WikiYATE: Extracción de Candidatos a Término utilizando la Wikipedia

5 de marzo de 2014

Tabla de contenido

1 Introducción

El objeto de este documento es explicar el procedimiento necesario para extraer los candidatos a término de un fichero de texto y obtener una lista valorada de estos candidatos.

2 Esquema general de funcionamiento

De manera general, se puede decir que los pasos a seguir son los siguientes:

a) Obtener el texto en formato electrónico: texto plano (codificación utf8)
b) Procesar lingüísticamente dicho texto
c) Extraer los patrones terminológicos
d) Analizar con Wikipedia
e) Visualización de los resultados
f) Análisis de los resultados y eventualmente mejora de la definición de dominio y repetición del análisis con Wikipedia

Estas etapas se muestran esquemáticamente en la Figura 1. Esta figura muestra también otras posibilidades del programa de análisis con la Wikipedia pero que no se utilizan en este procedimiento.

El directorio de trabajo puede ser cualquier punto del ordenador local. Es requisito imprescindible tener acceso a la red del IULA (en particular a P:\IULA\CORPUS\UTILS\) y al servidor vivaldi.upf.edu (inútil intentarlo desde fuera de la UPF, a menos que se utilice el VPN). Todos los programas que se utilizan en este procedimiento se ejecutan desde una ventana MS-DOS. Por defecto todos los ficheros de texto tienen extensión “txt” y los ficheros con texto procesado lingüísticamente extensión “vrt”.

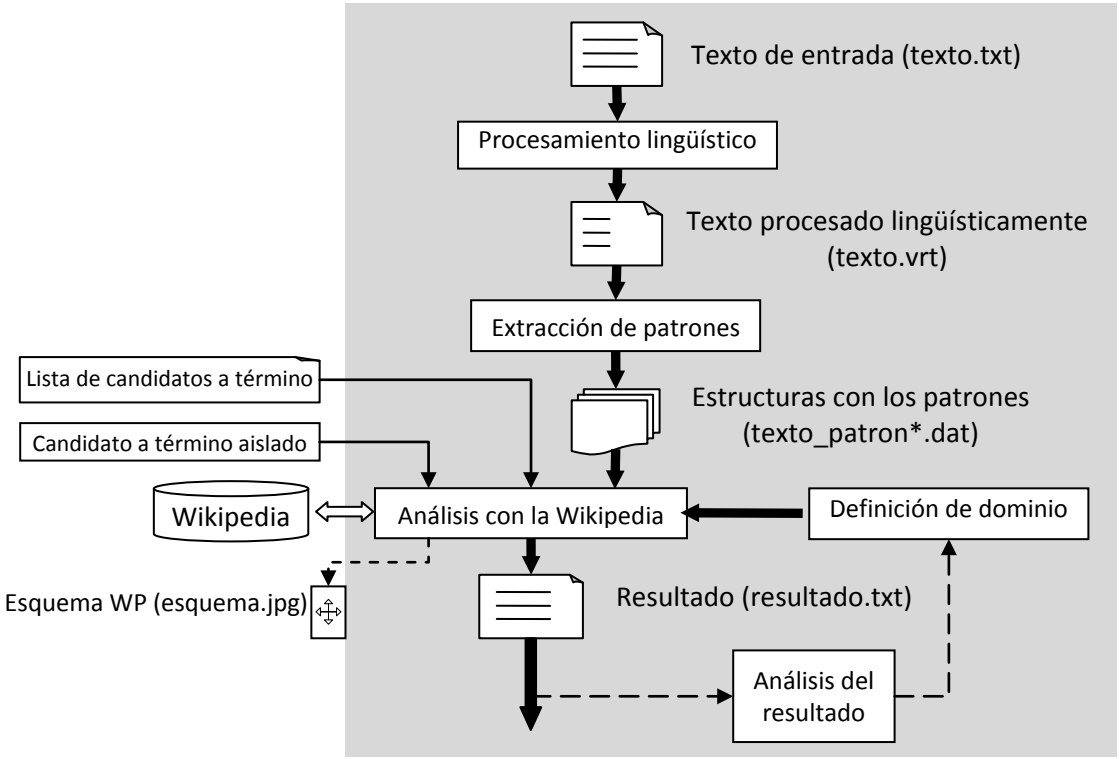
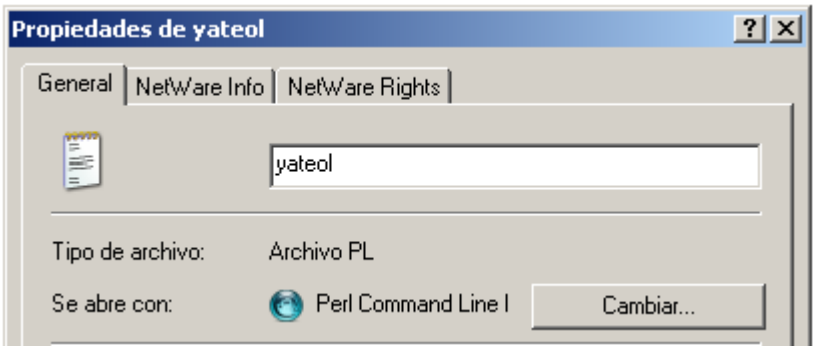


Figura 1. Esquema de procesamiento

3 Comprobaciones preliminares

Algunas de las comprobaciones preliminares se han realizado en el estadio anterior de procesamiento del texto (ver apartado 4 del documento *procesoCTaple2*), así pues ya se han aplicado. Aun así, en esta etapa de extracción de candidatos a término debemos comprobar algunos aspectos nuevos:

- Verificar que los archivos Perl (extensión `pl`) tienen asociado el intérprete Perl. Buscar con el administrador de archivos un programa `perl` (por ejemplo los que hay en `P:\IULA\CORPUS\UTILS\Yate`). En la ventana de propiedades (botón derecho del mouse y seleccionar Propiedades) se debe indicar que se abre con `Perl Command Line Interpreter`.



- Verificar que la variable `path` incluye el acceso al programa GraphViz: desde una ventana MS-DOS ejecutar el mandato `path`, el mensaje que devuelve debe incluir el string `P:\IULA\CORPUS\UTILS\graphviz\bin.` Útil para ver la estructura de la Wikipedia asociada a un candidato a término individual (sección 7.2.2).

```
C:\>path<return>

PATH=C:\Archivos de
programa\Graphviz2.36\bin;P:\IULA\SOFT\APLIIULA\Perl5\bin;C:\WI
NDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wbem;C:\WINDOWS\s
ystem32\nls;C:\WINDOWS\system32\nls\ENGLISH;c:\Archivos de
programa\Novell\ZENworks\;C:\Archivos de programa\
```

4 Selección de los textos y su conversión a texto plano

En principio, los textos que tenemos ya se han convertido en texto plano para el procesamiento lingüístico. Si necesitamos aplicar de nuevo la conversión de algunos textos podemos consultar el apartado 5 del documento anterior (*procesoCTaple2*).

5 Procesamiento lingüístico

Para obtener los patrones terminológicos de un texto es necesario procesar lingüísticamente dicho texto, aunque si hemos seguido los pasos del documento anterior (*procesoCTaple2*) estos pasos ya estarán completados y podremos pasar al apartado siguiente (6. Extracción de patrones terminológicos). Si no hemos procesado el texto, seguiremos las indicaciones siguientes.

Para procesar lingüísticamente el texto la orden a ejecutar es la siguiente:

```
C:\>p:<return>
P:\>cd \iula\corpus\utils\preproceso<return>
P:\IULA\CORPUS\UTILS\Preproceso>hector.pl -inputtext
doc.txt -outputtext doc.vrt -language lang -
annotationformat IULACT -keeptags<return>
```

Donde los parámetros más importantes de este programa son los siguientes:

- Input text: texto a procesar. Nombre del fichero con el texto plano a analizar (codificado como utf8 y con extensión "txt")
- outputtext: nombre del fichero resultado (será texto verticalizado con extensión "vrt")
- lang: idioma del texto a analizar (es, en o ca)
- annotationFormat: formato de la anotación (el valor debe ser IULACT)
- keeptags: mantiene eventuales tags SGML que se hubieran añadido manualmente

El programa tiene otras opciones; se puede obtener un listado completo de éstas ejecutando el programa: "`P:\IULA\CORPUS\UTILS\Preproceso\hector.pl -h`".

Para el fragmento de texto que se muestra a continuación:

> LA SITUACIÓN ECONÓMICA
>
> La pequeña localidad suiza de Davos se transforma cada año, hacia finales de enero, para acoger la reunión del World Economic Forum, uno de los encuentros de más alto nivel que se celebran a lo largo del año en Europa.

