TÖBBNYELVŰ TEZAURUSZ ÉPÍTÉSE ÉS SZOLGÁLTATÁSA WEBES KÖRNYEZETBEN

Förhécz András, fand_lev@freemail.hu
Mészáros Tamás, meszaros@mit.bme.hu
BME Méréstechnika és Információs Rendszerek Tanszék

Az Európai Unió polgárai szabadon dolgozhatnak és tanulhatnak bármely tagállamban. A munkaerőpiac átjárhatóságát növelendő az EU intézetei különböző eszközöket fejlesztettek a képesítések és diplomák hordozhatósága érdekében, ilyen például a Europass CV[1]. Ezen eszközök fő problémája, hogy az alapvető, a képességeket és képesítéseket leíró fogalmakra nincs nemzetközileg elfogadott terminológia.

A DISCO nemzetközi projekt[2] célja egy többnyelvű tezaurusz létrehozása a képességekről és képesítésekről, mely szabványt állítana egy egyezményes terminológia használatához, és szabadon hozzáférhető eszközöket nyújtana az információ lekérdezéséhez. A tezaurusz létrehozásában felhasználjuk a már létező, képességekkel és képesítésekkel kapcsolatos nemzeti szabványokat és ajánlásokat, ezeket egységes, angol nyelvű rendszerbe foglaljuk, majd elkészítjük az angol tezaurusz fordításait különböző nyelvekre (köztük magyarra is).

A közös terminológia kialakítható a meglévő tezauruszok egyesítésével (*thesaurus merging*): létre kell hozni egy tezauruszt, ami tartalmazza az összes fogalmat és azok fordítását valamennyi nyelvre. Mivel a munkát egy nemzetközi konzorcium végzi, a hagyományos, koncentrált tezaurusz építési megoldások esetünkben nem alkalmazhatók. Kifejlesztettünk egy web-alapú tezaurusz véleményező rendszert, mely támogatja a partnerek együttműködését az egyesítési folyamatban. A véleményező rendszer a tezaurusz szerkesztő szoftver adatai alapján követi a tezaurusz változásait, és a résztvevők megvitathatják a szükséges és már elvégzett módosításokat.

Az elkészített többnyelvű tezauruszt a DISCO Online Tool segítségével lehet elérni. Ez az eszköz egy adatbázis-alapú webrendszer, amely többféle felületen keresztül nyújt a tezauruszra épített szolgáltatásokat. A publikus web portál[3] támogatja a képességek és képesítések lekérdezését, böngészését és fordítását. A felhasználók elkészíthetik saját képesítés profiljukat, amit az önéletrajzukba illeszthetnek. Az Európai Uniós átjárhatóságot segítő portálok (*transparency tools*) webszolgáltatásokon keresztül csatlakozhatnak az Online Tool-hoz, így a Europass CV-hez hasonló szolgáltatások a saját webrendszerükbe integrálva használhatják fel a képességek és képesítések közös terminológiáját.

[1] The Europass Curriculum Vitae, URL: http://europass.cedefop.eu.int/
[2] European Dictionary on Skills and Competencies, URL: http://disco.youtrain.net/
[3] DISCO Portal, URL: http://disco.mit.bme.hu/

DEVELOPING AND PROVIDING MULTILINGUAL THESAURUS IN A WEB APPLICATION

András Förhécz, fand_lev@freemail.hu
Tamás Mészáros, meszaros@mit.bme.hu
BME Department of Measurement and Information Systems

In the European Union citizens are welcome to study or work abroad. In recent years EU institutions have developed tools to establish the transparency of qualifications, like the Europass CV.[4] But all these suffer from the same weakness: the core concepts of these tools—terms for skills and competencies—are neither standardised nor internationally compatible, thus they can only offer limited support.

