta modalidad de uso, *Vocindario* aporta una relación de términos relacionados con su contexto de uso, disciplinas o epígrafes de clasificación. Ello facilita la interpretación interdisciplinar del resultado, al permitir al usuario la visualización simultánea de los términos más utilizados para cualquiera de las áreas, así como las relaciones de cualquier de ellos con las disciplinas relacionadas en la base de datos. Así por ejemplo si se introduce el término «Arquitectura nazarí", el usuario sabrá que los resultados que obtendrá en la base de datos estarán clasificados en la base de datos en "Arqueología medieval", "Arte medieval" y "Teoría del Arte". Si se busca por "Transición política" (fig. 4), los resultados que obtendrá estarán clasificados casi en su totalidad en "Historia contemporánea", pero también en "Antropología social, económica y política", y en "Arte contemporáneo II (1945-Actualidad)". En caso de que se desee encontrar los resultados exclusivos de una clasificación concreta, se podrá posteriormente utilizando los filtros pertinentes en la base.

Fig. 6: Ejemplo de presentación de un término (Transición política) en la interfaz de consulta de Vocindario, con sus diferentes agrupaciones de clasificación de la base ISOC en las que aparece por uso

4.4 Exportación del vocabulario para su integración en otros lenguajes documentales.

Hoy en día, con la evolución de los lenguajes documentales tradicionales hacia herramientas de recuperación en el contexto de la web semántica (ontologías, tesauros con verbos,...), resulta recomendable disponer de opciones para la exportación en formatos de intercambio (Pérez Agüera, 2004). El programa *TemaTres* permite la extracción de listados en diferentes opciones: texto (txt), Zthes, Skos-Core, TopicMap (xtm), BS8723, IMS Vocabulary Definition Exchange (VDEX), Wordpress XML (WXR), Site Map y SQL (para función de Backup).

Skos (siglas de Simple Knowledge Organization System) es una iniciativa del W3C que puede emplearse para definir cualquier tipo de lenguaje documental en forma de aplicación tipo RDF. Este formato puede utilizarse igualmente en la importación de vocabularios en *TemaTres*.

La página de descripción del proyecto disponible en la web del Centro de Ciencias Humanas y Sociales[4] define la política en cuanto a la descarga o reutilización de los datos de este vocabulario

[4] http://www.investigacion.cchs.csic.es/isoc/es/node/25

bajo las condiciones de licencia Creative Commons Reconocimiento – No Comercial (by-nc). La exportación del vocabulario permite su posible integración en otras herramientas (gestión de bases de datos, catálogos bibliográficos, gestor documental, gestión de revistas electrónicas,...) o la suma con otras fuentes para constituir un corpus de conceptos relevantes en una disciplina.

Conclusiones

Vocindario está en fase de construcción y no es posible todavía hacer un balance general de los resultados obtenidos a partir de su puesta en marcha. Por el momento sólo podemos llegar a resultados parciales, siendo el más destacado la capacidad del léxico como herramienta para el control del vocabulario de indización. La creación de un vocabulario, aunque no se persiga el objetivo más ambicioso de construir un tesauro o una ontología, sí resulta ya de clara utilidad para sacar a la luz los problemas de desambiguación, otorgando más precisión y pertinencia a los términos de indización de la base de datos. De la misma manera, la posibilidad de absorber con facilidad los nuevos términos y conceptos que surgen en la investigación y que en las Humanidades es especialmente activa. En el futuro, este vocabulario y esta herramienta aspirarán a ser referente en el ámbito de la catalogación e indización institucional y nacional.

Frente a los trabajos con lenguajes documentales abordados anteriormente en la unidad ISOC, el proyecto *Vocindario* conlleva varias diferencias importantes:

- Cubre la totalidad de los campos de indización de la base de datos ISOC. Hasta ahora los tesauros generados a partir de la base de datos ISOC se limitaban a los descriptores, o trataban parcialmente el campo de topónimos. En este vocabulario se integran también los identificadores, se hace constar el tratamiento otorgado a los periodos históricos e incluso se reflejan las palabras clave de autor añadidas recientemente.
- Se plantea como un léxico multidisciplinar. Con ello, pretende ser una herramienta común a varias disciplinas que facilite la adopción de criterios comunes y la resolución de ambigüedades que pueden crearse cuando una entrada tiene diferentes sentidos en función del contexto. Por el contrario, los proyectos anteriores tuvieron un carácter disciplinar y ello no contribuyó a unificar criterios.
- Ubica cada término en la estructura jerárquica del vocabulario no por su sentido ontológico sino en relación con la disciplina en la que ha sido empleado. Esto supone una alternativa frente al tesauro tradicional: un concepto como "guerra civil española" en un tesauro quedaría ubicado exclusivamente como un término de Historia militar, mientras que el uso real muestra que se trata de una entrada indispensable en otras áreas como Historia política, religiosa o cultural.
- Establece filtros por uso, es decir no refleja necesariamente todos los términos de indización utilizados en la base de datos, sino solamente aquellos que superan un determinado número de registros (10 en descriptores y topónimos, 5 en identificadores y palabras clave de autor). Con ello, se puede asegurar que todas las entradas recogidas se corresponden con una búsqueda real en ISOC, mientras que en los tesauros se introducían algunos términos que no se correspondían con el uso real.

