Wikipedia2TBX: creación de glosarios terminológicos a partir de la Wikipedia

1. Introducción

La Wikipedia (http:wikipedia.org) es una enciclopedia libre que se ha creado de forma colaborativa. Esta enciclopedia es, además, multilingüe, es decir, hay versiones de la Wikipedia para un gran número de lenguas. Para cada entrada de la Wikipedia disponemos de enlaces interlingüísticos, que nos llevan directamente al mismo artículo en otra lengua.Los artículos de la Wikipedia, además, están organizados por categorías. Esto hace que se puedan construir glosarios terminológicos a partir de las entradas de la Wikipedia.

El programa que presentamos en este documento es capaz de crear glosarios terminológicos multilingües a partir de la Wikipedia. Para crearlos, el programa necesita saber de qué área o áreas temáticas y cuáles son las lenguas implicadas. Con estos datos mira todas las entradas de la Wikipedia inglesa que tengan asignadas las áreas temáticas dadas y mira si tiene enlaces interlingüísticos para las lenguas requeridas. A partir de los datos que va recogiendo confecciona un glosario terminológico multilingüe y crea un fichero en formato TBX (*Term Base Exchange*). También puede crear un glosario en formato tabulado para OmegaT. Estos glosarios son bilingües y por lo que sólo se tienen en cuenta las dos primeras lenguas dadas en el programa, la primera como lengua de partida, y la segunda como lengua de llegada. El resto de lenguas, para la creación del glosario en formato tabulado, las ignora.

En este documento explicaremos cómo se utiliza el programa Wikipedia2TBX. Su funcionamiento es muy sencillo, ya que dispone de una interfaz gráfica de usuario. También veremos qué códigos de lengua se utilizarán, así como explicaremos como conocer las clasificaciones temáticas de la Wikipedia.

Antes de utilizar la herramienta, hay que tener bien presente los siguientes aspectos:

- Los glosarios terminológicos se crean a partir de la información de la Wikipedia. Recuerda que la Wikipedia es un proyecto colaborativo, donde cualquier usuario puede añadir información. La riqueza del proyecto consiste precisamente en este aspecto colaborativo. La calidad del recurso es muy buena, pero todos los resultados obtenidos se deben entender como propuesta y por lo tanto queda a criterio del traductor utilizar o no una propuesta concreta en sus traducciones.
- El programa funciona contra una copia de la Wikipedia inglesa alojada en un servidor de la Universitat Oberta de Catalunya. Esta copia se actualiza regularmente con las últimas versiones, pero el contenido puede diferir ligeramente con la Wikipedia real.

2. Obtención, instalación y ejecución del programa

El programa es libre (bajo licencia GNU-GPL) y se puede descargar de http://lpg.uoc.edu/Wikipedia2TBX. El programa está escrito en Python y es por tanto multiplataforma (funciona sin problemas bajo Windows, Linux y Mac). Se distribuyen 2 versiones del programa:

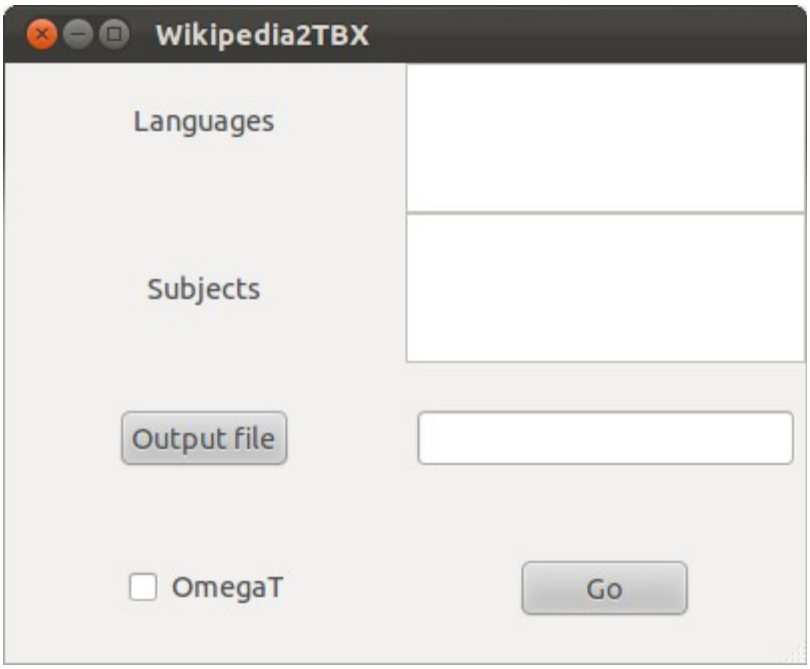
- Archivos fuente (Wikipedia2TBX-0.1.tar.gz): para ejecutarlo es necesario tener instalado un intérprete de Python en el sistema. Este hecho es el habitual en Linux y Mac y por lo tanto se podrá utilizar esta versión directamente bajo estos sistemas operativos si se dispone además de los prerequisitos. Si se quiere utilizar esta versión bajo Windows es necesario instalar un intérprete de Python (versión inferior a la 3.0) que se puede descargar libremente de www.python.org
- Versión ejecutable para Windows (Wikipedia2TBX-0.1-win.zip). Esta versión se puede ejecutar bajo Windows sin necesidad de tener el intérprete de Python instalado en el sistema.

