Universitat Oberta de Catalunya

DIGITHUM
Les humanitats en l'era digital

http://digithum.uoc.edu

Creació automàtica de diccionaris multilingües especialitzats en noves àrees temàtiques*

Joaquim Moré
Tècnic del Departament de Tecnologia Educativa de la UOC
jmore@uoc.edu

Data de presentació: gener de 2009
Data d'acceptació: abril de 2009
Data de publicació: maig de 2009

CITACIÓ RECOMANADA
MORÉ, Joaquim (2009). «Creació automàtica de diccionaris multilingües especialitzats en noves àrees temàtiques» [article en línia]. *Digithum.* Núm. 11. UOC. [Data de consulta: dd/mm/aa].
<adreça electrònica del document>
ISSN 1575-2275

Resum
En aquest article presentem una eina que genera automàticament diccionaris d'equivalències multilingües especialitzats en noves àrees temàtiques. L'eina explota recursos presents a la xarxa per a cercar les equivalències i verificar-ne la idoneïtat. Aquests recursos són, d'una banda, les viquipèdies, que es poden baixar i processar de manera lliure, i, de l'altra, els materials que institucions terminològiques de referència deixen disponibles. Aquesta eina pot ser útil per als docents que elaboren materials didàctics i per als investigadors que preparen tesis, articles o manuals de referència. També pot ser útil per als traductors i per als terminòlegs que s'ocupen de la normalització terminològica d'una nova àrea temàtica en una llengua determinada, els quals estan interessats a conèixer els conceptes que encara no tenen una denominació normalitzada.

Paraules clau
creació automàtica de diccionaris multilingües, Viquipèdia, noves àrees temàtiques, normalització terminològica, fiabilitat de la informació

Abstract
This article presents a tool to automatically generate specialised dictionaries of multilingual equivalents in new subject areas. The tool uses resources that are available on the web to search for equivalents and verify their reliability. These resources are, on the one hand, the Wikipedias, which can be freely downloaded and processed, and, on the other, the materials that terminological institutions of reference make available. This tool is of use to teachers producing teaching materials and

* Aquest treball es va portar a terme al Servei Lingüístic de la UOC durant els anys 2006 i 2007, gràcies als ajuts NORMA de l'Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR).

researchers preparing theses, articles or reference manuals. It is also of use to translators and terminologists working on terminological standardisation in a new subject area in a given language, as it helps them in their work to pinpoint concepts that have yet to receive a standardised denomination.

Keywords

automatic creation of multilingual dictionaries, Wikipedia, new subject areas, terminological standardisation, reliability of the information

Introducció

Actualment neixen moltes àrees temàtiques i, consegüentment, també neixen molts conceptes nous. Algunes de les noves àrees són tan recents que encara no hi ha hagut temps d'elaborar obres de referència amb els conceptes presentats de manera estructurada i amb les denominacions en diverses llengües. Les denominacions dels conceptes apareixen, sobretot, de manera dispersa a la xarxa en articles, estudis, blogs, etc., però moltes d'aquestes denominacions no estan recollides i organitzades en un recurs com un diccionari especialitzat, que podria ser útil per a elaborar materials didàctics, tesis, comunicacions per a congressos o articles de revistes. És cert, però, que el coneixement sobre moltes àrees temàtiques d'aparició recent està organitzat de manera enciclopèdica a la Viquipèdia, amb les denominacions del concepte en diverses llengües a cada entrada. És cert, també, que la Viquipèdia és una font d'informació assequible i que es pot explotar de manera lliure. Ara bé, tot i aquests avantatges, la recopilació manual d'entrades de la Viquipèdia per a fer un diccionari especialitzat multilingüe seria inviable, sobretot per la seva magnitud i per l'allau de conceptes referenciats a les entrades, dels quals s'hauria de destriar els conceptes rellevants i els que són d'àrees temàtiques diferents.