The DISCO international project[5] intends to fill this gap by providing a terminological support for these tools, that will be publicly available and accessible to non-experts and experts alike. A multilingual thesaurus is developed on the basis of already available national collections in the domain of skills and competencies. These are integrated into one unified English language thesaurus, which is translated to all other languages (including Hungarian).

Full integration can be achieved with thesaurus merging: construction of a common thesaurus with all concepts involved, and translated to all of the languages. As the integration process is accomplished by an international consortium, traditional centralized methods for thesaurus building can not be applied. We have developed a web-based thesaurus review system supporting collaboration during the merging process. The review system imports data from a thesaurus development tool tracking changes in the thesaurus and users can discuss changes made by the responsible partner.

The produced multilingual thesaurus can be accessed through an on-line translation tool for skills and competencies, the DISCO Online Tool. This tool is a database-driven web application providing services based on the thesaurus through multiple interfaces. A public web portal[6] is being developed for querying, browsing and automatic translation of qualifications. Users can combine skills into a competency profile for inclusion into their Curriculum Vitae. The portal will provide web services to European transparency tools, so portals like Europass CV can access the common terminology of skills and competencies integrated into their own web architecture.

[4] The Europass Curriculum Vitae, URL: http://europass.cedefop.eu.int/
[5] European Dictionary on Skills and Competencies, URL: http://disco.youtrain.net/
[6] DISCO Portal, URL: http://disco.mit.bme.hu/

TÖBBNYELVŰ TEZAURUSZ ÉPÍTÉSE ÉS SZOLGÁLTATÁSA WEBES KÖRNYEZETBEN

Förhécz András, fand_lev@freemail.hu
Mészáros Tamás, meszaros@mit.bme.hu
BME Méréstechnika és Információs Rendszerek Tanszék

BEVEZETÉS

Az Európai Unió polgárai szabadon dolgozhatnak és tanulhatnak bármely tagállamban. Egy külföldi tanulmány vagy munkahely megpályázása esetén azonban a nyelvi akadályok problémát jelenthetnek. Különösen fontos, hogy a résztvevő felek tisztában legyenek egymás képességeivel és elvárásaival. Mindkettejük érdeke, hogy egyrészt a munkavállaló tudja, a megpályázott munkakör megfelelő lehetőségeket biztosít a szakmai fejlődésben, másrészt eddigi tanulmányai és tapasztalata alkalmassá teszik a pozíció betöltésére. A jelenlegi gyakorlat szerint erre a szakmai önéletrajz (Curriculum Vitae) és a pályázott lehetőség hirdetésszerű leírása szolgál.

A munkaerőpiac átjárhatósága érdekében az EU intézményei különböző eszközöket fejlesztettek ki, melyek megkönnyítik a képességek és képesítések leírását. Ilyen például a Europass CV [1], mely elektronikus szakmai önéletrajzok elkészítésére szolgál. A felhasználó űrlapok kitöltésével megírhatja részletes, egységes formátumú önéletrajzát, melyet aztán minden jelentkezésénél felhasználhat. Az egységes szerkezet miatt a befogadó intézmény munkatársai is könnyebben megtalálják a számukra releváns információt.

A Europass CV-hez hasonló eszközök fő problémája, hogy a képességeket és képesítéseket leíró fogalmaknak nincs nemzetközileg elfogadott terminológiája. Az egyes szakterületek pontos megnevezése még egy országon belül sem feltétlenül egyértelmű. Ugyan a legtöbb országban léteznek hivatalos jegyzékek az egyes szakmákról és a szakterületen dolgozóktól megkövetelt képesítésekről – gyakran ezeknek törvényi szabályozása is adott –, eszközök hiányában használatuk nem általános.