El resultado de este proceso después de aplicar este mandato:

```
C:\>p:<return>
P:\>cd \iula\corpus\corpus\tmpdret<return>
P:\IULA\CORPUS\CORPUS\TMPDRET>P:\IULA\CORPUS\UTILS\Preproceso
\hector.pl -language es -inputtext e00051.txt -outputtext
e00051.vrt -annotationformat IULACT<return>
```

Es un fichero generado (e00051.vrt) tiene el aspecto siguiente:

```
##      TAG   <div1>
##      TAG   <head>
1       TOK   LA          BOS   el\AFS
2       TOK   SITUACIÓN         situación\N5-FS
3       TOK   ECONÓMICA   EOS   económico\JQ--FS
##      TAG   </head>
##      TAG   <p>
##      TAG   <s>
5       TOK   La          BOS   el\AFS
6       TOK   pequeña           pequeño\JQ--FS
7       TOK   localidad         localidad\N5-FS
8       TOK   suiza             suiza\ JQ--FS
9       TOK   de                de\P
##      TAG   <name>
10      TOK   Davos             Davos\N4666
```

```
##      TAG    </name>
11      TOK    se                  pr\R6EZZZZ
12      TOK    transforma          transformar\VDR3S-
13      TOK    cada                cada\JN--6S
14      TOK    año                 año\N5-MS
…
```

Finalmente es necesario completar el procesamiento indexando el fichero resultado del procesamiento lingüístico. Para esto es necesario ejecutar el mandato siguiente:

```
P:\IULA\CORPUS\UTILS\indexSGMLplus.pl -i doc.vrt -o doc.5dx
```

6 Extracción de patrones terminológicos

La extracción de los patrones terminológicos se hace con el programa "filtroCAT.pl". Esta acción consiste en leer el resultado del análisis lingüístico y extraer las secuencias potencialmente terminológicas. Este programa dispone de una serie de opciones que pueden obtenerse invocándolo con la opción –h. El resultado que se debería obtener es el indicado en el Listado 1. Esta manera de invocar el programa también proporciona información sobre ejemplos básicos de uso, nombre y posición del fichero de configuración y valor de la opción –DocNoCtBaseDir (útil para procesar ficheros aislados).

```
C:\>p:<return>

P:\> cd \IULA\CORPUS\UTILS\YatePlus<return>

P:\IULA\CORPUS\UTILS\YatePlus>filtroCAT.pl -h<return>



Atención! leo el fichero de configuración del directorio local.
C:\Documents and Settings\U1480\Mis documentos\Dropbox\YATE\YATEplus\filtroCAT.pl
filtroCAT filtro de CAT para WikiYATE
 -i docCT            # Documento del CT o fichero aislado (Ojo! en este caso
                     # el directorio se fija en la opción -DocNoCtBaseDir
                     # o en el fichero de configuración)
 -o fileName         # Nombre base de los ficheros resultado (ej. -o m105)
 -DocNoCtBaseDir dir # Directorio donde encontrar los ficheros que no son del CT
                     # (reeemplaza el fijado por el fichero de configuración)
 -lang lang          # Idioma (es, ca, en)
 -prep prep          # Preposiciones (+ de 1)
 -arts               # Admite artículos interiores a un CAT (por defecto
                     # no los permite)
 -conjs              # Admite conjunciones interiores a un CAT (por defecto no)
 -lemaParticipio file # Fichero con la información para cambiar el lema de las H
 -borders file       # Fichero con la definición de Fronteras de Dominio en EWN
 -nBorders num       # Número absoluto de FD a leer (por defecto todas)
 -bordersValue num   # Valor mínimo (o igual) de probabilidad asociada a FD
                     # (por defecto todas)(valor 0->1, ej. 0.5)
 -propAdj file       # Fichero con la definición de las propiedades referidas
                     # por los adjetivos calificativos
 -combNA file        # Fichero con la definición de las combinaciones admitidas
                     # de FD y Propiedad del Adjetivo
 -webuser usuario    # Nombre del usuario (le envía un mensaje cuando
                     # termina el procesamiento)
 -v                  # Verbosidad (ej. -v 1)
 -help

DocNoCtBaseDir (fijado en el fichero de configuraci¾n (filtroCAT.conf)=
c:/Documents                        and                        Settings/U1480/Mis
documentos/Local/TreeTaggerAplicacion/Espanyol/

Ejemplos:
- Sobre un documento del corpus (admite s¾lo dos preposiciones):
  P:\IULA\CORPUS\UTILS\YatePlus\filtroCAT.pl -i m00105.sgm   -o m105      -prep de
-prep con
- Sobre un documento cualquiera (ya procesado lingüísticamente):
  C:\Documents                        and                        Settings\U1480\Mis
documentos\Dropbox\YATE\YATEplus\filtroCAT.pl  -i prueba.vrt -o prueba -lang es -
prep de
```

Listado 1. Opciones del programa filtroCAT.pl

Ejemplo de aplicación:

Si suponemos que tenemos un documento en español con nombre e00051.vrt , el mandato a ejecutar es el siguiente:

```
C:\>p:<return>
P:\>cd \iula\corpus\corpus\tmpdret<return>
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\YatePlus\filtroCAT.pl -i e00051.vrt -o
e00051 -prep de -DocNoCtBaseDir "./" -lang es<return>
```

La ejecución de este mandato presupone lo siguiente:

- El texto ya está procesado lingüísticamente (`*.vrt`)

- El idioma del texto es español (`-lang es`)
- La única preposición que no corta las secuencias terminológicas es "de" (`-prep de`)
- Todos los resultados se guardaran en un directorio que tiene el nombre e00051 (esta cadena también formará parte de los ficheros resultado)(`-o e00051`)
- Por defecto se considera que las fronteras de dominio son las de medicina. Esto no es necesario cambiarlo cuando se procesen documentos de otro dominio ya que utilizamos este programa sólo para obtener los patrones potencialmente terminológicos.
- El resto de opciones tienen el valor por defecto (ver Listado 1).

El resultado de esta ejecución será la creación del directorio e00051 cuyo contenido será el

```
C:\>p:<return>
P:\>cd \iula\corpus\corpus\tmpdret<return>
P:\IULA\CORPUS\CORPUS\TMPDRET>dir e00051<return>
 El volumen de la unidad P es WXP
 El número de serie del volumen es: 8A0D-5AD0

 Directorio de P:\IULA\CORPUS\CORPUS\TMPDRET\e00051

27/01/2014  13:20    <DIR>          .
27/01/2014  13:20    <DIR>          ..
28/01/2014  15:54           232.825 e00051.trm
28/01/2014  15:54            30.283 e00051_catslist.cats
28/01/2014  15:54           114.310 e00051_N.dat
28/01/2014  15:54            60.885 e00051_NJ.dat
28/01/2014  15:54            31.696 e00051_NPN.dat
28/01/2014  15:54            35.570 e00051_PATS.dat
               6 archivos        505.569 bytes
               2 dirs  135.147.859.968 bytes libres

P:\IULA\CORPUS\CORPUS\TMPDRET>
```

siguiente:

Los ficheros importantes de este resultado (y que conviene verificar su existencia) son los cuatro que tienen extensión "`.dat`".

El comportamiento de ciertos aspectos del programa se puede controlar también mediante el fichero de configuración (`filtroCAT.conf`).

7 Análisis con Wikipedia

Esta etapa se realiza con un programa que se encarga de buscar en una copia (en un servidor local) de la Wikipedia los candidatos localizados en la etapa anterior. Este programa se llama `exploraWPenAPyPP_5.pl` y está en `P:\IULA\CORPUS\UTILS\Wikipedia`.

Este programa dispone de una serie de opciones que pueden obtenerse invocándolo con la opción –h. El resultado que se debería obtener es el indicado en el Listado 2. Esta manera de invocar el programa también proporciona información sobre ejemplos básicos de uso.

Las opciones más relevantes son las que se indican a continuación:

- Entrada de candidato/s a término (excluyentes):
    - `-doc`: documento procesado por YATE
    - `-file`: lista de candidatos a término
    - `-cat`: candidato a término aislado. El resultado de esta opción se guarda en el directorio `InCatResults` (si no existe, el sistema lo crea en el directorio corriente)[1]
- Salida de resultados (opcional, sólo con opciones): `-outputdir` (permite procesar un mismo documento con distintos parámetros y guardar los resultados en directorios separados)
- Idioma de los candidatos: `-lang`
- Dominio/s respecto al cual queremos evaluar los candidatos: una o más apariciones de la opción `–dominio`.

```
C:\>p:<return>

P:\>cd \IULA\CORPUS\UTILS\Wikipedia<return>

P:\IULA\CORPUS\UTILS\Wikipedia>perl exploraWPenAPyPP_5.pl -h<return>

Script que analiza Candidatos a Término a través de la Wikipedia.
 -doc doc        ## Documento a procesar: ficheros *.dat de YATE (N. NJ y NPN) o YATEplus
 -indir dir      ## Directorio donde están los ficheros con los CATs a explorar (sólo con opción -doc
 -file filename      ## Fichero de texto (directorios in: InFiles, out InFilesResults)
                     ## con listas de términos. Formato: cat(    frec(      pos)?)?
 -fileDir dirName    ## directorio donde leer el fichero indicado en -file
 -cat cat            ## CAT puntual (resultados en directorio InCatResults)
 -inTermsEncoding encoding   ## codificación de los términos de entrada (iso-8859-1* | utf8)
 -resultsEncoding encoding   ## codificación de los ficheros resultado (iso-8859-1* | utf8)
 -catPOS pos         ## POS del CAT puntual
 -modo modo          ## (AP*|PP) modo de exploración del árbol de).
 -modoAP modo        ## (0*|1) submodo de exploración específico para AP
                     ## (0: busca hasta encontrar el top del dominio; 1: búsqueda full)
 -categs             ## Busca los CATs de las opciones de entrada como categorías de la WP
                         (no  páginas)
 -outdir dir         ## Directorio alternativo donde guardar los ficheros generados (sólo con
                         opciones -doc y -file)
 -dominio dominio    ## Categoría/s de la WP que representan el dominio
 -lang lang          ## Idioma (es|ca|en|it|pt)
 -desambMethod SOURCE    ## (file|wikipedia*) páginas de desambiguación resueltas por
                                 fichero/wikipedia
 -desambWPtype type   ## (CD|lmin*) desambigüa con la WP usando los CDs o bien la long.
                             mínima al dominio
 -comp             ## Permite el análisis de los componentes (en especial para multipalabras)
 -noWikiCategs     ## Descarta todos los caminos al top que lleguen a una categoría con
                         nombre Wikipedia*
 -lmincat #        ## Longitud mínima de un cat para ser analizado (3 por defecto)
 -wp2010           ## Usa la versión WP2010 (ES: Luis Cabrera)
 -wp2013           ## Usa la versión WP2013 (ES: INEX)
 -help             ## Esta ayuda
*: opción por defecto
Genera:
```

[1] Si el candidato a término a analizar es poliléxico debe indicarse entre comillas. También hay que hacer lo mismo con las categorías de la Wikipedia que se usan como dominio.