S'ha publicat, molt recentment, una proposta de generació de diccionaris multilingües a partir de la Viquipèdia (Shahid *et al.*, 2009). Ara bé, la recopilació massiva d'entrades de la Viquipèdia s'ha de fer amb prudència, una prudència provocada pel fet que qualsevol usuari de la xarxa, especialista o no, amb bona fe o mala fe, hi pot incorporar entrades i alterar-ne el contingut. Malgrat els procediments de control de la fiabilitat de la informació de la Wikimedia Foundation, responsable del manteniment de la Viquipèdia, i malgrat que especialistes desvinculats de la Fundació considerin que les entrades són rigoroses i de qualitat (Anderson, 2006), les denominacions multilingües de les entrades s'haurien de contrastar amb fonts terminològiques de referència.

En aquest article presentem un generador automàtic de diccionaris especialitzats multilingües a partir de la Viquipèdia que consulta fonts terminològiques de referència disponibles a la xarxa. La generació té en compte les àrees temàtiques (categories) de les entrades i no els termes que apareixen a la definició que estan enllaçats amb altres pàgines, per exemple, com en la proposta de Shahid *et al.* (2009). D'aquesta manera evitem l'aparició en el diccionari de termes que fan referència a altres àrees temàtiques o bé a àrees que es desvien de l'interès de l'usuari.

Aquest article s'organitza de la manera següent: primer fem una presentació general de l'eina i expliquem què fa; seguidament descrivim a grans trets com funciona; en el tercer apartat ens ocupem de com s'obté i s'actualitza el diccionari multilingüe, tenint en compte que la Viquipèdia és dinàmica i es modifica contínuament amb noves aportacions; en el quart apartat expliquem alguns exemples concrets d'ús de l'eina; i l'últim apartat és dedicat a les conclusions i al treball futur.

1. Presentació general de l'eina

L'eina genera automàticament un diccionari d'equivalències multilingüe de l'àrea temàtica que vol l'usuari. L'usuari simplement escriu l'àrea temàtica (per exemple, *e-learning* o *web 2.0)* i l'eina confecciona un fitxer de text tabulat amb les denominacions en anglès, català, espanyol i francès dels conceptes d'aquesta àrea temàtica, indicant les fonts de referència de cada denominació. A la taula 1 es veu una mostra d'un diccionari d'*e-learning*. Si no ha trobat la denominació en alguna llengua, es deixa un espai buit. Aquest fitxer de text tabulat és fàcilment convertible en un fitxer Excel, Access, Calc, etc.

2. Funcionament de l'eina

A la figura 1 mostrem de manera esquemàtica com funciona l'eina.

L'usuari escriu l'àrea temàtica sobre la qual vol crear el diccionari. El component d'interfície amb la Viquipèdia presenta a l'usuari les categories de la pàgina que té com a títol aquesta àrea. Per exemple, a la pàgina de la Viquipèdia anglesa que té com a títol *e-learning* apareix la categoria general *Education* ('Educació') i una categoria més específica *Virtual Learning Environment* ('Entorn d'aprenentatge virtual'). L'usuari ha de seleccionar les categories que l'interessen.

Quan l'usuari selecciona les categories, el motor de creació de diccionaris cerca a la Viquipèdia anglesa les pàgines que tenen alguna de les categories escollides i guarda el títol de la pàgina com

Taula 1: Mostra d'un diccionari sobre *e-learning*

Anglès	Català	Castellà	Francès
Cyberschool	–	–	–
Digication	–	–	–
E-learning	Aprenentatge electrònic *Font: TERMCAT* Aprenentatge virtual *Font: Viquipèdia*	E-learning *Font: Viquipèdia* E-aprendizaje; aprendizaje en línea; aprendizaje por medios electrónicos *Font: IATE*	Apprentissage en ligne *Font: Viquipèdia* *Font: IATE*
Knowledge Machine	–	máquina basada en conocimiento *Font: IATE*	machine à base de connaissances *Font: IATE*
M-learning	–	Aprendizaje electrónico móvil *Font: Viquipèdia*	–
Moodle	Moodle *Font: Viquipèdia*	Moodle *Font: Viquipèdia*	Moodle *Font: Viquipèdia*
Virtual Campus	campus virtual *Font: TERMCAT*	–	campus virtuel *Font: IATE*

la denominació en anglès. Cada pàgina correspon a una entrada. La Viquipèdia anglesa és la font principal de creació del diccionari perquè és la més extensa i amb continguts més actualitzats.

