



Az ETO alapú számítógépes információkeresés egy új megközelítése

PIROS Attila

Bevezetés

Az Egyetemes Tizedes Osztályozás egyike a világ legnagyobb és legelterjedtebb osztályozási rendszereinek. Különböző kiadásai legalább harminckilenc nyelvre lefordítva jelentek meg és a világ százhuszonnégy országában használják¹. Csak Európában több mint 140 000 könyvtár állományát indexelték vele,² és olyan nagy gyűjtemények feltárására használták, mint például a Magyar Országos Közös Katalógus (MOKKA) tagkönyvtárainak állománya, az Orosz Tudományos Akadémia Tudományos és Műszaki Információs Intézet (VINITI RAS), a Szlovák Nemzeti Könyvtár vagy a Svájci Könyvtárak és Információs Központok Hálózathoz (NEBIS) tartozó 140 könyvtár állománya.³ Ez a széles körű elterjedtség többek között annak

köszönhető, hogy a világ legelső analitikus-szintetikus információkereső nyelveinek egyikeként jelent meg több mint száz évvel ezelőtt, és azóta is folyamatos revízió alatt áll, ami biztosítja a benne foglalt ismeretek naprakészségét. Másrészt éppen az analitikus-szintetikus jelleg az, ami miatt a jelenlegi számítógépes információkereső rendszerek többnyire nem képesek az ETO lehetőségeinek maradéktalan kiaknázására.

A legfejlettebb ETO-alapú szoftveres megoldások alapelvei általában a következők: válasszuk ki azokat a jelzetelemeket, amelyeket önállóan visszakereshetővé szeretnénk tenni, a kontextusból kiemelve építsünk belőlük egy indexet, majd tegyük lehetővé a kereső számára, hogy azokat a Boole-féle logikai operátorokkal összekapcsolva posztkoordinált keresést végezzen.

Ez a megközelítés majdnem ötven éve került kidolgozásra, tökéletesen alkalmazkodva az akkori legmodernebb technológiákhoz.

A jelzetek ennél kifinomultabb elemzésén alapuló keresőrendszerekre azóta sem nagyon található működő példa (valójában még a fentiek szerint működőkre is alig), pedig az informatika változásai, a modern leíró és programozási nyelvek, az adatbázis-kezelés módszereinek a fejlődése ma már lehetővé tennék ezt – és az információkeresés minden korábbinál szélesebb körű elterjedése és hétköznapivá válása idején szükség is lenne – az ETO szerinti gépi keresés modern elvek szerinti újragondolására.

A jelenlegi megoldásokról

Jelenleg is számos kutatás folyik az ETO OPAC-okban való használatával kapcsolatban, illetve világszerte sok ETO-t használó kereskedelmi szoftver is létezik. Ezek többsége vagy sima tárgyszavakként kezeli a teljes ETO-jelzeteket, vagy legfeljebb arra ad lehetőséget, hogy azok egyes elemeit kiemelve indexet készíthessünk.

ETO-jelzetek kezelése tárgyszavakként

A legtöbb könyvtári rendszer a teljes ETO-jelzeteket – akár csak a tárgyszavakat mindenféle előzetes feldolgozás nélkül – egyszerű szövegmezőkben tárolja, általában helyettesítő karakterek használatát is lehetővé téve a kereséskor.

A csonkolás javíthatja az ETO lehetőségeinek kihasználását, azonban ha még ez sem adott, nem lehetséges hatékony, ETO-jelzetek szerinti keresés. Például tekintsük a következő problémát: ha a Magyarország történetével foglalkozó műveket szeretnénk összegyűjteni, a 943.9 osztályra történő keresés csak azokat a jelzeteket fogja visszahozni, amelyek ezt az egy számot tartalmazzák, anélkül, hogy akár másik főcímszám (művelettel), akár korlátozottan vagy általánosan közös (pl. idő) alosztás kapcsolódna hozzá, nem is beszélve a hierarchikus leszármazottakról. A szóba jöhető kapcsolódási lehetőségek száma – a sorrendi kötetlenségeket is figyelembe véve – lényegében végtelen, tehát nem lehet olyan keresőkérdést összeállítani, amellyel ne szenvednénk jelentős információvesztést.

De szerencsére már csak elvétve van olyan rendszer, amely ne engedné meg legalább az egy és a tetszőleges számú karaktert helyettesítő jelek használatát; bár a tetszőleges számú karaktert helyettesítő jel használata sok esetben csak korlátozottan lehetséges (pl. a keresett kifejezés végén, egy már eleve leszűkített találati halmazban stb.).

