Wikipedia2TBX: creació de glossaris terminològics a partir de la Vikipèdia

1. Introducció

La Vikipèdia (http:wikipedia.org) és una enciclopèdia lliure que s'ha creat de forma col·laborativa. Aquesta enciclopèdia és, a més, multilingüe, és a dir, hi ha versions de la Vikipèdia per a un gran nombre de llengües. Per a cada entrada de la Vikipèdia disposem d'enllaços interlingüístics, que ens porten directament al mateix article en una altra llengua. Els articles de la Vikipèdia, a més, estan organitzats per categories. Això fa que es puguin construir glossaris terminològics a partir de les entrades de la Vikipèdia.

El programa que presentem en aquest document és capaç de crear glossaris terminològics multilingües a partir de la Vikipèdia. Per crear-los, el programa necessita saber de quina àrea o àrees temàtiques i quines són les llengües implicades. Amb aquestes dades mira totes les entrades de la Vikipèdia anglesa que tinguin assignades les àrees temàtiques donades i mira si té enllaços interlingüístics per les llengües requerides. A partir de les dades que va recollint confecciona un glossari terminològic multilingüe i crea un fitxer en format TBX (*Term Base Exchange*). També pot crear un glossari en format tabulat per a OmegaT. Aquests glossaris són bilingües i per la qual cosa només es tenen en compte les dues primeres llengües donades al programa, la primera com a llengua de partida, i la segona com a llengua d'arribada. La resta de llengües, per a la creació del glossari en format tabulat, les ignora.

En aquest document explicarem com es fa servir el programa Wikipedia2TBX. El seu funcionament és molt senzill, ja que disposa d'una interfície gràfica d'usuari. També veurem quins codis de llengua s'han de fer servir, així com explicarem com conèixer les classificacions temàtiques de la Vikipèdia.

Abans de fer servir l'eina, cal tenir ben present els següents aspectes:

- Els glossaris terminològics es creen a partir de la informació de la Vikipèdia. Recordeu que la Vikipèdia és un projecte col·laboratiu, on qualsevol usuari pot afegir informació. La riquesa del projecte consisteix precisament en aquest aspecte col·laboratiu. La qualitat del recurs és molt bona, però tots els resultats obtinguts s'han d'entendre com a proposta i per tant queda a criteri del traductor fer servir o no una proposta concreta en les seves traduccions.
- El programa funciona contra una còpia de la Vikipèdia anglesa hostatjada a un servidor de la Universitat Oberta de Catalunya. Aquesta còpia s'actualitza regularment amb les darreres versions, però el contingut pot diferir lleugerament amb la Vikipèdia real.

2. Obtenció, instal·lació i execució del programa

El programa és lliure (sota llicència GNU-GPL) i es pot descarregar de http://lpg.uoc.edu/Wikipedia2TBX. El programa està escrit en Python i és per tant multiplataforma. Es distribueixen 2 versions del programa:

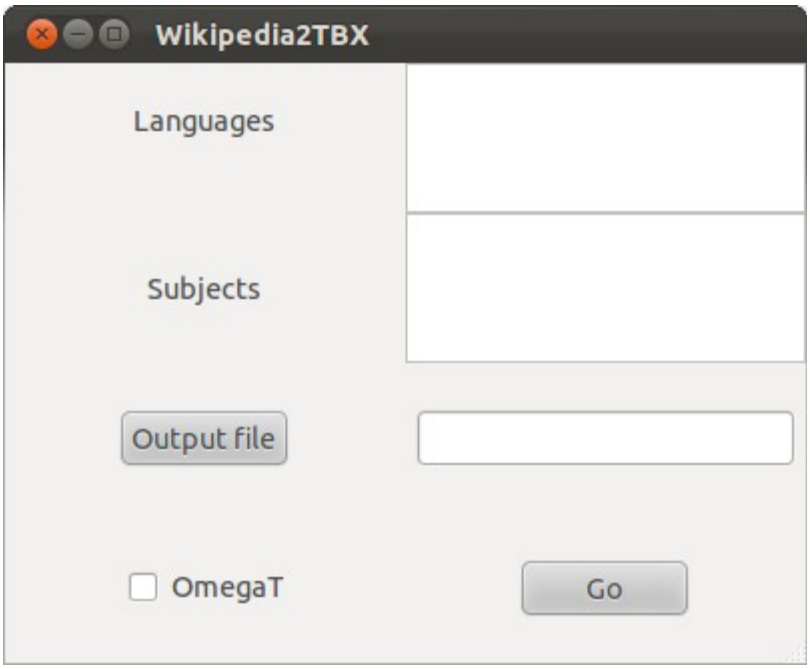
- Arxius font (Wikipedia2TBX-0.1.tar.gz): per executar-lo és necessari tenir instal·lat un intèrpret de Python en el sistema. Aquest fet és l'habitual en Linux i Mac i per tant es podrà fer servir aquesta versió directament sota aquests sistemes operatius si es disposa a més dels prerequisits. Si es vol fer servir aquesta versió sota Windows caldrà instal·lar un intèrpret de Python (versió inferior a la 3.0) que es pot descarregar lliurement de www.python.org
- Versió executable per Windows (wikipedia2tbx-win.zip). Aquesta versió es pot executar sota Windows sense necessitat de tenir l'intèrpret de Python instal·lat al sistema.