Per cada pàgina que s'ha trobat, el motor de creació de glossaris activa el cercador d'equivalents, el qual cerca les denominacions que són els títols de les pàgines corresponents de la Viquipèdia en català, castellà i francès. A més, també busca l'equivalent en fonts d'institucions terminològiques de referència. Quan busca l'equivalent en castellà i francès, fa consultes a la base de dades terminològica multilingüe de la Unió Europea IATE (InterActive Terminology for Europe) per mitjà de la interfície disponible al seu lloc web.[1] Quan busca l'equivalent en català, cerca l'equivalent en repertoris terminològics que el centre de terminologia Termcat ha posat a disposició del públic i que es poden baixar lliurement.[2] S'ha de dir que quan el cercador no troba l'equivalent en castellà o francès a la base de dades de la IATE, consulta en els repertoris terminològics del Termcat l'equivalent en aquestes llengües.

La denominació que s'ha trobat a la Viquipèdia apareix sola quan coincideix amb la denominació fixada per la institució terminològica, o quan no s'ha trobat una denominació establerta per una institució. En canvi, quan la denominació fixada per la institució terminològica de referència és diferent de la que surt a la Viquipèdia, en el diccionari es presenten totes dues denominacions, indicant-ne la font, tal com es pot veure la taula 1. Vàrem decidir fer-ho així perquè és possible que la denominació fixada per les institucions no sigui la denominació que té més èxit entre els especialistes. La presentació de les dues denominacions i de les fonts permet a l'usuari saber que hi ha denominacions alternatives i que a la pràctica haurà de triar la que li sembli més fiable o convenient.

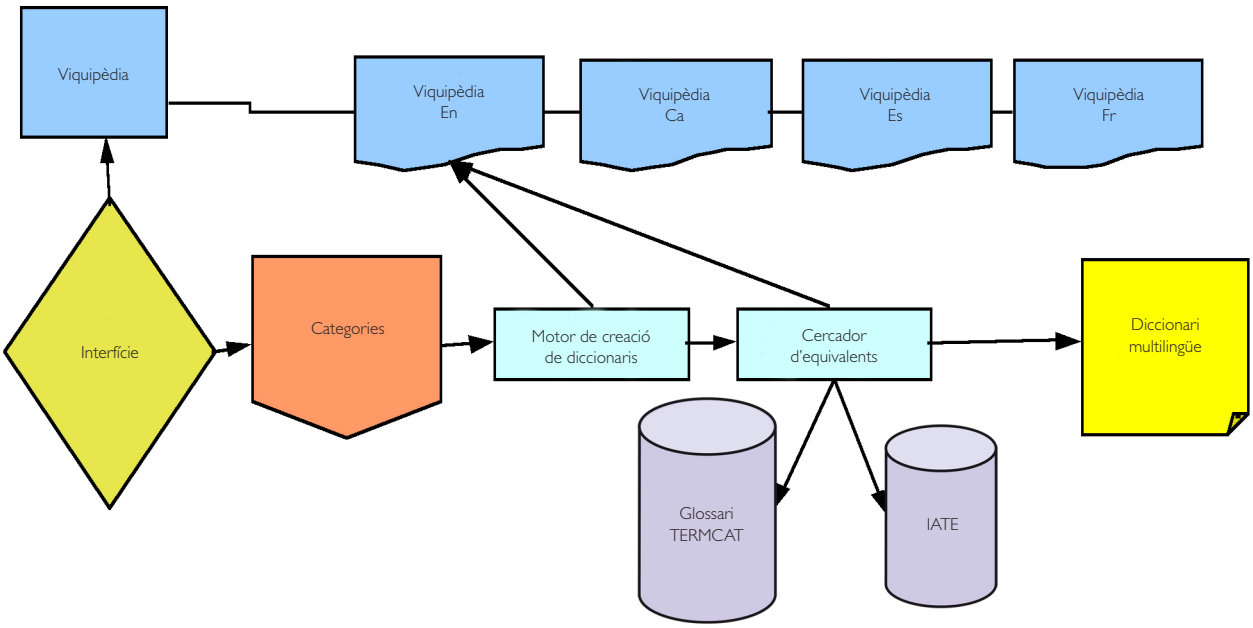
Quan finalitza el procés de recopilació de títols de pàgines amb els equivalents, l'eina presenta a l'usuari les categories dels títols recopilats que encara no ha presentat i espera que l'usuari seleccioni alguna d'aquestes categories per a realitzar el procés de recopilació de títols i cerca d'equivalents amb les noves categories seleccionades. Així es va creant un diccionari més focalitzat a les necessitats d'un usuari que, suposem, està més interessat a obtenir denominacions d'entorns d'aprenentatge virtual que a tenir denominacions de conceptes generals sobre educació. La generació es para quan l'usuari no selecciona cap categoria.

