Többnyelvűség a digitális könyvtárban. Szakirodalmi szemle

Bevezetés

A globalizáció erősödésének és az internet terjedésének következtében a digitális könyvtárak nemcsak a határokon, hanem a nyelvi korlátokon is átlépnek. A többnyelvű tartalmat kínáló, illetve a több nyelven is kereshető gyűjtemények száma folyamatosan növekszik, akárcsak az igény és az érdeklődés iránt. Jó néhány publikáció is megjelent már, melyek ezeknek a szolgáltatásoknak a sajátosságait: technikai, nyelvészeti, együttműködési kérdéseit tárgyalják, vagy bemutatnak – jellemzően még kísérleti fázisban levő – projekteket, de a téma szakirodalmát áttekintő összefoglaló eddig még nem készült. Ez a cikk arra tesz kísérletet, hogy summázza a keresztnyelvi információvisszakeresés, az információkhoz való többnyelvű hozzáférés és a digitális könyvtárak kapcsolatával foglalkozó kutatások eredményeit, az azokról beszámoló hírek és tanulmányok alapján. A szemlézés az *ACM*, az *ERIC*, a *Library Literature* és a *Library, Information Science & Technology Abstracts* adatbázisokból kikeresett, angol, holland, német, illetve angolra fordított egyéb nyelvű publikációkból történt.

A *cross-language information retrieval* (*CLIR*) szakkifejezés olyan technológiák összefoglaló elnevezése, amelyekkel a felhasználók nyelveken átnyúló kereséseket végezhetnek, például egy dokumentumtárból hollandul, kínaiul, vagy éppen arabul írt releváns tételeket találhatnak meg egy forrásnyelven (pl. angolul) megfogalmazott keresőkérdéssel, sőt ezeket a találatokat akár géppel le is fordíttathatják maguknak. A *multilingual information access* (*MLIA*) ennél szélesebb fogalom – magában foglalja ugyan a CLIR technikákat is, de jelentése nem csupán a visszakeresésre korlátozódik, hanem olyan területeket is ide sorolnak, mint például az automatikus kivonatkészítés különböző nyelvű dokumentumokból, vagy a nyelvek között is működő kérdésértelmező és válaszadó rendszerek.

A többnyelvű digitális könyvtárak különféle nyelveken íródott, digitalizált vagy digitálisan született dokumentumokat szolgáltatnak, vagy ha egynyelvű is a tartalmuk, de egynél több nyelven kereshetők, illetve érhetők el. Ezek a szolgáltatások különböző országok, régiók és kultúrák értékeit gyűjtik össze, gyorsan és egyszerűen teszik hozzáférhetővé világméretekben az információk egy széles körét, biztosítják a kulturális örökség megőrzését, és elősegítik az együttműködést, a nemzetek közötti megértés elmélyítését. Tartalmukat tekintve igen sokfélék lehetnek, találunk közöttük orvosi, gazdaságtudományi és jogi információforrásokat, gyermekirodalmat, újságkivágás-gyűjteményt, ősi spanyol térképeket, szerb kulturális értékeket, indiai disszertációkat stb., és ennek megfelelően a felhasználóik köre is változatos és széles.