- (-doc|-file|-cat) + _patronesProcesadosOUT.txt;
- (-doc|-file|-cat)_Wikipedia.dat (utilizado por experimentoAIJplot.pl)
- (-doc|-file|-cat).log

Ejemplo de llamada:

- documento en español procesado con YATE/YATEplus
  exploraWPenAPyPP_5.pl -doc e00066 -lang es -noWikiCategs -dominio economía
- documento en inglés (no del CT) procesado por YATE
  exploraWPenAPyPP_5.pl -doc ecologyMcMillanEvaluation -lang en -noWikiCategs - dominio climatology - dominio Nature -dominio biology -dominio ecology -desambWPtype lmin -indir "C:\Documents and Settings\All Users\Documentos\My Dropbox\YATE\YATEplus\ecologyMcMillanEvaluation" -outdir ecologyMcMillanEvaluation_EcoBioClimNat
- Lista de CATs
  exploraWPenAPyPP_5.pl -file prueba.txt -lang es -noWikiCategs -dominio matemáticas
- Término individual en español
  exploraWPenAPyPP_5.pl -cat "memoria ram" -lang es -noWikiCategs -dominio informática -dominio electrónica

Listado 2. Opciones del programa `exploraWPenAPyPP_5.pl`

7.1 Selección de categorías

Este programa de análisis permite analizar candidatos a término en cualquier dominio a condición de que dicho dominio se pueda expresar como una combinación de una o más categorías de la Wikipedia. Esto puede hacer necesario repetir el experimento con diversas combinaciones de categorías a fin de optimizar los resultados.

En principio la definición del dominio puede limitarse simplemente al dominio más genérico con el cual asociamos el texto a analizar (p.ej., derecho, economía, etc.). En este caso, si existe una categoría de la Wikipedia con este nombre ya podemos realizar una primera prueba (ver sección 7.2.1) y el correspondiente análisis de los resultados obtenidos (ver sección 7.3).

En las secciones siguientes se indica con cierto detalle cómo realizar el procesamiento tanto para un documento completo (ver sección 7.2.1) como para un candidato a término aislado (ver sección 7.3) y el análisis del resultado obtenido.

Este programa se puede aplicar a textos en los siguientes idiomas: catalán, castellano, inglés, italiano y portugués.[2]

7.2 Procesamiento

7.2.1 A partir del resultado obtenido con la extracción de patrones

Si suponemos que tenemos un documento en español con nombre e00051 del cual ya hemos obtenido los patrones terminológicos y queremos analizar con la Wikipedia, el mandato a ejecutar es el siguiente:

```
C:\>p:<return>
P:\>cd \iula\corpus\corpus\tmpdret<return>
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\exploraWPenAPyPP_5.pl -doc e00051
-lang es -dominio derecho -indir "./e00051/" -wp2013<return>
```

[2] En el caso de italiano y portugués se supone que el texto tiene el mismo formato que los textos del CT del IULA y las etiquetas han sido transformadas a las del español (etiquetario IULA).

7.2.2 De un candidato a término aislado

Cuando el sistema no ha considerado como término una secuencia que consideramos es terminológica podemos analizarla individualmente para obtener su relación jerárquica con otras categorías para evaluar si, añadiendo alguna categoría a la definición de dominio, podemos mejorar el resultado.

Para hacer este tipo de análisis se debe utilizar el programa de análisis con la Wikipedia con la opción de entrada puntual de CATs (`-cat`) en lugar de hacerlo con el resultado de aplicar YATE. Usando el programa de esta manera los resultados se almacenan en el directorio `InCatResults`. Otra característica peculiar de usar el programa de esta manera es que genera un grafo con las categorías de la Wikipedia que puede ayudar a visualizar la estructura de categorías y su influencia en la aceptación/rechazo de una página en un dominio [3] [4].

Para hacer este tipo de análisis partimos de una ventana MS-DOS y se procede a analizar este candidato como indicamos a continuación:

```
C:\>p:<return>
P:\>cd \iula\corpus\corpus\tmpdret<return>
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\wikipedia\exploraWPenAPyPP_5.pl -cat
"derecho procesal" -lang es -dominio derecho -modo PP -wp2013
<return>
```

Como podemos ver en el script superior, si el candidato a término (`-cat`) que hemos aislado es poliléxico, los marcamos entre comillas, mientras que si se trata de un término monoléxico, no hacen falta las comillas. En el caso de la categoría de Wikipedia (`-dominio`) también la marcamos entre comillas si es poliléxica.

De acuerdo con estos resultados, y para mejorar la extracción de patrones (sección 7.2.1), podemos aplicar más de un dominio para detallar más la búsqueda. En el caso del ámbito del derecho canónico el sistema no consideraba terminológicas secuencias como *familia* o *heterosexual.* Si analizamos aisladamente una de ellas (por ejemplo, *heterosexual*) observamos que la categoría *comportamiento humano* podría funcionar como categoría complementaria para detectar estos términos.

[3] Para visualizar el grafo es imprescindible tener instalado (o tener acceso) al programa GraphViz. La variable PATH debe incluir el camino siguiente: `P:\IULA\CORPUS\UTILS\graphviz\bin`.

[4] Código de colores asociados a cada nodo categoría: rojo, define el dominio; amarillo, en el camino hacia el top; verde claro, nodos entre la definición del dominio y el top; sin color, no interviene en los caminos. Los nodos que representan página de la Wikipedia son rectangulares mientras que aquellos que representan categorías tienen forma de óvalo.

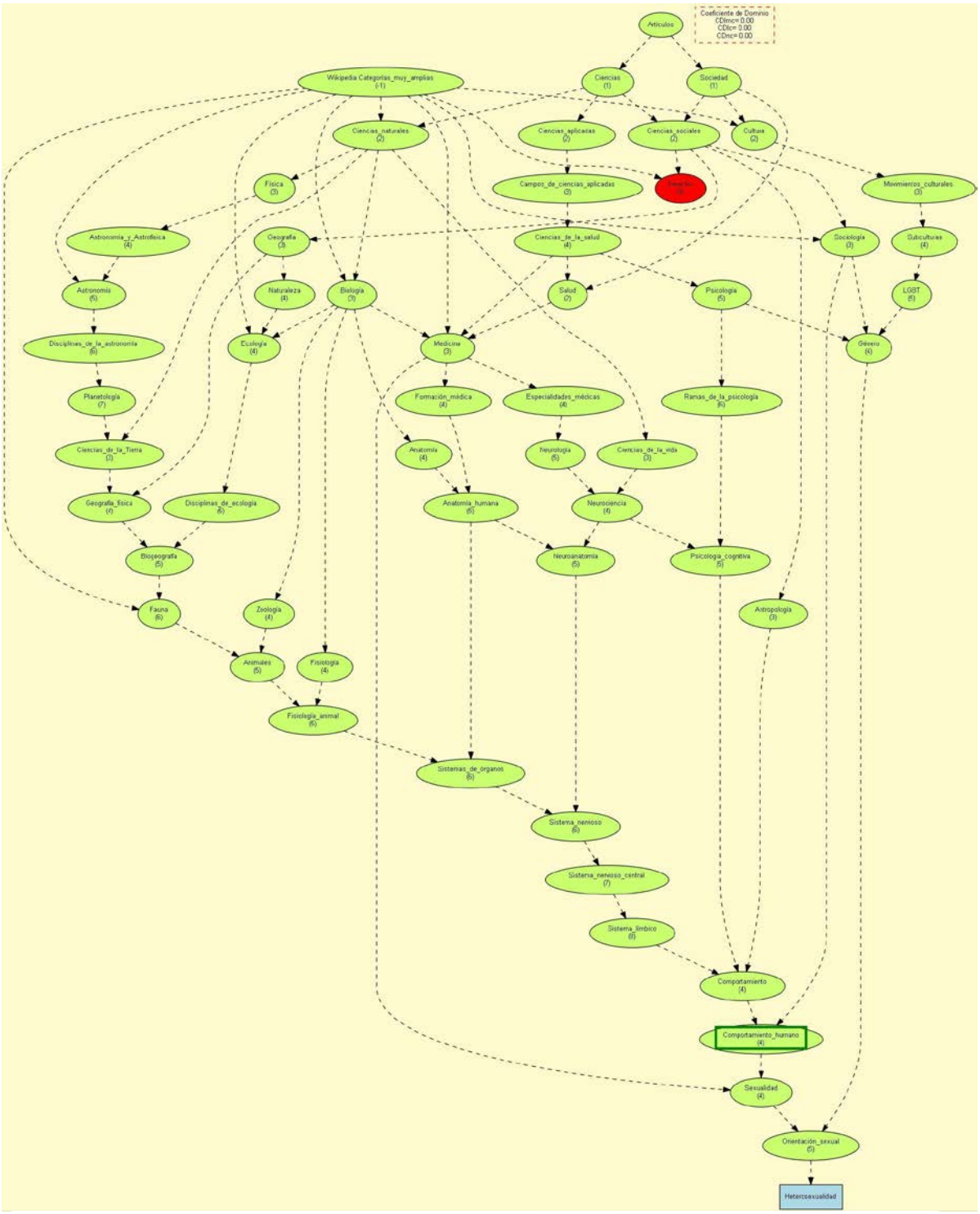


Figura 2. Resultados del candidato *heterosexual* si definimos el dominio derecho únicamente con la categoría *derecho*

En la Figura 2 se muestran los resultados de *heterosexual* dentro de la categoría *derecho* en modo PP. Como podemos ver en la parte superior, obtiene 0 puntos porque en la Wikipedia *heterosexual* no aparece vinculado con esta categoría. Si observamos el resto de categorías (en los caminos hacia el top) relacionadas con este término, encontramos *comportamiento humano* muy cerca del término. Parece lícito considerar que *comportamiento humano* es una

categoría vinculada con el derecho[5]. Así pues, podemos aplicar esta nueva categoría junto con la del derecho para comprobar si de esta manera algunos términos que no se detectaban o recibían una puntuación baja ahora sí se recogen como términos del derecho canónico. La mejora que se obtiene en este caso es pequeña (CWwp_n=0,1) pero suficiente para que este candidato no quede completamente descartado.