3. Buidatge de la Viquipèdia i manteniment dels diccionaris

La Viquipèdia és una enciclopèdia dinàmica que té unes dimensions molt grans. L'anglesa, per exemple, tenia un total d'1.418.145 articles el 30 de novembre de 2008. Havíem de plantejar-nos com

1. Vegeu l'enllaç: <http://iate.europa.eu>.
2. Vegeu l'enllaç: <http://www.termcat.cat/productes/toberta.htm>.

Figura 1: Esquema del funcionament de l'eina



faríem, de manera ràpida, la recopilació de tots els articles de la Viquipèdia sobre una àrea temàtica i també com faríem el manteniment dels diccionaris d'especialitat multilingües, amb noves incorporacions, a partir de la Viquipèdia més actualitzada.

Buidatge de la Viquipèdia

La interfície amb la Viquipèdia en diferents llengües (vegeu l'apartat 2) és un recurs informàtic, escrit en Perl,[3] que només permet fer consultes a una pàgina. Per a recopilar totes les pàgines que tenen una determinada categoria i extreure'n les denominacions en altres llengües, vàrem plantejar-nos el buidatge de la Viquipèdia anglesa sencera.[4]

Perquè la recopilació fos més ràpida, vàrem generar una versió reduïda de la Viquipèdia anglesa. Aquesta versió es limita a registrar, per cada pàgina, els camps que ens interessen, que són les categories i els títols de les pàgines corresponents en les viquipèdies de totes les llengües disponibles. La versió reduïda es va fer amb un script de Perl que va ser una adaptació de l'script anomenat *WikiPrep* (Gabrilovich *et al.*, 2007). WikiPrep és un script de codi obert i de distribució lliure, creat especialment per Evgeniy Gabrilovich per a la seva tesi doctoral (Gabrilovich, 2006), l'accés al qual la Wikimedia Foundation ha deixat disponible.[5] L'script buida la informació de les pàgines de la Viquipèdia en format XML (figura 2) i obté de manera automàtica el títol de cada pàgina, les seves categories i els equivalents en diverses llengües.