Példák a többnyelvű digitális könyvtárakra és technológiákra

Bár a téma szakirodalma jelentős részben tervezési vagy prototípus szinten levő rendszerekkel, lehetséges jövőbeli projektekkel, vagy együttműködési javaslatokkal foglalkozik, azért akadnak a publikációk közt már működő szolgáltatásokról szóló beszámolók is. *Chen* és *Ruiz* 2009-ben 150 digitális könyvtárat vizsgált meg az Egyesült Államokban, és csupán öt többnyelvűt talált (a *Meeting of Frontiers*, a *France in America*, a *Parallel Histories*, a *Perseus Digital Library* és az *International Children's Digital Library*). Utóbbi, vagyis az *ICDL* (*childrenslibrary.org*) gyűjteményében levő digitalizált gyermekkönyvek több mint ötvenféle nyelven íródtak. A *University of Maryland* és az *Internet Archive* közös vállalkozása nemcsak a gyerekek és szüleik, valamint a gyermekirodalmat kutatók számára értékes forrás, hanem a többnyelvű hozzáféréssel foglalkozó szakembereknek is hasznos kísérleti platform. A *World Digital Library* (*wdl.org*) egy nemzetközi vállalkozás, melyet az *UNESCO* és a *Library of Congress* működtet nemzeti könyvtárak és más intézmények bevonásával, és többek között a Google és a Microsoft szponzorálásával. A tartalma nagyon széles kört fed le, mind kulturálisan, mind pedig a dokumentumok típusát tekintve. Keresőfelülete hét nyelven használható. A *Digital Library of the Caribbean* (*dloc.com*) a karibi térség kulturális, történeti és tudományos anyagait gyűjti. Finanszírozása az USA Oktatási Minisztériuma által létrehozott alapból és a partnerintézmények hozzájárulásaiból történik. Az európai digitális könyvtár, az *Europeana* (*europeana.eu*) több fejlődési szakaszon ment át, míg elérte a jelenlegi formáját; az EU tagországok nyelvein használható felületén át számos nemzet kulturális, illetve tudományos jellegű digitális dokumentumai között kereshetünk. Ugyancsak uniós támogatással, a *Network of European Economists Online* projekt

keretében működik a bibliográfiai adatokat, nyílt hozzáférésű publikációkat és adatállományokat tartalmazó, közgazdasági témájú *Economists Online* (*economistsonline.org*) szolgáltatás. A *Virtual Catalogue for Art History* (korábbi nevén: Virtueller Katalog Kunstgeschichte) pedig egyfajta metakatalógus, mellyel európai művészeti intézmények adatbázisaiban keresgélhetünk (*artlibraries.net*).

Többnyelvű digitális könyvtárak létrehozásához és fenntartásához együttműködésekre van szükség, nemcsak országok, hanem szakterületek (pl. számítástechnika, könyvtártudomány, muzeológia, művészettörténet, nyelvtudomány, természetes nyelvi feldolgozás), illetve intézménytípusok (pl. közgyűjtemények és informatikai vállalkozások) között. Európában több program is segíti ezeket az együttműködéseket. Érthető, hogy itt fokozott figyelmet kapnak az ilyen irányú kutatások és kooperációk, hiszen az EU működtetése során mindennapos igény a keresztnyelvi információ-visszakeresés. A *CACAO* (*Cross-Language Access to Catalogues and On-line Libraries*) projekt célja egy olyan infrastruktúra kialakítása, amellyel a felhasználók online katalógusokat és digitális könyvtárakat kérdezhetnek le valamelyik európai nyelven, miközben azok tartalma más nyelveken íródott. A projekt keretében kifejlesztett technológia azután beépülhet olyan szolgáltatásokba, mint amilyen például a *TEL* (*The European Library*). A CACAO az európai *CLEF* (*Cross-language Evaluation Forum*) kezdeményezés egyik résztvevője. A *DELOS Network of Excellence* szintén egy európai program, mely a digitális könyvtárak kutatásával és az ezekhez kapcsolódó műszaki megoldások fejlesztésével foglalkozik. A DELOS egyik eredménye a *DelosDLMS* nevű moduláris digitális könyvtári rendszer, amelyben többnyelvűséget támogató funkciók is vannak.

Említést érdemel még néhány olyan nyelvi eszköz, amelyek digitális könyvtárakba is integrálhatók. Ilyen például a kétnyelvű (kínai és angol) szótárra épülő *MTIR* információ-visszakereső rendszer. Ez nemcsak a keresőkérdést fordítja le (megengedve a tulajdonnevek transzliterált beírását is), hanem a találatokat is visszafordítja a felhasználó nyelvére. Az MTIR HTML címkéket használ a gépi fordításhoz és HTTP protokollon át kommunikál, így könnyen beépíthető webes alkalmazásokba. A *SPIRIT* (*Syntactic and Probabilistic Indexing and Retrieval of Information in Texts*) technológia előzményei még a nyolcvanas évekre nyúlnak vissza. Egynyelvű (francia, illetve angol) keresőrendszerből fejlesztették tovább egy keresztnyelvi információ-visszakereső eszközzé. Az *Eurovision* képek megtalálását segíti: a keresőkérdéseket angolra fordítja, majd lefuttatja őket a képaláírások angol nyelvű adatbázisában. A *SIS-TMS* nevű eszközzel többnyelvű tezauruszokat kapcsolhatunk össze, így ezek hasznos tudásbázisul szolgálhatnak CLIR alkalmazásokhoz. A *SyDoM* pedig egy olyan technológia, amely többnyelvű ontológiát használ annak eldöntésére, hogy a digitális dokumentumokból milyen szavakat/kifejezéseket gyűjtsön ki és indexeljen le a későbbi visszakereséshez.