A DISCO nemzetközi projekt [2] célja egy többnyelvű tezaurusz létrehozása a képességekről és képesítésekről, mely szabványt állít egy egyezményes terminológia használatához. A fogalmi jegyzék tartalmazza a fogalmak nyelvi változatait, mely lehetővé teszi az automatikus fordítást egyik nyelvről a másikra. A projekt során elkészítettük a fogalomtár fejlesztésekor használt véleményező rendszer, valamint a tezaurusz képességeit bemutató minta alkalmazást. Utóbbi bemutatja a közös terminológia használatának előnyeit, alapvető eszközöket adva a felhasználók kezébe az információk lekérdezéséhez és a fordításhoz. A projektben résztvevő partnereknek megfelelően a tezaurusz rendelkezésre áll majd hat különböző nyelven (angol, cseh, francia, litván, magyar, német).

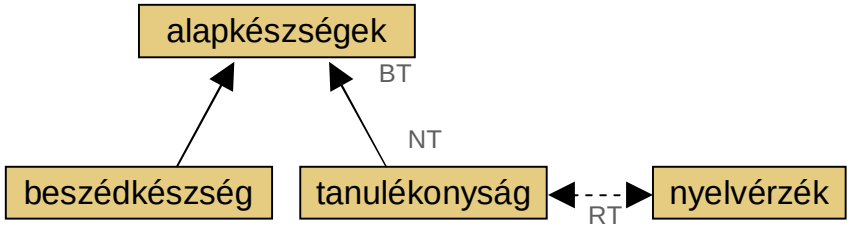
A tezaurusz létrehozásában felhasználtuk a már létező, képességekkel és képesítésekkel kapcsolatos nemzeti szabványokat és ajánlásokat. Szinte valamennyi országban jogi szabályozása van az egyes szakemberektől megkövetelt kompetenciáknak. Bár ezek részletessége nagyon eltérő, mégis az egyik legfontosabb forrásai egy összevont jegyzék

készítésének. Felhasználtunk továbbá két szorosan kapcsolódó szabványos gyűjteményt. Az ISCED [3] az UNESCO által létrehozott jegyzék az oktatás színvonalának és rendszerének nemzetközi összehasonlítására. Tartalmazza az oktatás által lefedett szakterületek hierarchiáját. Az ISCO [4] a foglalkozásokat rendszerezi, ami ugyan csak közvetve kapcsolódik a képesítésekhez, a kategóriák és a felsőbb szintek rendszere átültethető volt a DISCO készítésekor.

TEZAURUSZ

Milyen struktúrát is takar tulajdonképpen egy tezaurusz? Fogalmak és a köztük lévő szemantikus relációk leírására szolgál. Az alapja tulajdonképpen egy taxonómia, egy olyan fogalmi struktúra, ahol a fogalmak értelmének leírására az „általánosabb értelmű” (*broader term, BT*) és „szűkebb értelmű” (*narrower term, NT*) bináris relációkat használjuk. A taxonómiával ellentétben megengedjük, hogy egy fogalomnak több őse, általánosabb értelmű megfelelője legyen, amit polihierarchiának hívünk, ahogy arra a DISCO építése során szükség van. Például a „fordítási és tolmácsolási képesség” a „kommunikációs és nyelvi készségek” és a „művészetek” kategória alá is besorolható.

A tezauruszok további, de előre rögzített számú és fajtájú relációt is tartalmazhatnak, például gyakori a kapcsolódó fogalmak (*related term, RT*) használata. Ezen felül az egyes fogalmakhoz attribútumokat rendelhetünk, a DISCO esetében például definíciót és szinonimákat. Egy ilyen struktúra látható az 1. ábrán.



1. ábra: Tezaurusz részlete

Amint azt korábban említettük, a projekt célja a képességek és képesítések többnyelvű tezauruszát létrehozni a már létező nemzeti gyűjtemények és néhány egyéb forrás alapján. A tezaurusz valójában egy megállapodás a közös fogalmak egyezményes megnevezéséhez [5]. A projektben résztvevő partnereknek már rendelkezésükre áll ezen közös fogalmak egy része, a DISCO tezaurusz megteremti az átjárhatóságot a terminológiák között.