La versió reduïda és un fitxer de text tabulat, on en cada línia s'organitza la informació de la manera següent:

> *Nom entrada* <TAB> *Domini Temàtic1#Domini Temàtic2#...#Domini Temàticn* <TAB>
> *Denominació Llengua1#Denominació Llengua2#Denominació Llengua3...#Denominació Llenguan*

Com a exemple, mostrem a continuació la informació relativa a l'entrada *e-learning* de la Viquipèdia anglesa:

> E-learning<TAB>Alternative education#Computing and society#Educational psychology#Educational technology#Learning#Technical communication#Distance education#Virtual learning environments<TAB>pt:E-learning#tr:E-ö?renme#es:E-learning#da:E-learning#no:E-læring#gl:E-learning#sl:E-zobraževanje#fr:Apprentissage en ligne#nn:E-læring#eo:E-lerno#pl:E-learning#nl:E-learning#de:E-Learning

3. **Vegeu l'enllaç:** <http://search.cpan.org/dist/WWW-Wikipedia/>.
4. Es poden baixar les viquipèdies senceres des d'aquest enllaç <http://download.wikimedia.org/>.
5. Vegeu l'enllaç: <http://meta.wikimedia.org/wiki/Data_dumps#WikiPrep_Perl_script>.

Figura 2: Fragment d'una pàgina de la Viquipèdia anglesa. En blau hi ha marcades les categories temàtiques i en vermell, els títols de les pàgines corresponents en diverses llengües.

```
<page>
    <title>Anarchism</title>
    <id>12</id>
    <restrictions>move=sysop</restrictions>
    <revision>
      <id>49514072</id>
      <timestamp>2006-04-21T23:04:31Z</timestamp>
      <contributor>
        <username>Flexxx</username>
        <id>522242</id>
      </contributor>
      <comment>/* The fight against fascism */</comment>
      <text xml:space="preserve">{{Anarchism}}
'''Anarchism''' is [[etymology|derived from]] the [[Greek language|Greek]]...
[[Category:Anarchism| ]]
[[Category:Political ideology entry points|Anarchism]]
[[Category:Political theories|Anarchism]]
[[Category:Social philosophy|Anarchism]]
[[bs:Anarhizam]]
[[ca:Anarquisme]]
[[cs:Anarchismus]]
[[da:Anarkisme]]
[[de:Anarchismus]]
[[fr:Anarchisme]]
```

Manteniment dels diccionaris

La Viquipèdia és una enciclopèdia dinàmica perquè contínuament s'hi afegeixen entrades i es revisen i s'amplien les entrades ja existents. Per a mantenir els diccionaris amb la informació tan actualitzada com sigui possible, s'hauria de generar la versió reduïda dels continguts actualitzats de la Viquipèdia anglesa amb una periodització que, per exemple, podria ser anual. De tota manera, la periodització pot ser més curta, si així es creu convenient, ja que l'script de Perl facilita la creació de versions reduïdes pel fet que fa aquest procés d'una manera completament automàtica.

4. Exemples d'ús de l'eina

L'eina, com ja hem comentat, pot ser útil per al docent que ha de preparar un material didàctic o per a un investigador que fa un article. També ho és per a un doctorand que necessita una relació dels conceptes bàsics per a preparar el capítol de l'estat de la qüestió de la seva tesi. La rapidesa amb què l'eina elabora el diccionari compensa els dies i els mesos que demanaria fer aquesta relació de manera no automàtica.

Altres professionals també poden fer-ne ús, com els traductors que han de traduir documents especialitzats en un domini temàtic concret. Amb l'eina podrien crear bases de dades terminològiques multilingües per a un sistema de traducció assistida. D'aquesta manera, s'estalviarien la cerca per la xarxa de fonts terminològiques i documentals fiables perquè l'eina ja hauria fet aquesta feina.

Els terminòlegs que s'ocupen de la normalització terminològica de les àrees temàtiques d'aparició recent també podrien fer servir l'eina per a tenir dades sobre la cobertura terminològica d'un domini en una llengua. Per exemple, vàrem agafar una mostra de 40 denominacions angleses d'un diccionari sobre *e-learning* fet amb el nostre mètode i vàrem comprovar que només 6 (el 15%) tenien denominacions normalitzades en català.