Si hacemos una prueba con otros candidatos como *familia*, por ejemplo, observamos que se encuentra relacionado con esta categoría también y por lo tanto sube su puntuación (en este

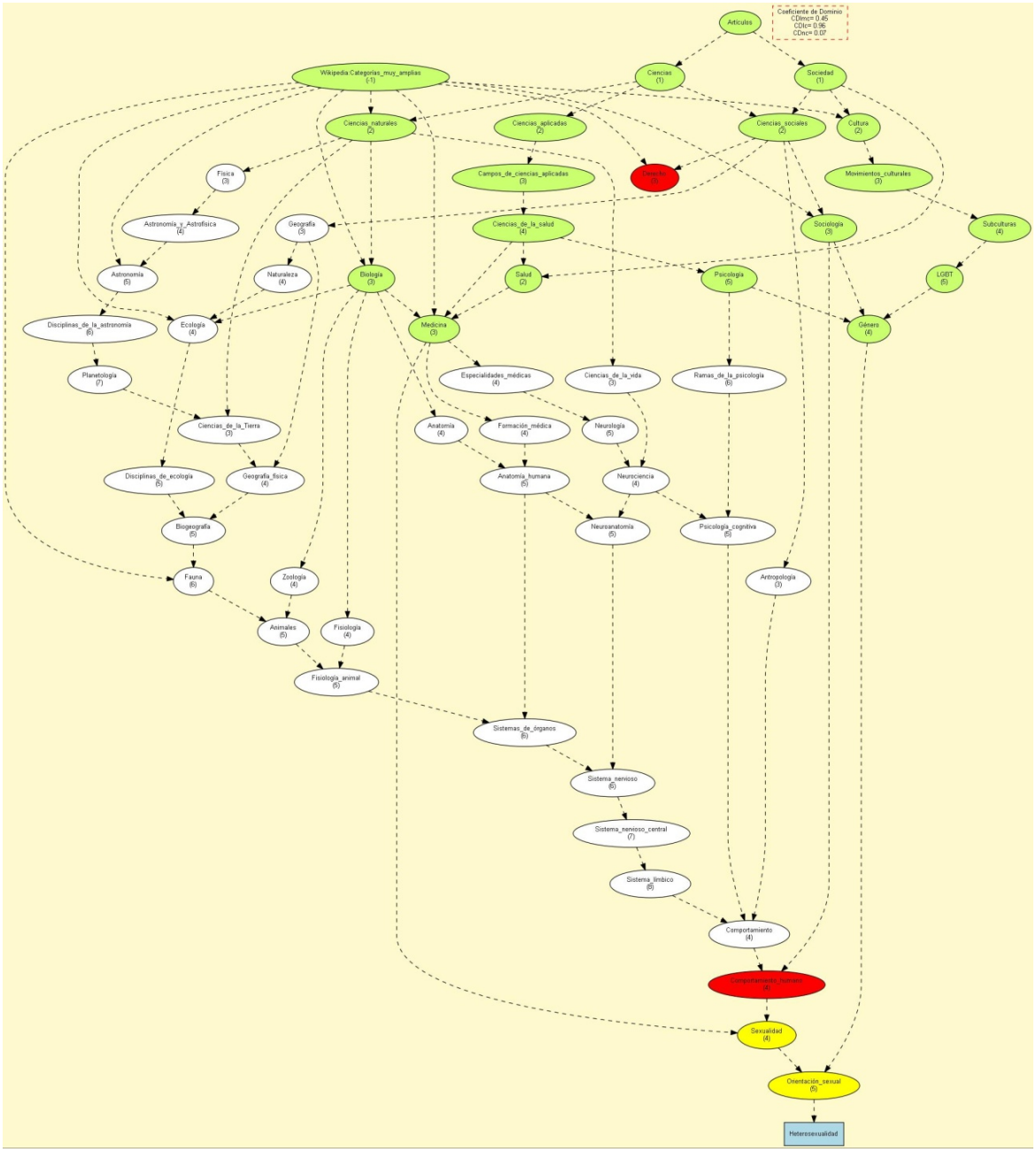


Figura 3. Resultados del candidato *heterosexual* si definimos el dominio derecho con las categorías *derecho* y *comportamiento humano*

[5] Hay que tener en cuenta que el hecho de añadir esta categoría a la definición de dominios hará que todas las subcategorías (y páginas vinculadas) recibirán a partir de este momento una puntuación superior y por lo tanto se alterará el fichero resultado.

caso el valor de CDwp_nc sube de 0.0009 a 0,1).

En la Figura 3, al añadir la categoría *comportamiento humano*, los colores cambian para indicar la nueva relación entre la categoría y el término, y ya obtiene puntuación.

En lo que respecta a este tipo de procesamiento, es necesario indicar el "modo" (de exploración del árbol de categorías de la Wikipedia). En el script superior, nos dará los resultados que buscamos utilizando el modo AP (exploración en amplitud prioritaria), pero también podemos detallar que nos muestre los resultados en modo PP (exploración en profundidad prioritaria). Como podemos observar en los grafos de la Figura 4, el modo PP ofrece más información porque no se limita a las subcategorías situadas entre el término y la categoría que hemos detallado, sino que aporta información jerárquica más elevada y nos indica la distancia entre ellos a partir de los números. Sin embargo, puede complicar la supervisión de las relaciones porque el grafo se complica. Por otro lado, el modo AP no profundiza en la jerarquía superior a la categoría definida, sino que aporta la información situada entre dicha categoría o categorías y el candidato a término. Así pues, la información es más concisa.

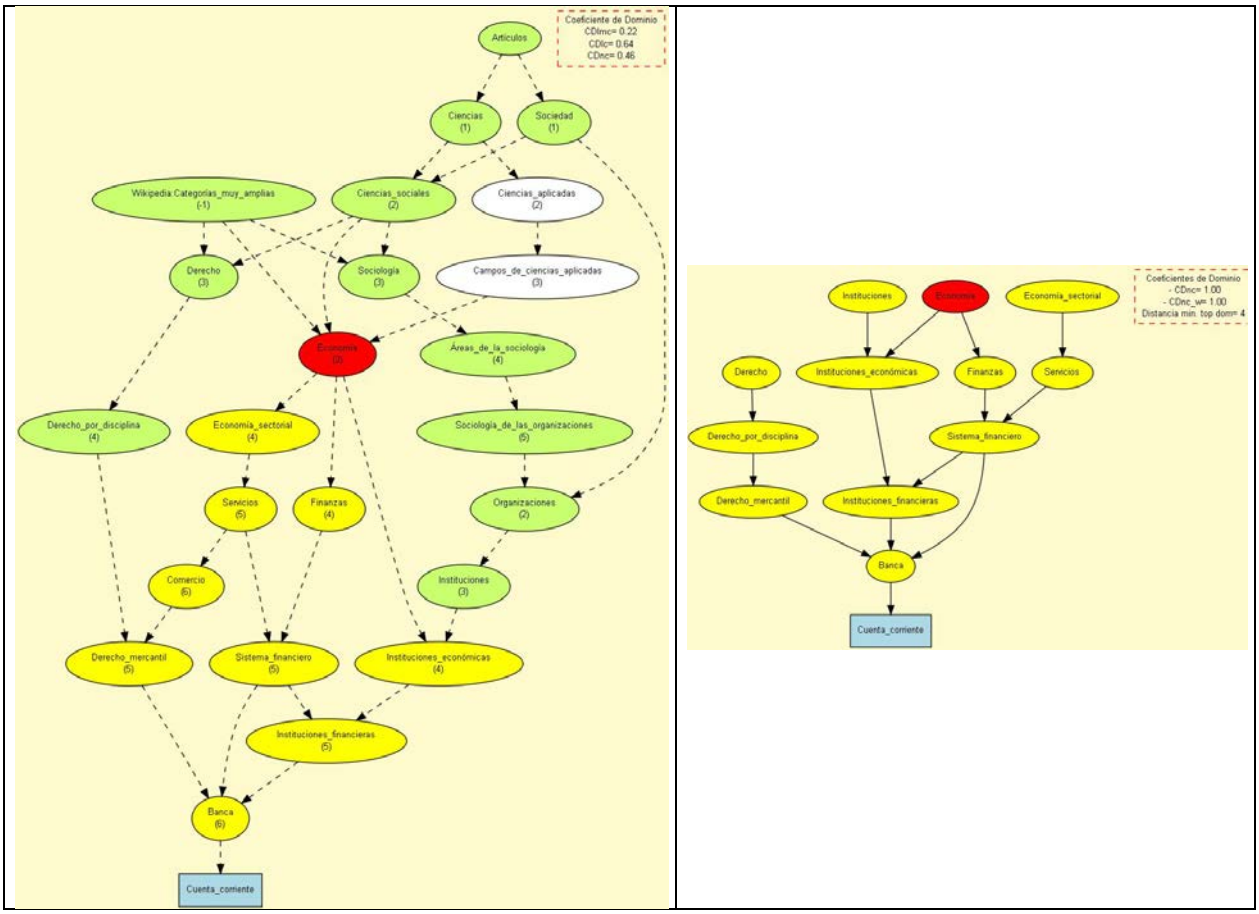


Figura 4. Grafos de resultados en modo PP y AP del término "cuenta corriente" en la categoría economía