Fogalomtárak összevonására alapvetően két különböző módszer használható. A tezauruszok összekapcsolása (*thesaurus mapping*) során megkeressük az azonos vagy hasonló jelentésű fogalmakat és relációkat az egyes tezauruszok között, és ezek segítségével egységes keretbe foglalhatjuk őket. Az átjárhatóság függ a megtalált kapcsolatok számától, és attól, hogy az egyes források mennyire fedik át egymást. A mi esetünkben ha például egy szakterület hiányzik valamelyik ország jegyzékéből, nem tudjuk lefordítani a fogalmakat arra a célnyelvre.

A teljes egyesítés (*thesaurus merging*) során egy közös tezauruszt építünk, mely tartalmazza az összes releváns fogalmat, miközben felhasználjuk a forrásként használt terminológiák fogalmait és kapcsolatait. Az egyesítés lényegesen összetettebb feladat, mivel az egyesítendő tezauruszok közötti eltéréseket fel kell oldani. Az egymásnak ellentmondó kapcsolatok úgy

szüntethetőek meg, ha az eredeti jelentéseket csak közelítjük, és közös megegyezésre új egyezményes terminológiát fektetünk le.

Az összevonás módszerének megválasztásánál figyelembe kell venni a többnyelvű tezauruszok fejlesztésének sajátosságait is. Az egyik legfontosabb kérdés, hogy a fordítások ugyanarra a struktúrára épülnek, vagy az azonos fogalmak közötti relációk eltérnek a különböző nyelvi változatok között. Ha egységes a struktúra, és a többnyelvűség csak a fogalmak lefordítását, nyelvenként eltérő szinonimákat jelent, akkor szimmetrikus tezauruszról beszélünk [6]. Aszimmetrikus tezaurusznál a kapcsolatok is eltérhetnek, vagy akár egyes fogalmak csak bizonyos nyelveken léteznek.

A DISCO esetében nyelvfüggetlen, egységes jelölésrendszert hozunk létre a képesítések leírására. Mivel a partnerek saját fogalomtárai nagyon eltérő részletességűek, általában nem teljesítik a kész tezaurusztól elvárt minőséget, az egyesítés módszerével új szimmetrikus fogalomtárat hozunk létre. A szimmetria ellen két jelenség szól. Egyrészt egyes fogalmak nem fordíthatóak le bizonyos nyelvekre. Ilyenkor megpróbálunk új kifejezéseket, szókapcsolatokat alkotni, vagy ha ez nem lehetséges, a fogalmat annak definíciójával reprezentáljuk. A másik problémát a fogalmak megszokott rendszerezésében előforduló eltérések jelentik, amik eltérő relációkat eredményeznének. Például a nyomdát Magyarországon az ipar, Nyugat-Európában általában a média területére sorolják. Mivel az ilyen eltérések viszonylag ritkák, a feladat egyszerűsítése végett egy struktúrát alakítunk ki, amely közös lesz minden nyelvi változatban, és az ilyen eltéréseket a fogalmak leírásában jelöljük csak.

VÉLEMÉNYEZŐ RENDSZER

A tezaurusz fejlesztését egy nemzetközi konzorcium végzi, ezért a hagyományos, koncentrált tezaurusz építési megoldások önmagukban nem alkalmazhatóak. Több olyan eszközt is kifejlesztettek, ami tudásbázisok elosztott építését teszi lehetővé, különösen az ontológia szerkesztés területén (például *Ontolingua*, *KAON Engineering Server*). Az ilyen eszközökkel a felhasználók egy időben módosíthatják a tudásbázis tartalmát egy kliens-szerver architektúrával rendelkező keretrendszer segítségével.