5. Conclusió i treball futur

En aquest article hem presentat una eina de creació automàtica de diccionaris multilingües d'àrees temàtiques d'aparició recent. És una eina que permet obtenir automàticament una relació de les denominacions dels conceptes d'una àrea temàtica, les quals solen estar disperses en revistes digitals, blogs i altres fonts d'informació. Per aquesta raó és una eina útil per a docents i investigadors. També pot ser útil per a traductors que han de traduir documents

especialitzats sobre diferents disciplines molt recents. També és una eina útil per als terminòlegs que volen tenir dades sobre la cobertura terminològica d'una nova àrea temàtica.

Com a treball futur ens plantegem la introducció de l'àrea temàtica en la llengua que vulgui l'usuari. Ara per ara, el terme introduït ha de coincidir amb el títol d'una pàgina de la Viquipèdia en anglès, que és la llengua pivot de l'eina. D'altra banda, en cas que una denominació anglesa pertanyi a diverses àrees temàtiques, en el diccionari s'haurien de presentar només els equivalents multilingües de l'àrea temàtica introduïda per l'usuari. L'enllaç a pàgines de les viquipèdies en diferents llengües garanteix aquesta coherència, però s'ha de resoldre en les consultes a les fonts d'institucions terminològiques de referència. Finalment, el nostre procés de buidatge de la Viquipèdia permet obtenir totes les denominacions multilingües del concepte que dóna títol a una pàgina, per la qual cosa en el futur es pot plantejar la generació de diccionaris especialitzats amb un ventall més ampli de llengües. Tampoc no descartem la possibilitat que s'hi puguin afegir les definicions per a cada denominació multilingüe.

Bibliografia

ANDERSON, N. (2006). «Experts rate Wikipedia's accuracy higher than non-experts» [article en línia]. *Ars Technica*. [Data de consulta: 10 de gener de 2009]. <http://arstechnica.com/news.ars/post/20061127-8296.html>

GABRILOVICH, E. (2006). *Feature Generation for Textual Information Retrieval Using World Knowledge*. Tesi doctoral presentada al Technion - Israel Institute of Technology. Haifa (Israel).

GABRILOVICH, E.; MARKOVITCH, S. (2007). «Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis». A: *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*. Hyderabad (Índia).

SHAHID, A.; KAZAKOV, D. (2009). «Automatic Multilingual Lexicon Generation using Wikipedia as a Resource». A: *International Conference on Agents and Artificial Intelligence*. Porto (Portugal).

Joaquim Moré
Tècnic del Departament de Tecnologia Educativa de la UOC
jmore@uoc.edu

Servei Lingüístic
Universitat Oberta de Catalunya
Rambla del Poblenou, 156
08018 Barcelona

Investigador de l'IN3 i tècnic del Departament de Tecnologia Educativa de la Universitat Oberta de Catalunya especialitzat en tecnologies lingüístiques. És llicenciat en Filologia Anglesa i màster en Lingüística computacional per la Universitat de Barcelona. És especialista en tecnologia aplicada a la traducció (traducció automàtica, sistemes de traducció assistida, extracció automàtica de terminologia, etc.) i coautor de publicacions de divulgació sobre aquest tema. També ha treballat en l'explotació automàtica d'informació disponible a la xarxa per a l'escriptura i la traducció de textos (correctors gramaticals i generació de recursos terminològics). Actualment desenvolupa la tesi doctoral entorn de l'avaluació de la traducció automatitzada.

SOME RIGHTS RESERVED

Aquesta obra està subjecta a la llicència de **Reconeixement-No comercial-Sense obres derivades 3.0 Espanya** de Creative Commons. Podeu copiar-la, distribuir-la i comunicar-la públicament sempre que n'especifiqueu l'autor i la revista que la publica (*Digithum*); no en feu un ús comercial i no en feu obra derivada. La llicència completa es pot consultar a http://creativecommons.org/licenses/by-nc-nd/3.0/es/deed.ca.

UOC Universitat Oberta de Catalunya
www.uoc.edu

Núm. 11 | Maig de 2009 ISSN 1575-2275