a) Modo PP	b) Modo AP

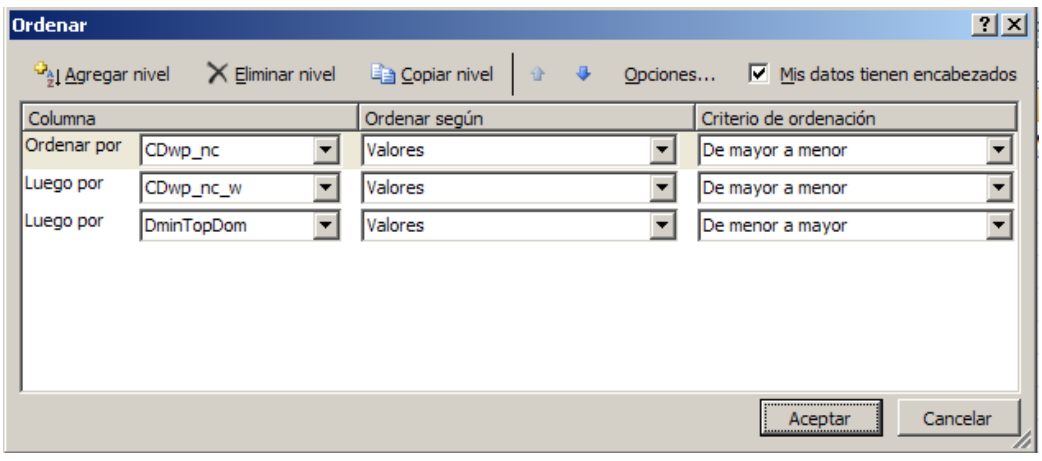
7.3 Interpretación de los resultados

El resultado que proporciona esta herramienta en una tabla donde cada CAT tiene asociada una serie de informaciones que se describen a continuación:

#	Parámetro	Significado
1	CAT	Candidato a término
2	Lema	Lema
3	Patrón	Patrón morfosintáctico
4	Control	Parámetro interno de control
5	ValidTerm	Indicación si es un término validado (permite calcular precisión y cobertura)
6	CDwp_lc	Coeficiente de dominio en base a la longitud de caminos
7	CDwp_lc_w	Coeficiente de dominio (ponderado) en base a la longitud de caminos
8	CDwp_nc	Coeficiente de dominio en base al número de caminos
9	CDwp_nc_w	Coeficiente de dominio (ponderado) al número de caminos
10	CDwp_lmc	Coeficiente de dominio en base a la longitud media de caminos al top del dominio
11	CDwp_lmc_w	Coeficiente de dominio (ponderado) en base a la longitud media de caminos al top del dominio
12	DminTopDom	Distancia mínima a un top del dominio
13	deteccion	Mecanismo de detección. Valores posibles: - `page`: existe como página de la Wikipedia - `category`: existe como categoría de la Wikipedia - `desamb`: información obtenida por desambiguación - `pageredir`: información obtenida a través de un redireccionamiento - `components`: información obtenida mediante análisis de sus componentes - `noCategs`: la página existe pero no tiene asignada categoría alguna (error en la BD) - `nil`: no encontrado en esta versión de la Wikipedia
14	CATalternativo	Candidato a término efectivo (en caso de redirección o desambiguación)
15	Categorías	Todas las categorías de la Wikipedia asociadas al CAT

Figura 5. Ordenación de los resultados

La manera óptima de interpretar los resultados es abrir el fichero de texto con una hoja de cálculo (ej. Excel o Open Office Calc) y ordenar los resultados siguiendo tres criterios (ver Figura 5). Seleccionar *Ordenar y filtrar/orden personalizado*: i) CDwp_nc (mayor → menor), ii) CDwp_nc_w (mayor → menor) y iii) DminTopDom (menor → mayor).



La visualización resultante de este ordenamiento nos debería facilitar la valoración de los candidatos resultantes. En práctica, lo que estamos haciendo es dividir la lista resultante en cuatro zonas tal como se indica a continuación:

zona	CAT		CDwp_nc	CDwp_lc_w	DminTopDom	deteccion	
1	…		0,99-0,01				
2	…		1				
3	…		0				
4	…		-1				

Aspectos que debemos observar en cada zona:

1) Se trata de candidatos que tienen una valoración entre 0 y 1 en la columna `CDwp_nc`. Se pueden distinguir varios subcasos:
   a. En la columna Deteccion se indica *page* o *categ*. Si el valor indicado en la columna CDwp_lc_w es alto combinado con un valor de `DminTopDom` bajo el candidato se puede considerar válido.
   b. En la columna Deteccion se indica *components*, lo cual indica que se trata de una unidad poliléxica y que el candidato que no está en la Wikipedia pero sí alguno o algunos de sus componentes. Podría tratarse de un paratérmino (ej. *caso de asma bronquial* o *signos de insuficiencia cardiaca*) en el cuál habría que centrarse en identificar el término (*asma bronquial y insuficiencia cardiaca*). Si este término ya aparece en la lista de términos podemos descartar esta línea.
2) Candidatos con la valoración máxima. Es importante buscar potenciales errores. Una indicación de un candidato potencialmente erróneo es cuando `DminTopDom` tiene un valor más alto de lo habitual (valores mayores que 5-6 son sospechosos porque no son del dominio o son generales). Otra causa de error son categorías de la Wikipedia demasiado permisivas. En este caso habría que examinar el candidato individualmente como se indica en De un candidato a término aislado7.2.2 y eventualmente mejorar la definición de dominio.
3) Se trata de candidatos que el sistema no valora como pertenecientes al dominio. Si consideramos que debería estar incluido entre los termino del dominio examinar individualmente como se indica en De un candidato a término aislado7.2.2. por si fuera necesario agregar/eliminar/cambiar alguna categoría de la Wikipedia a la definición de dominio.
4) Se trata de candidatos que no se pueden detectar con la Wikipedia ya sea porque no están incluidos o bien hay algún error interno de la base de datos (en este caso en la columna detección debería mostrar el mensaje `NoCategs`)

Una situación común a todas las zonas son los candidatos detectados por desambiguación (`desamb` en la columna `deteccion`). Se trata de palabras ambiguas y que el programa de evaluación intenta desambiguar por métodos endógenos. Este es un proceso delicado y sujeto a errores por lo tanto es necesario verificar estos candidatos. El mensaje `nil` en la columna `deteccion` es una indicación que hay una ambigüedad pero el sistema no ha sido capaz de resolverla.

En cualquier caso hay que tener en cuenta que si se ha detectado la necesidad de agregar/eliminar/cambiar categorías de la Wikipedia a la definición de dominio es imprescindible repetir el proceso de análisis de la lista de candidatos (ver 7.2.1 y 7.3).

7.3.1 Ejemplo de revisión de resultados

Para guiar la interpretación y revisión de los resultados, a continuación comentamos el listado de términos obtenido a partir del análisis de un texto del ámbito del derecho canónico[6].

Zona 1: 0,1-0,99

En este caso, todos los términos de esta zona tienen la etiqueta *components* porque se trata de una unidades poliléxicas que no están en la Wikipedia pero sí que está registrado alguno de sus componentes. En este punto encontramos dos posibles errores (*a* y *b*) y en *c)* realizamos una observación sobre los candidatos con puntuación:

a) *Candidatos a término con carácter general*

Contiene nacionalidades, nombre de poblaciones y países: *normativa legal vigente italiana*

Empiezan con la palabra ***carácter*** (*carácter multicultural*, *carácter previo*, *carácter tradicional*), ***característica*** (*característica morfológica*), ***caso*** (*caso límite, caso concreto, caso de pareja*), ***criterio*** (*criterio equivalente*, *criterio de interpretación*), ***interés*** (*interés bastardo, interés político, interés superior*), ***modo*** (*modo de ejemplo*, *modo de* introducción) ***serie*** (*serie de problema*, *serie de vínculo*), etc.

b) *Candidatos a término considerados paratérminos*

Empiezan con la palabra ***carácter*** (*carácter heterosexual*), ***caso*** (*caso de matrimonio homosexual, caso de adopción*), ***condición*** (*condición de acogida*, *condición de ejercicio*), ***cuestión*** (*cuestión de capacidad*, *cuestión de constitucionalidad*), ***forma*** *(forma de matrimonio, forma de oposición*), ***materia*** (*materia matrimonial, materia de derecho*), ***motivo*** (*motivo de inconstitucionalidad*), ***objeto*** (*objeto de exequátur*), ***procedimiento*** (*procedimiento de adopción*), ***proceso*** (*proceso de reconocimiento de maternidad*), ***regulación*** (*regulación jurídica*), ***supuesto*** (*supuesto de familia*, *supuesto de pareja estable*), ***tipo*** (*tipo de discriminación*, *tipo de gestación, tipo de ley*).

c) *Candidatos con puntuación baja* (0,2-0,3)

Entre los candidatos con puntuación muy baja también hay términos correctos: *adopción conjunta*, *conviviente homosexual*, *enunciado legal, fundamento jurídico, carácter registral, filiación materna, madre gestante, órgano consultivo,* etc.