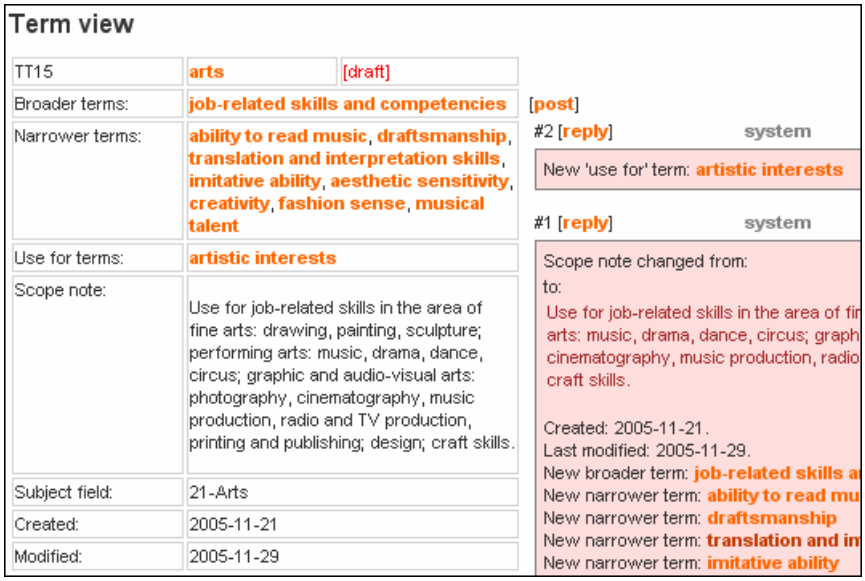
Ugyan ezek az eszközök rendelkezésre állnak, a projekt során nem alkalmaztunk elosztott szerkesztőt a tezaurusz mérete és bonyolultsága miatt. A kész fogalomtár várhatóan 2-3000 fogalmat és nyelvenként körülbelül 1000-5000 szinonimát fog tartalmazni. A konzisztencia biztosítása – például duplikátumok elkerülése – miatt célszerűbb, hogy egy felelős partner szerkessze a tezauruszt, míg a többiek csak módosítási javaslatokat tehetnek, figyelemmel kísérve a változásokat.

Először a forrásként szolgáló nemzeti gyűjteményeket a közös nyelvre, angolra fordítjuk le, majd egy megbízott partner szakterületenként megpróbálja azokat egyesíteni. Természetesen előfordulhatnak hibák és hiányosságok, elsősorban az egyes országok sajátosságai esetében, amit a partnerek módosítási javaslatként jeleznek. A végleges terminológiát egyesítő és véleményező szakaszok sorozatával alakítjuk ki. Ha egy szektor elkészült, az angol tezaurusz fogalmait valamennyi nyelvre lefordítjuk.

A fenti folyamatot támogató kész terméket nem találtunk, ezért kifejlesztettünk egy web-alapú tezaurusz véleményező rendszert, mely támogatja a partnerek együttműködését az egyesítési folyamatban. A véleményező rendszer a tezaurusz szerkesztő szoftver adatai alapján követi a tezaurusz változásait. Így a véleményezés független a tezaurusz fejlesztésénél használt

professzionális eszköztől, csak a szerkesztésért felelős partnernek kell tisztában lennie utóbbi használatával.

A véleményező rendszer segítségével áttekinthető a tezaurusz szerkezete, a „szűkebb” és „általánosabb értelmű” relációkon keresztül bejárható a fogalmak hierarchiája. Egy fogalmat kiválasztva megtekinthető annak minden tulajdonsága, valamint a szerkesztés során végrehajtott változtatások története. A résztvevők gyakorlatilag minden egyes fogalom esetén az internetes fórumokon megszokott módon vitathatják meg a szükséges és már elvégzett módosításokat. Az átláthatóság kedvéért a hozzászólások és a rendszer által generált, a változtatásokat leíró üzenetek időrendben egymás alatt jelennek meg. A 2. ábrán a „foglalkozásokhoz kötődő kompetenciák” fogalom látható az egyezményes angol nyelvű tezauruszban, a véleményező keretrendszerben.