Módszerek a nyelvi akadályok leküzdésére

A CLIR szakirodalomban többféle megoldással is találkozunk az egy adott nyelven megfogalmazott információs igény (a keresőkérdés) és a más nyelve(ke)n íródott tartalom (a dokumentumok) összepároztatására. Az egyik lehetőség a kérdés átfordítása a dokumentum nyelvére, a másik a dokumentum lefordítása a keresőkérdés nyelvére, a harmadik pedig mindkettőnek egy köztes alakra (ún. interlingvális reprezentációra) való átalakítása. A többnyelvű digitális könyvtárak kutatói egy negyedik lehetséges megoldást is említenek: a leíró metaadatok lefordítását, ami kétségtelenül hatékonyabb megoldás lehet, mint a teljes dokumentum gépi fordítása, sőt ha nem szöveges anyagokat tartalmaz a gyűjtemény, akkor különösen hasznos (feltéve, hogy dokumentumszinten vannak leírva a digitális objektumok). Egy további, részleges megoldás lehet az, amikor csak a közös szavakat, vagyis amelyek mindkét nyelvben azonosak vagy hasonlók (pl. a tulajdonneveket) veszik figyelembe a keresés során.

A fordításhoz szükséges háttértudás származhat többnyelvű szótárakból, tezauruszokból és gépi fordítórendszerekből, illetve statisztikai módszerekkel is előállítható nagyobb szövegkorpuszokból. Mindegyikre vannak példák a digitális könyvtárak esetében is. A könyvári katalógusokban használt tárgyszavak és az olyan nagy *KOS* (*knowledge organization system*) rendszerek, mint amilyen a Library of Congress Classification vagy a Library of Congress Subject Headings, jó kiindulási alapot jelenthetnek többnyelvű tudásbázisokhoz. Ezek biztosan relevánsabbak, mint a bibliográfiai adatokból vagy a teljes szövegből automatikus módszerekkel kigyűjthető kulcsszavak, hiszen a fogalmakat szakemberek választják ki és rendelik hozzá a dokumentumokhoz. Hasonló okból készítették el a MeSH (Medical Subject Headings) tárgyszó-

rendszer kínai fordítását az orvosi témájú webhelyek közötti keresés megkönnyítésére. Automatikus tudástár-építésről is vannak beszámolók a szakirodalomban: például kínai és angol nyelvű szövegekből előállított kétnyelvű fogalomtár, illetve különböző nyelvű, de azonos tartalmú jogi dokumentumokból generált „hasonlósági tezaurusz". Egyes kutatók pedig kevert módszerekkel kísérleteznek: például szótárak és ontológiák együttes használata eltérő nyelvű digitális könyvtárakban való föderált kereséseknél, illetve weboldalakon végzett szövegbányászat a fordítási szótár szókészletének bővítéséhez.

Bár a többnyelvű digitális könyvtárak nyelvi korlátainak áttöréséről szóló publikációk elsősorban a fordításhoz szükséges tudásbázissal foglalkoznak, a CLIR szakirodalomban gyakran esik szó egyéb problémákról is, amelyek negatív hatással lehetnek a gépi fordítás, és ezáltal az információ kinyerés pontosságára. A többjelentésű szavak és a szinonimák még egy egynyelvű rendszerben is megnehezítik a visszakeresést, és minden újabb nyelv beépítésével hatványozódnak a gondok. A fordítás során három komolyabb hibaforrás léphet fel: bizonyos szavak (pl. szakkifejezések, rövidítések, tulajdonnevek) hiánya a másik nyelvben; a nem összetett mondatok helytelen feldarabolása a nyelvi elemzés során; és a többféleképpen fordítható szavakból származó bizonytalanság. Mivel a digitális könyvtárakban a felhasználók rendszerint csak néhány keresőszót írnak be, ezért ha ezek közül egyet vagy esetleg többet nem sikerül lefordítani, a visszakeresési folyamat teljesen kudarcba fulladhat. Ezért fontosak azok a kísérletek, amelyek a szótárak automatikus módszerekkel való bővítésére irányulnak.