[6] Este es un ejemplo concreto y no significa que todos los listados que se obtengan tengan exactamente las mismas características en cuanto a distribución de coeficientes de dominio.

Si procesamos uno de estos candidatos como se indica en 7.2.1 o bien lo buscamos directamente en la Wikipedia, vamos a ver que, por un lado, el término no se encuentra registrado como entrada poliléxica y, por otro lado, uno o varios de los elementos que lo componen tampoco se encuentran como entradas bien situadas dentro de la categoría del derecho y, por este motivo, reciben una puntuación baja.

Zona 2: 1 punto

a) Candidatos con puntuación baja en* DminTopDom *que no son términos del derecho

Aunque obtienen una puntuación baja en *DminTopDom* no son términos del derecho. A veces, este error ocurre porque el candidato se utiliza como parte del lenguaje general en el texto (*en este sentido, entre otras cosas, intereses políticos y propagandísticos,* etc.), pero coincide con un término del derecho (*sentido, cosa* e *interés,* respectivamente).

- *dato* (1) [Comunicación,Datos_informáticos,Estadística,Programación…]
- *fin* (1) [Comunicación,Datos_informáticos,Estadística,Programación…]
- *interés* (3) [Banca,Contabilidad,Derecho_mercantil,Interés]
- *lectura* (1) [Comunicación,Datos_informáticos,Estadística,Programación…]
- *referencia* (4) [Filosofía_del_lenguaje,Referencias,Terminología_filosófica]
- *génesis* (4) [Génesis,Libros_adaptados_a_la_televisión…]
- *cosa* (1) [Derecho_de_cosas,Términos_jurídicos]
- *cuestión* (1) [Comunicación,Gramática]
- *extensión* (1) [Desarrollo_personal,Educación_en_Estados_Unidos]
- *sentido* (1) [Discriminación,Peyorativos] (desamb)

b) Candidatos con puntuación alta en* DminTopDom *que son correctos

Aunque tienen una puntuación alta y, por tanto, es posible que no sean términos del derecho, algunos son correctos. La puntuación es alta porque tienen una categoría asignada en la Wikipedia que no se corresponde con el dominio en cuestión. Por ejemplo, en el caso de *recurrente* (persona que interpone un recurso), obtiene una puntuación alta porque si lo buscamos en la Wikipedia está registrado con otro significado.

- *domicilio conyugal* (7) [Películas_de_1970,Películas_de_Francia…]
- *recurrente* (7) [Gramática_generativa,Programación,Terminología]

c) Candidatos con puntuación alta en* DminTopDom *que no son términos

En general, a partir de 5-6 puntos en el campo DminTopDom podemos deducir que no son términos del ámbito o bien son muy generales y, por tanto, se deben eliminar.

- *sensación* (7) [Neurociencia,Percepción,Sistema_nervioso]
- *especie humana* (5) [Antropología,Fósiles_del_Pleistoceno,Homo…]
- *época* (5) [Nacionalismo_catalán,Organizaciones_terroristas]

- *merma* (9) [Contabilidad,Economía_de_la_producción]

Zona 3: 0 puntos

Los siguientes candidatos a término reciben 0 puntos, pero se consideran términos de acuerdo con el ámbito del texto (derecho canónico):

- *alcalde* [Administración_local,Ciencia_política]
- *convivencia* [Psicología]
- *institución* [Instituciones,Sociología_de_la_cultura]
- *mujer* [Género,Mujer,Sociología_de_la_cultura]
- *nacimiento* [Género,Mujer,Sociología_de_la_cultura]
- *precepto* [Judaísmo]
- *premisa* [Lógica]
- *proyecto* [Comunidad,Desarrollo_social,Proyectos]
- *religión* [Religión]
- *vulneración* [Desastres]
- *libertad religiosa* [Libertad_de_culto]
- *sexo femenino* [Mujer,Reproducción,Términos_botánicos]
- *libertad de consciencia* [Libertad_de_culto]
- *constancia registral*
- *progenitor biológico*
- *proyecto legislativo*

En el caso de *alcalde*, por ejemplo, si lo procesamos como candidato aislado observamos que en la Wikipedia no se encuentra dentro del árbol del derecho, sino de la economía. Así pues, aunque se considerarían términos del ámbito del derecho no los clasifica como tales porque en la Wikipedia tienen asignada otra categoría.

Zona 4: -1 punto

Los siguientes candidatos a término reciben -1 punto porque no existe una entrada en la Wikipedia o hay un error de desambiguación. Aun así, podemos observar que algunos deberían considerarse términos del ámbito aunque no estén registrados en la Wikipedia.

- *aborto* NoCategs
- *adopción* NoCategs
- *adoptante* NoCategs
- *antecedente* desamb - nil
- *contratante* nil
- *contrayente* nil
- *conveniencia* NoCategs
- *in_fine* nil
- *inconstitucionalidad* nil
- *ius_canonicum* nil
- *obligatoriedad* nil
- *ope_legis* nil

- *precedente* NoCategs
- *primo_comma* nil
- *progenitor* NoCategs
- *sometimiento* NoCategs
- *subterfugio* nil
- *garantía constitucional* NoCategs
- *laguna legal* NoCategs
- *ordenamiento constitucional* NoCategs
- *sexo masculino* NoCategs
- *gestación de sustitución* NoCategs

8 Integración de los resultados en una dase de datos para su explotación

Los resultados obtenidos con los documentos procesados se pueden integrar en una base de datos mysql para posteriormente hacer una explotación de dichos resultados.

8.1 Estructura interna de la Base de datos de resultados

La BD está físicamente alojada en `vivaldi.upf.edu`. Los datos necesarios para su creación están en el fichero `wikiyateresultsdb.mysql` y está en el directorio `/home/vivaldi/Perl/wikiYATE`. El script de creación es el siguiente:

¡¡ATENCIÓN!! Ejecutar este script implica borrar TODOS los datos

```
vivaldi@vivaldi:~/Perl/wikiYATE$ mysql -u root -p <
wikiyateresultsdb.mysql
```

Para generar las tablas con el MySql Workbench el procedimiento a seguir desde el menú principal es el siguiente:

1) File → Export → Forward Engineer SQL create script
2) Next
3) Next
4) Copy to Clipboard
5) Pegar en un editor de texto y utilizar como script de generación de las tablas

En el Anexo I se muestra la estructura interna de esta BD.

8.2 Procedimiento de integración

Esta etapa consiste en integrar el resultado del análisis de un documento del corpus (sección 7) a la base de datos de resultados (BDR) de Wikipedia para su explotación. Esta integración se realiza con el programa `guardaResulWP2DB.pl` y está en `P:\IULA\CORPUS\UTILS\Wikipedia`.

Este programa dispone de una serie de opciones que pueden obtenerse invocándolo con la opción –h. El resultado que se debería obtener es el indicado en el Listado 2. Esta manera de invocar el programa también proporciona ejemplos básicos de uso.

Las opciones más relevantes son las que se indican a continuación:

- Selección del documento `-doc`: documento procesado por YATE
- acción a realizar sobre el documento indicado en `-doc`. `–accion`: (insertar/borrar/explorar los CAT de un documento, comprobar el fichero índice)
- Idioma del documento: `-lang`
- Dominio del CT al que pertenece el documento.

```
C:\Users\U1480\Dropbox\wikipedia\WikiCD\ExperimentoIJCAI> perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -help
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl
Programa para guardar el resultado de explorar un fichero con la Wikipedia
en la base de datos "wikiYATEresultsDB" (vivaldi.upf.edu)
  -accion acción        ## acción a realizar: insertar | explorar | docsBD | test5DX
  -doc docCT            ## documento a procesar: ficheros *.dat de YATE (N, NJ, NPN y PATS)
  -lang idioma          ## idioma de los documentos (es*|ca|en)
  -indir dir            ## directorio donde ir a buscar los ficheros con los CATs a explorar (sólo con la
   opción -doc)
  -user usuario         ## nombre de usuario
  -dominio dominio      ## dominio del CT
  -cdwpnc número        ## umbral de CDwp_nc [-1*, 0-1]
  -cdwpncw número       ## umbral de CDwp_nc_w [-1*, 0-1]
  -dminTD número        ## distancia mínima al top del dominio [0-10]
  -validacion val       ## método de validación (page|category|catSplit|pageredir|pagedesamb|nil)
  -help                 ## Ayuda

Ejemplos de uso:
- Insertar todos los candidatos a término a término de un documento de medicina en castellano
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi -accion  insertar -indir
   C:\Users\U1480\Documents\__Proyectos\_2012_Aple2\Proves -doc informeMedicinaInternaD003
   -lang es -dominio medicina
- Mostrar todos los candidatos a término de un documento:
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -accion  explorar -user vivaldi -indir './' -
   doc informeMedicinaInternaD003
- Mostrar todos los candidatos a término de un documento con el patron NJ:
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -accion  explorar -doc
   informeMedicinaInternaD003 -user vivaldi -indir './' -pattern NJ
- Mostrar todas las ocurrencias de un candidatos a término (dolor)
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -accion  explorar -user vivaldi -indir './' -
   cat dolor
- Mostrar todos los candidatos a término de un documento que cumplan ciertas condiciones (
   CDwp_nc_w > 0.5 y DminTopDomain < 4)
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -accion  explorar -doc
   informeMedicinaInternaD003 -user vivaldi -indir './' -cdwpncw 0.5 -dminTD 4
- Borrar un documento de la BD
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi -accion borrar -doc
   Llibre1_ES -lang es
- Comprobar el fichero índice de un documento antes de ingresarlo a la BD
  P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi -accion test5DX -doc
   Llibre3_ES -lang es -inDir P:\IULA\CORPUS\CORPUS\TMPDRET
```

Listado 3. Opciones del programa `guardaResultWP2DB.pl`

A continuación detallamos la información básica y necesaria para realizar algunas de las operaciones posibles con este programa.