A helyettesítő karakterek használatának előnye, hogy nem csak az előzetes feldolgozás függvénye lesz, mi kereshető vissza, és mi nem. Más szóval, nem csak azon elemek alapján kereshetők vissza a jelzetek, amelyek indexelésekor – manuálisan vagy automatikusan – ki lettek emelve. Hátránya ugyanakkor az, hogy a keresést végző felhasználónak az ETO beható ismeretén kívül magát a szoftvert, annak lehetőségeit és korlátait is ismernie kell.

Ennél is nagyobb gond, hogy a fenti probléma részben itt is fennáll. Igaz, hogy a csonkolás által adott bővebb lehetőségeket felhasználva, de rá kell keresni az adott jelzet valamennyi előfordulási formájára, ami különösen a sorrend kötetlensége miatt okoz problémát. Az önálló általánosan közös alosztások sorrendjére vonatkozóan ugyanis csak ajánlás létezik, amelytől el is lehet térni. Valójában a főcímszámok és az önálló alosztások sorrendje tetszőleges, az utóbbiak akár az előbbiekre ékelve is elhelyezkedhetnek. Hasonlóan az összekötés és viszonyítás tényezőinek sorrendje is felcserélhető. Az adatvesztés elkerülése érdekében valamennyi lehetőséget számba kell venni.

A *UDC Summary Linked Databan* szereplő példa szerint az Amerikai Egyesült Államok külkereskedelmének jelzete a (73)339.5, a Francia Igazságügyi Minisztériumé a 354(44)51 formában is felírható⁴.

Eszerint, ha a Francia Igazságügyi Minisztériumra vonatkozó műveket szeretnénk összegyűjteni, a következő keresőkérdést kellene összeállítanunk:

* (44)*354.51* VAGY *354.51*(44)* VAGY
 3(44)54.51 VAGY *35(44)4.51* VAGY
 354(44).51 VAGY *354(44)51* VAGY
 354.5(44)1

És akkor még nem számoltunk azzal a lehetőséggel, hogy esetleg több alosztás is be van ágyazva a főtáblázati számba. Egy $3*(44)*54.51*$ keresőkérdés ugyanis már tekintélyes zajt is okozhatna (mivel a csillagok helyén számjegyek, illetve műveletek is szerepelhetnek).

Ha az egyszerűség kedvéért a kérdést $354.51*$ -ra változtatjuk, akkor pedig olyan művek is bekerülhetnek a találati listába, amik valójában nem érdekelnek bennünket, mert más ország minisztériumáról szólnak.

A kiterjesztések még komolyabb problémákat vetnek fel. Például az 519.7 (Matematikai kibernetika) számra keresve nem fogjuk visszakapni az 519.6/519.8-al (Numerikus matematika/Operációkutatás) jelölt műveket, amik pedig a fenti témakört is tartalmazzák (különböző összekötés szerepelne a jelzetben). Ha az ilyen jellegű információvesztést el szeretnénk kerülni, a keresett fogalmat tartalmazó, vele azonos hierarchiaszinten összeállítható összes intervallumot is hozzá kellene adni a kérdéshez. A fenti példa esetén ez az alábbi kifejezést eredményezné:

519.7 VAGY 519.1/519.8 VAGY 519.2/519.8 VAGY 519.6/519.8 VAGY 519.7/519.8 VAGY 519.1/519.7 VAGY 519.2/519.7 VAGY 519.6/519.7.

A kérdés hossza az adott helyen elérhető továbbosztások számától is függ, de az összes lehetőség számba vétele mindenképpen nagy terhet ró a keresőre.

A korlátozottan közös alosztások használata a kereséskor a kontextusra vonatkozó információ elvesztése miatt nem lehetséges. Nincs értelme annak, hogy csak egy speciális alosztás jelzetére keressünk, mivel az számos különböző jelentésű számhoz kapcsolódhat (pl. a -24 a 62-24 és 791-24 alatt is előfordulhat, egymástól teljesen különböző jelentéssel). Meg lehet próbálkozni a főtáblázati számmal együtt történő szerepeltetésével, pl. 511-37 (számítógép alkalmazása a számelméletben) formában, de akkor a főtáblázati szám hierarchikus leszármazottai fognak hiányozni az eredmények közül (pl. 511.3-37), nem is beszélve arról, hogy az ajánlás szerint az általános ismérvek szerinti alosztás is megelőzi a kötőjeles speciális alosztást (pl. 511-027.22-

37). Tehát ez a trükk információvesztéssel járhat. Ha a két jelzet közé csillagot teszünk, akkor pedig a zaj lesz nagy, hiszen nem biztos, hogy nincs művelet a főtáblázati szám és az alosztás között – ha van, akkor az alosztás másik fogalomra vonatkozik.