A dokumentumok és metaadataik tárolása és szolgáltatása is felvet megoldandó nyelvi feladatokat a digitális könyvtárakban. Le kell például fordítani az útmutatókat, a metaadatok űrlapjait, a különböző listákat, és a könyvtári rendszer kezelőfelületét is (szoftverlokalizálás). Az eltérő nyelvű dokumentumok indexelése sem egyszerű, mivel minden nyelvnek megvannak a maga sajátosságai és szabályai, amelyek alapján eldönthető, hogy mely szavakat érdemes belevenni az indexekbe és melyeket célszerű kizárni belőlük. Az optikai karakterfelismerés szintén sajátos problémákkal jár, különösen a nem latin betűket használó nyelveknél.

A szövegek számítógépes tárolása, feldolgozása (pl. a visszakereséshez szükséges indexelése) és képernyőn való megjelenítése megfelelő karakterkódolást igényel. Bár sokféle szabvány van érvényben, az interneten a leggyakoribb az UTF-8-as Unicode kódolás. Az Unicode kódtábla elvileg bármilyen írott nyelv karaktereinek reprezentálására alkalmas, de a gyakorlatban még nincs benne minden létező nyelv. Az elterjedt böngészőprogramok már mind támogatják az UTF-8 kódolást, és egyes digitális könyvtári szoftverek (pl. a Greenstone) is képesek nem latin betűs szövegek kezelésére. A különböző kódszabványok keveredése, és az, hogy egyik sem fedi le az összes nyelvet, mindenesetre gondokat okozhat egyes gyűjteményekben.

Mivel a többnyelvű digitális könyvtári projektek gyakran nemzetközi együttműködések keretében zajlanak, nem szabad elfeledkezni a kulturális különbségekből fakadó nehézségekről sem. Több esettanulmányt is olvashatunk a szakirodalomban, melyek arról számolnak be, hogy a szövevényes és a kívülállók számára nehezen érthető kulturális sajátosságok milyen módon befolyásolják például a szoftverfejlesztést, a szolgáltatások használatát, a digitális tartalmak értelmezését, vagy akár a projektek finanszírozását.

Az együttműködési képesség alapvető feltétel a digitális könyvtárak sikeréhez. Többnyelvű, nemzetközi projekteknél nemcsak műszaki szintű interoperabilitásra van szükség (pl. egy közös kereső megvalósításához), hanem társadalompolitikai és szemantikus szinten is meg kell teremteni az együttműködés feltételeit. Szemantikus interoperabilitás olyankor szükséges, amikor különböző tezauruszokat használó gyűjteményeket egyesítenek, vagyis eltérő tudásstruktúrákat kell összefésülni.

Kutatási területek

A többnyelvű digitális könyvtárakkal foglalkozó publikációk jelentős része a rendszerek értékelésével foglalkozik. Egyes kutatók prototípusokat építenek, hogy kipróbálhassák rajtuk az ötleteiket. Mások a már működő rendszereket tesztelik különböző feladatokkal, majd rangsorolják őket. Az első olyan értékelési kampány, amely kifejezetten a keresztnyelvi információkeresésre fókuszált, az NTCIR Workshop keretében zajlott 1999-ben. Ezt követték azután a már említett CLEF által meghirdetett tesztelési akciók, melyek az európai nyelvekre korlátozódnak és egyre „realisztikusabbak", vagyis közelítenek a valós felhasználói szokásokhoz és igényekhez. Az évek során nagy mennyiségű kísérleti adat gyűlt össze, ezek hasznosítha-

tók a további kutatásokhoz, illetve ösztönözhetik a fejlesztéseket.