8.2.1 Insertar los CATs de un documento

Para insertar todos los candidatos a término de un documento en la BDR el mandato a ejecutar es el siguiente:

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi -lang es -accion
insertar -dominio derecho -indir P:\IULA\CORPUS\CORPUS\TMPDRET -doc d00360
```

Listado 4. Inserción de los candidatos a término de un documento en la BDR

El sistema devolverá un mensaje como el que se indica a continuación:

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi
 -lang es -accion insertar -dominio derecho -indir
P:\IULA\CORPUS\CORPUS\TMPDRET -doc d00360
Conexión ok! (vivaldi.upf.edu:wikiYATEresultsDB).
Se han recuperado un total de 385 (385) CATs del patrón N
Se han recuperado un total de 215 (215) CATs del patrón NJ
Atención! P:\IULA\CORPUS\CORPUS\TMPDRET/d00360/d00360_JN.dat no existe
Se han recuperado un total de 69 (69) CATs del patrón NPN
Se han recuperado un total de 12 (12) CATs del patrón NJJ
Se han recuperado un total de 4 (4) CATs del patrón NJPN
Se han recuperado un total de 9 (9) CATs del patrón NPNJ
Se han recuperado un total de 1 (1) CATs del patrón NJPNJ
Se han recuperado un total de 1 (1) CATs del patrón NPNPN
Se han recuperado un total de 1 (1) CATs del patrón NPNPNJ

Conexión ok! (vivaldi.upf.edu:wikiYATEresultsDB).
Patron: NJPNJ
Atención no hay definidas categorÍas de WP para "pena privativo de
libertad superior".
Patron: NJPN
Atención no hay definidas categorÍas de WP para "responsabilidad objetivo
de pago".
Atención no hay definidas categorÍas de WP para "identidad sustancial de
cliente".
Atención no hay definidas categorÍas de WP para "pena privativo de
libertad".
Atención no hay definidas categorÍas de WP para "delito correspondiente
de conformidad".
Patron: N
Categorías de "dominio": 1.
Atención no hay definidas categorÍas de WP para "incumplimiento".
Categorías de "consumidor": 5.
Categorías de "laura": 2.
Categorías de "peligrosidad": 1.
Categorías de "corrupción": 4.
Atención no hay definidas categorÍas de WP para "inicio".
Categorías de "transformación": 0.
Categorías de "ley": 2.
Categorías de "uso": 0.
Atención no hay definidas categorÍas de WP para
"societas_delinquere_potest".
Categorías de "accionista": 2.
Categorías de "defraudación": 0.
Categorías de "oportunidad": 4.
Categorías de "regulación": 3.
…
```

```
…
Patron: NPNPN
Atención no hay definidas categorÍas de WP para "delito de
quebrantamiento de condena".
Patron: NPNJ
Atención no hay definidas categorÍas de WP para "establecimiento de pena
relativo".
Atención no hay definidas categorÍas de WP para "exigencia de
responsabilidad penal".
Atención no hay definidas categorÍas de WP para "cumplimiento de forma
solidario".
Atención no hay definidas categorÍas de WP para "punto de vista
procesal".
Atención no hay definidas categorÍas de WP para "venta de producto
alimenticio".
Atención no hay definidas categorÍas de WP para "ejercicio de actividad
social".
Atención no hay definidas categorÍas de WP para "pago de manera directo".
Atención no hay definidas categorÍas de WP para "medio de prueba
pertinente".
Atención no hay definidas categorÍas de WP para "establecimiento de
medida eficaz".

Inserción completada de "d00360".

P:\IULA\CORPUS\CORPUS\TMPDRET>
```

¡¡ATENCIÓN!!

Si el fichero de indexación del documento que estamos procesando existe pero corresponde a una versión anterior del documento el programa dará errores. Si pasa esto, dejar terminar el proceso de inserción y a continuación borrar el documento (sección 8.2.2) antes de volver a insertarlo (sección 8.2.1) en la BDR.

Si el documento ya existe en la BDR el sistema devolverá un mensaje como el que se indica a continuación:

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi
 -lang es -accion insertar -dominio derecho -indir
P:\IULA\CORPUS\CORPUS\TMPDRET -doc d00360
Conexi¾n ok! (vivaldi.upf.edu:wikiYATEresultsDB).
Se han recuperado un total de 385 (385) CATs del patr¾n N
Se han recuperado un total de 215 (215) CATs del patr¾n NJ
Atención! P:\IULA\CORPUS\CORPUS\TMPDRET/d00360/d00360_JN.dat no existe
Se han recuperado un total de 69 (69) CATs del patr¾n NPN
Se han recuperado un total de 12 (12) CATs del patr¾n NJJ
Se han recuperado un total de 4 (4) CATs del patr¾n NJPN
Se han recuperado un total de 9 (9) CATs del patr¾n NPNJ
Se han recuperado un total de 1 (1) CATs del patr¾n NJPNJ
Se han recuperado un total de 1 (1) CATs del patr¾n NPNPN
Se han recuperado un total de 1 (1) CATs del patr¾n NPNPNJ

Conexión ok! (vivaldi.upf.edu:wikiYATEresultsDB).
Atención! el documento indicado (d00360) ya existe en la BD de wikiYATE.

P:\IULA\CORPUS\CORPUS\TMPDRET>
```

Esto significa que para actualizar el contenido de la BDR con los datos de un documento, primero hay que borrarlo y a continuación insertarlo nuevamente en la BDR.

8.2.2 Borrar los CATs de un documento

Para insertar todos los candidatos a término de un documento en la BDR el mandato a ejecutar es el siguiente:

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi -lang es -accion
borrar -doc d00360
```

Listado 5. Borrado de los candidatos a término de un documento en la BDR

El sistema devolverá un mensaje como el que se indica a continuación:

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi
 -lang es -accion borrar -doc d00360


Conexión ok! (vivaldi.upf.edu:wikiYATEresultsDB).
Recoge información sobre los registros a borrar...
        términos...  encontrados 697 términos.
        contextos
                tabla intermedia... 1767
                datos... 1767
        componentes
                tabla intermedia... 546
                datos... 321
        categorías WP de los términos (tabla intermedia) ... 875
        categorías WP del resultado (tabla intermedia) ... 1
        resultado ... 1
Borrado efectivo...
Borra 1767 ids de contextData...
Borra 1767 ids de Contexts...
Borra 546 ids de TermCandidates_has_TCcomponents...
Borra 321 ids de TCcomponents...
Borra 875 ids de TermCandidates_has_WPcategories...
Borra 697 ids de TermCandidates...
Borra 1 ids de Results_has_WPcategories...
Borra 1 ids de Results...
P:\IULA\CORPUS\CORPUS\TMPDRET>
```

8.2.3 Comprobar el fichero índice de un documento

Para comprobar que el resultado de la indexación (resultado de la ejecución del mandato `indexSGMLplus.pl`, ver sección 5) existe y permite recuperar todos los elementos textuales de un documento a integrar en la BDR el mandato a ejecutar es el siguiente:

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi -lang es -accion
test5DX --dominio derecho -indir P:\IULA\CORPUS\CORPUS\TMPDRET -doc d00360
```

Listado 6. Comprobación del fichero índice de un documento a integrar en la BDR

Este mandato es necesario si existen dudas sobre la existencia del fichero resultado de la indexación (*.5dx). Si se integra un documento actualizado en la BDR pero la indexación no se ha realizado (o corresponde a una versión anterior del documento) el mandato “insertar” de `guardaResultWP2DB.pl` dará un error (ver sección 8.2.1).

El sistema devolverá un mensaje como el que se indica a continuación

```
P:\IULA\CORPUS\CORPUS\TMPDRET>perl
P:\IULA\CORPUS\UTILS\Wikipedia\guardaResultWP2DB.pl -user vivaldi  -
lang es -accion test5DX -dominio derecho -indir
P:\IULA\CORPUS\CORPUS\TMPDRET -doc d00360

Conexi¾n ok! (vivaldi.upf.edu:wikiYATEresultsDB).
Procesalmente se exige que se adopten con audiencia previa de los
titulares de las empresas o instituciones o de sus representantes
legales .
12 .
3 .
1
Ámbito de aplicación
Entre estos últimos autores cabe citar a Mir_Puig , que considera que
las consecuencias accesorias son medidas de seguridad peculiares "
sui_generis " , con rasgos propios de las penas .
La responsabilidad penal de las personas jurídicas no excluye la de
las personas físicas autores o cómplices de los mismos hechos " ( art.
121 ) .
5 .
La doctrina y la jurisprudencia españolas han acogido tradicionalmente
de forma mayoritaria este principio.
Por_lo_tanto estas consecuencias se hacen efectivas desde la firmeza
de la sentencia , a·n cuando no se hubiera iniciado la ejecuci¾n de la
pena impuesta , para poder lograr los fines preventivos que las
sustentan .
7 .
Extinción
Muy discutida entre la doctrina es la cuestión de la naturaleza
jurídica de las consecuencias accesorias recogidas en el art. 129 CP.
Las posturas las podemos agrupar en dos grandes grupos .
…
…
el acceso sin autorización a sistemas informàticos ;
1 .
En primer lugar cabe que nos preguntemos acerca_de la naturaleza
jurídica de la multa que se impone a la persona jurídica .
327- Delitos contra el medio ambiente . Art .

--> Comprobar que el texto que acaba de imprimirse en pantalla sea
coherente y sin errores.

P:\IULA\CORPUS\CORPUS\TMPDRET>
```

Es importante hacer la comprobación que se indica en la última línea del mensaje. Normalmente cuando se intenta integrar a la BDR un resultado con una versión antigua del fichero de indexación el texto resultante es inconexo.