2. ábra: Véleményező rendszer

A partnerek szektoronként készítik el a DISCO tezauruszt egy vagy két véleményező fázis beiktatásával. Ha megfelelő a fogalmak száma és minősége, elkészülhetnek a nyelvi fordítások. A tezaurusz szemantikája szerint minden fogalomnak van egy hivatalos megnevezése (*preferred name*), valamint tartozhatnak hozzá szinonimák, melyek ugyanazt a jelentést hordozzák, de segítik a felhasználókat az orientációban és a keresésben. A tezaurusz fordítása során a hivatalos megnevezést kell lefordítani, pontosabban ugyanannak a fogalomnak az országon belüli hivatalos nevét megtalálni. A szinonimákat minden nyelvre külön-külön össze kell gyűjteni, a más-más nyelvű szinonimák nincsenek megfeleltetve egymásnak.

DISCO PORTÁL

Az elkészített többnyelvű tezauruszt a DISCO Online Tool segítségével lehet elérni. Ez az eszköz egy adatbázis-alapú webrendszer, amely többféle felületen keresztül nyújt a tezauruszra épített szolgáltatásokat.

A publikus web portál[7] támogatja a képességek és képesítések lekérdezését, böngészését és fordítását. A tezaurusz a portál nyelvétől függetlenül tetszőleges nyelven bejárható, a fogalmak megjelenítésénél (3. ábra) a fontosabb kategóriáknál egy leírás mező segíti a felhasználót. A portál két fontos szolgáltatása a fogalmak fordítása és a személyes profil készítése. A felhasználók elkészíthetik saját képesítés profiljukat, amit az önéletrajzukba illeszthetnek. A fogalomtárat böngészve kiválaszthatják azokat a képességeket vagy képesítéseket, melyekkel rendelkeznek. Az önéletrajz írásakor a személyes profil tartalmát tetszőleges nyelven exportálhatják.

3. ábra: DISCO Portál, fogalom nézet

Az Európai Uniós átjárhatóságot segítő portálok (*transparency tools*) webszolgáltatásokon keresztül csatlakozhatnak az Online Tool-hoz, így a Europass CV-hez hasonló szolgáltatások a saját webrendszerükbe integrálva használhatják fel a képességek és képesítések közös terminológiáját. A Europass CV oldala jelenlegi formájában a kompetenciák kitöltését szabad szöveges űrlapként kínálja fel, helyette a DISCO Portálon elérhető személyes profilhoz hasonló szolgáltatást nyújthatna.

ÖSSZEFOGLALÁS

A DISCO nemzetközi tezaurusz a képességek és képesítések hivatalos jegyzékévé válhat az Európai Unióban. A szakterületen dolgozók számára egyezményes megnevezéseket vezet be a kompetenciák megjelölésére, és lehetővé válik automatikus nyelvi fordításuk. A projektben elkészítettünk egy véleményező rendszert, amivel a partnerek elosztott tezaurusz szerkesztő

[7] DISCO Portál, URL: http://disco.mit.bme.hu/

szoftverek nélkül építhetik a fogalomtárat. A DISCO Online Tool a tudásbázisra építhető szolgáltatásokat demonstrálja, kihangsúlyozva a gépi fordítás és az önéletrajz készítésnél nyújtott támogatás előnyeit.

Hivatkozások

[1] The Europass Curriculum Vitae, URL: http://europass.cedefop.eu.int/

[2] European Dictionary on Skills and Competencies, URL: http://disco.youtrain.net/

[3] International Standard Classification of Education, ISCED 1997, URL: http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm

[4] International Standard Classification of Occupations, ISCO-88, URL: http://www.warwick.ac.uk/ier/isco/isco88.html

[5] M. Doerr, "Semantic Problems of Thesaurus Mapping," *Journal of Digital Information*, 1(8), Apr. 2001. URL: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

[6] IFLA, Working Group on Guidelines for Multilingual Thesauri, "Guidelines for Multilingual Thesauri", Apr. 2005