El programa no cal instal·lar-lo, simplement cal descomprimir l'arxiu descarregat i fer doble clic sobre l'arxiu Wikipedia2TBX.sh (per Linux o MAC) o Wikipedia2TBX.exe (per Windows). En tots tres sistemes operatius, si es disposa de l'intèrpret de Python, es pot executar des de línia de comades fent:

```
python Wikipedia2TBX.py
```

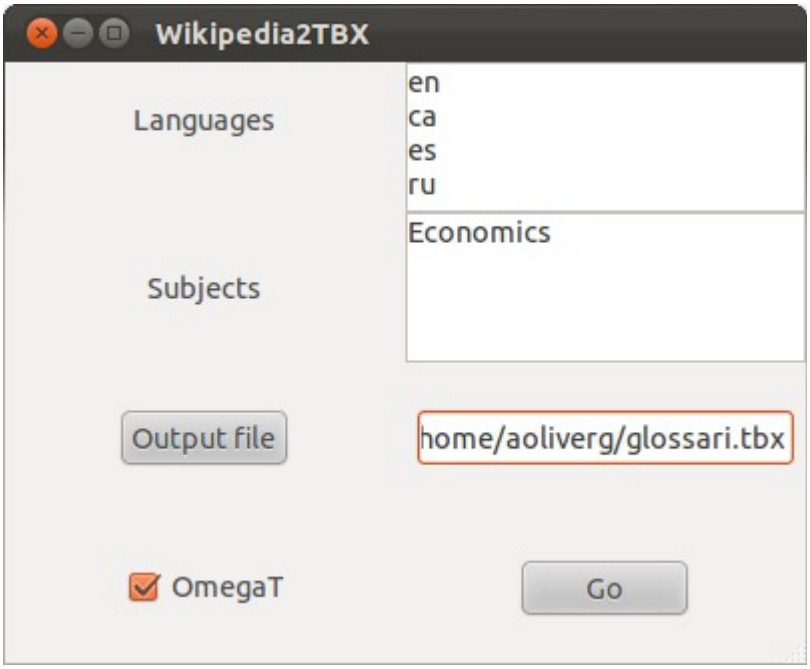
3. Funcionament del programa

Un cop fem doble clic sobre Wikipedia2TBX.py o Wikipedia2TBX.exe s'obre una interfície gràfica molt senzilla.



En aquesta interfície hem de fer:

- En el quadre de text **Languages** hem de posar els codis de llengua (ISO de 2 lletres, mireu l'apartat 4 d'aquest document). Com a mínim hem de posar dos codis de llengua. S'ha de posar un codi per línia
- En el quadre de text **Subjects** hem de posar les àrees d'especialitat que volem (s'han d'expressar en anglès, mireu l'apartat 5 d'aquest document). S'ha de posar una àrea d'especialitat per línia.
- Cal seleccionar la ubicació i el nom del fitxer de sortida amb el botó **Output File**. No és necessari donar l'extensió, ja que automàticament es posarà extensió .tbx.
- El quadre de selecció **OmegaT** cal marcar-lo si volem obtenir la sortida també en format tabulat per OmegaT. En aquest cas es crearà un arxiu en la mateixa ubicació i amb el mateix nom que el fitxer de sortida, però en aquest cas amb extensió .utf8
- Un cop fetes totes les seleccions cal fer clic al botó **Go**. En aquest moment es comença a crear el glossari. El procés pot durar una bona estona i és imprescindible estar connectat a Internet.



A continuació podem observar un fragment del fitxer de sortida 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>
```

i en format tabulat per 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. Codis de llengua

Els codis de llengua que es fan servir són els corresponents als codis ISO de dues lletres (ISO 639-1). Aquests codis es poden consultar al següent enllaç: http://www.sil.org/iso639-3/codes.asp?order=639_1&letter=%25

5. Àrees d'especialitat

Les àrees d'especialitat són les pròpies de la Vikipèdia i s'han d'expressar en anglès. Recordem, però, que les àrees d'especialitat són lliures, és a dir, qualsevol usuari en pot crear. Per poder conèixer quines àrees d'especialitat hi ha és útil consultar el següent enllaç: http://en.wikipedia.org/wiki/List_of_academic_disciplines

6. Conclusions

Amb Wikipedia2TBX podem crear d'una manera ràpida glossaris terminològics per a una gran quantitat de llengües i 'àrees d'especialitat. Aquests glossaris s'han de fer servir amb precaució, però pot ser un gran recurs complementari a recursos terminològics més clàssics.