No es necesario instalar el programa, simplemente hay que descomprimir el archivo descargado y hacer doble clic sobre el archivo Wikipedia2TBX.py o Wikipedia2TBX.exe. En todos los sistemas operativos, si se dispone del intérprete de Pythojnm se puede ejecutar desde lina de comando haciendo:

```
python Wikipedia2TBX.py
```

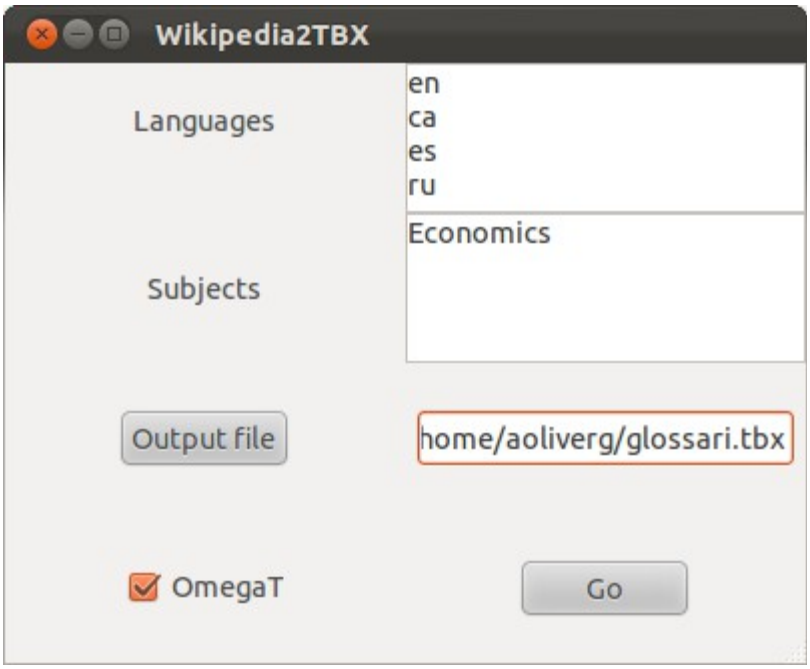
3. Funcionamiento del programa

Una vez hacemos doble clic sobre Wikipedia2TBX.py o Wikipedia2TBX.exe se abre una interfaz gráfica muy sencilla.



En esta interfaz tenemos que hacer:

- En el cuadro de texto **Languages** tenemos que poner los códigos de lengua (ISO de 2 letras, ver el apartado 4 de este documento). Como mínimo tenemos que poner dos códigos de lengua. Hay que poner un código por línea
- En el cuadro de texto **Subjects** tenemos que poner las áreas de especialidad que queremos (deben expresarse en inglés, ver el apartado 5 de este documento). Hay que poner un área de especialidad por línea.
- Hay que seleccionar la ubicación y el nombre del archivo de salida con el botón **OutputFile**.No es necesario dar la extensión, ya que automáticamente se pondrá extensión .tbx.
- El cuadro de selección **OmegaT** hay marcarlo si queremos obtener la salida también en formato tabulado para OmegaT. En este caso se creará un archivo en la misma ubicación y con el mismo nombre que el archivo de salida, pero en este caso con extensión .utf8
- Una vez hechas todas las selecciones hay que hacer clic en el botón **Go.** En este momento se empieza a crear el glosario. El proceso puede durar un buen rato y es imprescindible estar conectado a Internet.



A continuación podemos observar un fragmento del archivo de salida en TBX:

```
<?xml version="1.0" ?>
<martif type="TBX" xml:lang="en">
  <text>
    <body>
      <termEntry id="1">
        <descrip type="subjectField">
          Economics
        </descrip>
           <langSet xml:lang="en">
             <tig>
               <term>
                 Game theory
               </term>
             </tig>
           </langSet>
           <langSet xml:lang="ca">
             <tig>
               <term>
                 Teoria dels jocs
               </term>
             </tig>
           </langSet>
           <langSet xml:lang="es">
             <tig>
               <term>
                 Teoría de juegos
               </term>
```

```
            </tig>
          </langSet>
          <langSet xml:lang="ru">
            <tig>
              <term>
                Теория игр
              </term>
            </tig>
          </langSet>
        </termEntry>
```

y en formato tabulado para OmegaT:

```
Game theory   Teoria dels jocs    Economics
Human rights  Drets Humans  Economics
Smuggling     Contraban     Economics
Means of production Mitjans de producció       Economics
Poverty       Pobresa       Economics
Innovation    Innovació     Economics
Optimism      Optimisme     Economics
Millionaire   Milionari     Economics
Liquidity trap      Trampa de liquiditat       Economics
Break-even    Punt mort (economia)       Economics
```

4. Códigos de lengua

Los códigos de lengua que se utilizan son los correspondientes a los códigos ISO de dos letras (ISO 639-1). Estos códigos se pueden consultar en el siguiente enlace: http://www.sil.org/iso639-3/codes.asp?order=639_1&letter =% 25

5. Áreas de especialidad

Las áreas de especialidad son las propias de la Wikipedia y se expresarán en inglés. Recordemos, sin embargo, que las áreas de especialidad son libres, es decir, cualquier usuario puede crear una nueva área. Para poder conocer qué áreas de especialidad hay es útil consultar el siguiente enlace: http://en.wikipedia.org/wiki/List_of_academic_disciplines

6. Conclusiones

Con Wikipedia2TBX podemos crear de una manera rápida glosarios terminológicos para una gran cantidad de lenguas y de áreas de especialidad. Estos glosarios se tienen que utilizar con precaución, pero puede ser un gran recurso complementario a recursos terminológicos más clásicos.