A rendszerekre koncentráló kutatások mellett jóval kisebb arányban ugyan, de vannak azért olyan vizsgálatok is, amelyek a felhasználókra vonatkoznak. Az International Children's Digital Library esetében például gyerekekkel véleményeztették a szolgáltatás külalakját és megnézték azt is, hogy hogyan keresnek a könyvtárban. Egy másik kutatásban a kétnyelvű tezauruszra épülő, *Searchling* nevű keresőfelületet teszteltették 15 felhasználóval, akiknek három keresési feladatot kellett megoldaniuk. Az *Eurovision* képkereső rendszert szintén alávetették a fejlesztői egy ilyen tesztelésnek, itt két feladatot kaptak a felhasználók. Mindezen példák ellenére elmondható, hogy a szakirodalomban nem sok információt találni arról, hogy kik, hogyan és milyen mértékben használják a többnyelvű digitális könyvtárakat. Minél több olyan kutatásra lenne szükség, amelyek valós szituációkra, valódi felhasználókra és üzemszerűen működő szolgáltatásokra vonatkoznak.

/DIEKEMA, Anne R.: Multilinguality in the digital library. A review. = The Electronic Library, 30. köt. 2. sz. 2012. p. 165–181./

(Drótos László)

Személyes tudásmenedzsment együttműködésen alapuló és szemantikus technológiákkal

A közelmúlt új trendjei újradefiniálják a tudásmenedzsmentről alkotott képünket:
- Az ipari társadalomból tudásgazdaság lett.
- A szemantikus és tudástechnológiák gyorsan fejlődnek, és kiterjednek a 2.0-ás, valamint a 3.0-ás webre.
- A munka és a munkahelyi környezet virtuálissá válik.
- A hangsúly az alkalmazásalapú és információcentrikus architektúrákról és technológiákról a nyílt és tudásközpontú architektúrákra és technológiákra tevődik át.

A web 2.0 elősegíti a formába nem öntött, lektorálatlan, véletlenül felfedezett, vagy még nagyon nyers tudás feltűnésmentes összegyűjtését. Az együttműködésen alapuló új környezetekben sokkal könnyebb a tudás bármely formájának rögzítése életciklusának korai fázisában.

A szemantikus technológiákban benne rejlik annak a lehetősége, hogy a tudástranzakciók léptékét és körét megnöveljük. Lehetővé teszik, hogy több tudást és gyorsabban rögzítsünk és hasznosítsunk, továbbá, hogy kihasználjuk a gépi rendszerek képességeit.

Az új, virtuális környezet elősegíti az emberek közötti dinamikus kapcsolattartást. Ebben a környezetben tevékenységük túlmutat egy-egy intézmény információkezelési rendszerén vagy információtechnológiai infrastruktúráján, és tudásuk túlcsordul az eredetileg neki szánt, azt tároló alkalmazásokon.

Alapvető elmozdulást tapasztalhatunk a hagyományos információs infrastruktúráktól és információkezelési technológiáktól a tudásépítészeti és a tudásmenedzsment-technológiák irányába. Jelenlegi infrastruktúránk az 1980-as évektől a 2000-es évek elejéig kifejlesztett technológián alapul. Jól szolgált bennünket, azonban a szemantikus és tudásközpontú gondolkodás előtti szemléletet tükrözi, amely szerint az információkból csomagokat hozunk létre, tároljuk és elzárva tartjuk őket, ahelyett, hogy a tudásra úgy tekintenénk, mint ami dinamikus, folytonosan formálódik és szabadon áramlik.

A tudás összetett tárggyá válik. Ebben a kontextusban világossá lesz, hogy az emberek tudástárgyakká válnak, így kihívást jelent, hogy meghatározzuk, miként reprezentálhatjuk az embereket mint tudástárgyakat, továbbá megértsük, hogy miként kezelhető, tartható fenn, férhető hozzá, mozgósítható és tehető fogyasztásra alkalmassá az egyéni tudás.

Alapkérdések

A tudásmenedzsment érdeklődési köre eredetileg a világ, illetve a nemzetgazdaságok szintjére terjedt ki. Az elmúlt évtizedben ezt követte a közös-