Por todas estas aportaciones, consideramos que *Vocindario* es un proyecto con un gran potencial, tanto para el uso interno de la base de datos ISOC como para su explotación externa. Las bases de datos bibliográficas precisan herramientas para el control del vocabulario, en donde reflejar aquellos criterios del análisis de contenido que pueden ayudar a mejorar la eficacia de las búsquedas en los casos en los que existen ambigüedades o imprecisiones por sinonimias y polisemias.

Referencias

ABEJÓN PEÑA, T. (1997). Normalización del lenguaje documental para la base de datos ISOC-Arte. *Museo. Revista de la Asociación Profesional de Museólogos de España; nº2* , 251-259. *www.apme.es/revista/museo02_251.pdf*

ABEJÓN PEÑA, T., MALDONADO MARTÍNEZ, Á., RODRÍGUEZ YUNTA, L., & RUBIO LINIERS, M. C. (2009). La base de datos ISOC como sistema de información y fuente para el análisis de las Ciencias Humanas y Sociales en España. *El Profesional de la Información , 18* (5), 521-528. HYPERLINK <http://hdl.handle.net/10261/28769" http://hdl.handle.net/10261/28769>.

ANTA CABREROS, Ceferina: Divulgación de la producción científica española a través de las Bases de Datos Bibliográficas del CSIC: La Base ISOC. En L. Rodríguez-Yunta, & E. Giménez-Toledo, *La documentación como servicio público. Estudios en homenaje a Adelaida Román* (págs. 177-200). Madrid: CSIC. HYPERLINK <http://hdl.handle.net/10261/40986" http://hdl.handle.net/10261/40986>.

ARANO, S. (2005). Los tesauros y las ontologías en la biblioteconomía y la documentación. *Hipertext.net* . HYPERLINK <"http://www.upf.edu/hipertextnet/numero-3/tesauros.html" http://www.upf.edu/hipertextnet/numero-3/tesauros.html>.

CODINA, L., & PEDRAZA JIMÉNEZ, R. (2011). Tesauros y ontologías en sistemas de información documental. *El Profesional de la Información* , 555-563. <www.lluiscodina.com/ontologiaTesauros_2011.pdf>.

CURRÁS, E. (2005). *Ontologías, taxonomía y tesauros: manual de construcción y uso.* Gijón: Trea.

Gonzáles AGUILAR, A., RAMÍREZ POSADA, M., & FERREYRA, D. (2012). TemaTres: Software para gestionar tesauros. *El Profesional de la Información , 21* (3), 319-325.

LANCASTER, F. (2002). *El control de vocabulario en la recuperación de información.* Valencia: Universidad.

LANCASTER, F. (1996). *Indización y resúmenes: teoría y práctica.* Buenos Aires: EB Publicaciones.

LÓPEZ HERNÁNDEZ, M. A. (2000). El Tesauro de Patrimonio Histórico Andaluz. Un reto institucional y metodológico. *PH: Boletín del Instituto Andaluz del Patrimonio Histórico* , nº 31, págs. 130-133. HYPERLINK <"http://www.iaph.es/revistaph/index.php/revistaph/article/view/1003/1003" http://www.iaph.es/revistaph/index.php/revistaph/article/view/1003/1003>.

MOREIRO, J. A. (2013). Hacia la primacía de los conceptos sobre los términos en los vocabularios para la Web semántica. *Anuario ThinkEPI , 7*, 173-177. HYPERLINK <"http://www.thinkepi.net/hacia-primacia-conceptos-terminos-vocabularios-web-semantica" http://www.thinkepi.net/hacia-primacia-conceptos-terminos-vocabularios-web-semantica >.

PÉREZ AGÜERA, J. R. (2004). Automatización de tesauros y su utilización en la web semántica. *BiD: textos universitaris de biblioteconomia i documentació* , (13). HYPERLINK <"http://www.ub.edu/bid/13perez2.htm" http://www.ub.edu/bid/13perez2.htm>.

RODRÍGUEZ-YUNTA, L. (2009). Las bases de datos documentales del CSIC en el desarrollo histórico del mercado de la información en España (desde sus antecedentes hasta 2008). En L. Rodríguez-Yunta, & E. Giménez-Toledo, *La documentación como servicio público. Estudios en homenaje a Adelaida Román* (págs. 133-174). Madrid: CSIC. HYPERLINK <"http://hdl.handle.net/10760/14820" http://hdl.handle.net/10760/14820>.

Tesauros editados ligados a la experiencia de la base de datos ISOC

ALCAIN PARTEARROYO, María Dolores (coord.) (1992). *Tesauro ISOC de psicología*. Madrid: CINDOC. Reeditado con actualizaciones en 1995.

MOCHÓN BEZARES, Gonzalo; SORLI ROJO, Ángela (2002). *Tesauro de Biblioteconomía y Documentación*. Madrid: CINDOC.

RUBIO LINIERS, María Cruz (1999). *Tesauro de historia contemporánea de España*. Madrid: CINDOC; ANABAD Castilla-La Mancha; Junta de Comunidades de Castilla La Mancha.

Tesauro ISOC de topónimos (1993). Madrid: CINDOC. Reeditado con actualizaciones en 1994, 1996 y 2003. Edición abreviada en 2004.

Tesauro ISOC de urbanismo (1992). Madrid: CINDOC – Instituto Vasco de Administración Pública.

VALVERDE LUNA, Ana María (coord.) (1992). *Tesauro ISOC de Economía*. Madrid: CINDOC. Reeditado con actualizaciones y edición multilingüe en 1995.

VILLAGRÁ RUBIO, Ángel (coord.) (2008). *Tesauro ISOC de Economía*. Madrid: CINDOC.