8.2.4 Exploración desde línea de comandos del contenido de la Base de datos de resultados

Este programa dispone de varias opciones que permiten seleccionar los datos de algunos de los CATs de la BDR en función de sus características. Estas opciones no se explican en este documento ya que se pueden realizar más cómodamente desde la interficie web (ver Sección 8.3).

8.3 Exploración desde la interficie web del contenido de la Base de datos de resultados

La comprobación de los CAT de uno o más documento que se hayan procesado con wikiYATE y se hayan incorporado a la BDR es a través de la página web http://vivaldi.upf.edu/WikiYATE_0/wikiYateBase.html.

Es muy importante tener en cuenta que esta página sólo es accesible desde dentro de la UPF. La pantalla inicial tiene el aspecto que se muestra en la Figura 6.

Los pasos a seguir para realizar una consulta básica son los siguientes:

1) Selección de idioma y dominio.
2) Selección entre los documentos disponibles aquellos objeto de la consulta y hacer clic sobre el botón Busca!
3) Selección del modo de ordenación de los CATs
4) Selección del patrón morfosintáctico de los CATs
5) hacer clic sobre el CAT buscado (Atención! los CATs se muestran lematizados siempre que es posible)
6) Examinar la información disponible (datos del CAT y contextos de aparición en los documentos seleccionados)
7) Eventualmente, es posible actualizar la información del CAT seleccionado: validez en el dominio y de los contextos válidos)[7]. Para ello sólo es necesario marcar/desmarcar las casillas correspondientes y activar el botón “Actualiza!”.

[7] Por defecto: i) sólo se consideran CATs válidos aquellos que tienen un CDwp_nc = 1, ii) ningún contexto es considerado válido

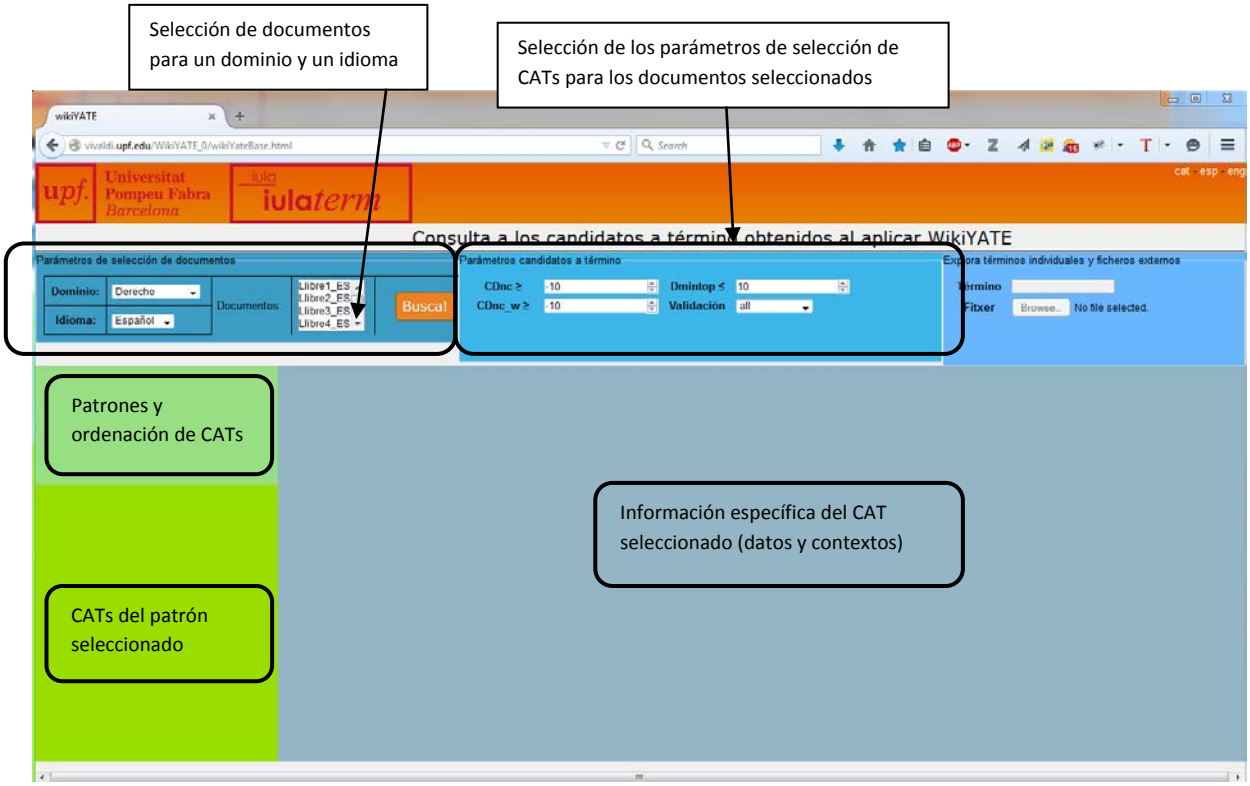


Figura 6. Pantalla principal de la interficie web al contenido de la BDR: áreas relevantes

Anexo I: Estructura de la Base de Datos de resultados

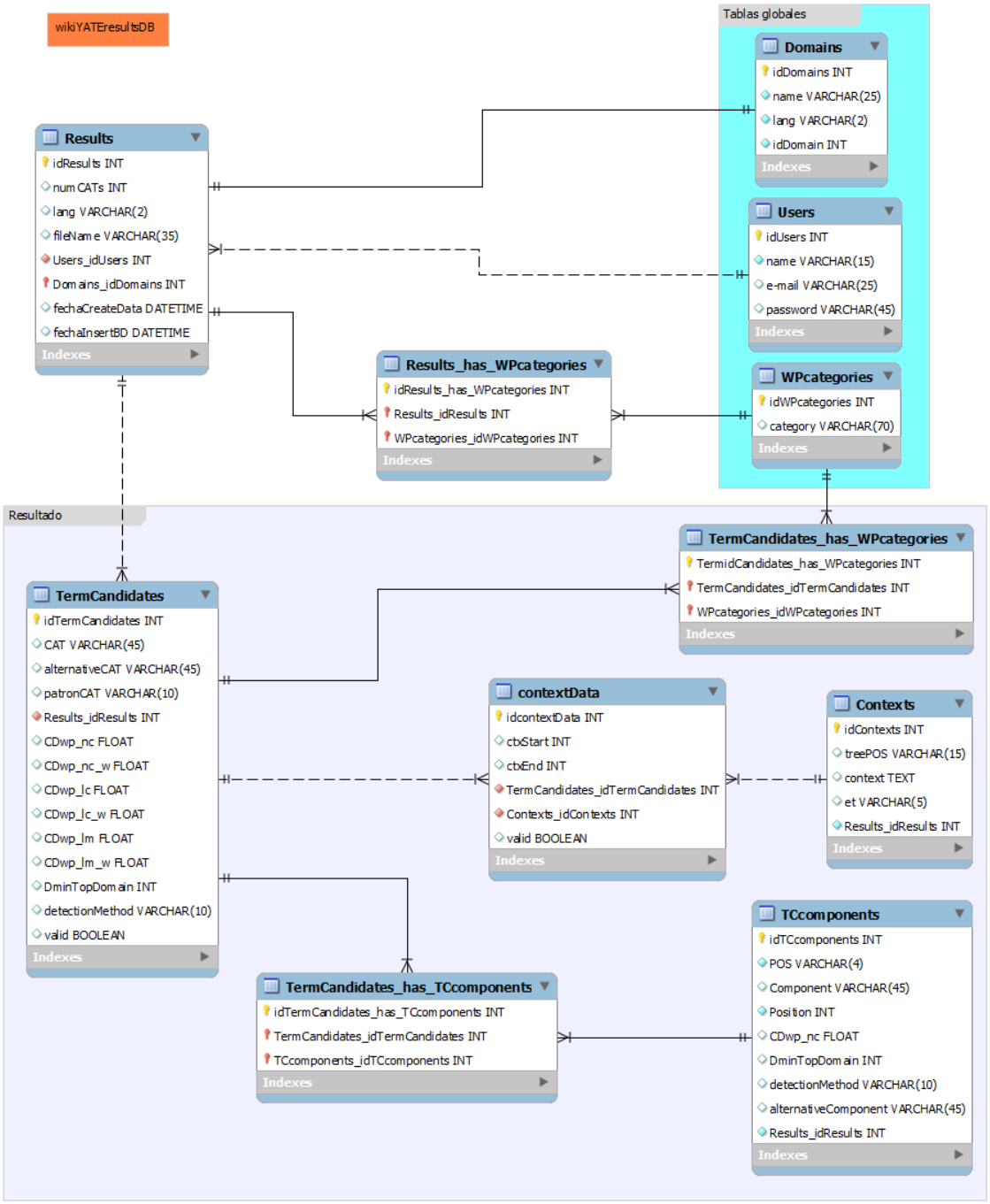


Figura 7. Estructura de la BD con los resultados.