Valamivel jobb megoldást kaphatunk akkor, ha a rendszer egy dokumentumhoz több ETO-jelzet hozzárendelését is lehetővé teszi. Ebben az esetben ugyanannak a jelzetnek a különböző permutációi is rögzíthetők. Ahogy fentebb említettem, a jelzetszerkesztési szabályok nagy teret engednek az ilyen típusú számítógépes indexelésnek, azok hetvenes évek végi változtatását, a láncszabály eltörlését valószínűleg nem kis részben éppen ennek támogatása motiválta.

Helyettesítő karakterek nélkül az így épített index csak annyival jobb, mint a sima tárgyszavas rendszer, hogy valamivel nagyobb az esélye a jelzet pontos eltalálásának. Ha csonkolást is használunk, az invertált elemekre keresve jóval egyszerűbb kérdést is megfogalmazhatunk. A fenti példánál maradva, ilyen esetben elég a következő keresőkérdést használnunk, mivel valószínűsíthető, hogy a főtáblázati szám megjelenik kereső pozícióban:

$354.51*(44)*$

Ezzel a rövidebb keresőkérdés használata miatt esetlegesen fellépő információvesztést kiküszöböltük. Az intervallumok és a zaj problémáját azonban továbbra sem oldottuk meg. Itt is ugyanúgy lehetséges például, hogy műveletek vannak a csillagok helyén és az alosztás egy másik főtáblázati számhoz kapcsolódik.

Itt szeretném megjegyezni, hogy a többszörös indexelést szokás felhasználni arra is, hogy ugyanahhoz a dokumentumhoz több azonos vagy hasonló jelentésű jelzetet rendeljünk a szinonimák kezelése vagy a több szempontú visszakeresés megkönnyítése céljából.

KWOC-indexek használata

A KWOC-rendszerű indexelést ETO-számokra elsőként *Klaus Schneider* és *Karl-Heinz Koch* alkalmazta a hatvanas években. A megoldás lényege, hogy az önálló indexfogalomként al-

kalmazható jelzetelemeket a szövegkörnyezetből kiemelve indexet építünk belőlük, mely a kiemelt jelzetelem mellett a teljes jelzetet és a dokumentum azonosítóját tartalmazza.⁶

Például a

55(439)(035) Kézikönyv Magyarország geológiájáról

indextételei a következők lesznek:

55 - 55(439)(035) – dokumentum azonosító

(439) - 55(439)(035) – dokumentum azonosító

(035) - 55(439)(035) – dokumentum azonosító

A jelenlegi megoldások is a fenti elveken alapulnak, kisebb-nagyobb eltérésekkel. A jelzetelemek kiemelése történhet automatikusan, de manuálisan is, például az ETO-számok elemenkénti rögzítésével.

Az ilyen jellegű feldolgozás elsősorban a posztkoordinált keresésnek kedvez, amikor a felhasználó az egyes ETO-számokat a Boole-operátorokkal („és”, „vagy”) összekapcsolva állítja össze a keresése tárgyát.

Mivel a kiemelés független a sorrendtől, ezért azzal nem kell foglalkoznia, ez előny a korábbi megoldásokhoz képest. Mivel csak egyes jelzetek vannak kereső pozícióban, a rendszer kezelheti automatikusan (vagy helyettesítő karakter megadása esetén) a hierarchiát anélkül, hogy ez felesleges zajt okozna, mint például a 354.51*(44) keresőkérdés használatakor.

A speciális alosztások feldolgozását és visszakereshetőségét szokták megkérdőjelezni, de erre is lehet megoldás. *Mándy Gábor* le is írt – és meg is valósított – egy elegáns módszert erre.⁷ Ez is előnye az invertált indexnek a sima tárgyszavas tároláshoz képest.

Mindezek mellett hátrányai is vannak ezeknek a megoldásoknak.

Az intervallumok kezelése például nem oldható meg. De ennél is nagyobb gond, hogy a kontextus információ nem érvényesíthető a keresés során. Egyáltalán nem tudjuk, melyik jelzetelem melyikhez kapcsolódik, és azt sem azt, milyen művelettel és milyen sorrendben. Márpedig egy komplex jelzet esetén egyáltalán nem mindegy, melyik alosztás, melyik fő táblázati számhoz kapcsolódik, sorrendrögzítés esetén a két jelzet sorrendje is számít stb. Nem kezelhető továbbá

a csoportképzés sem, és az olyan árnyalatnyi különbségek is elvesznek, mint a nemzeti és liberális mozgalmak kapcsolata (329.17:329.12), illetve a nemzeti liberalizmus (329.17'12) közötti különbség.

Egy új megközelítés

A fentiek alternatívája lehetne egy olyan rendszer, amely a jelzeteket indexeléskor analizálja, feltárja azok belső struktúráját, és nemcsak az egyes jelzetelemeket emeli ki, hanem azok kapcsolódásait is rögzíti. Az így kapott reprezentációk kerülnének be az adatbázisba.

Így kereséskor nem Boole-műveletekkel összekötött táblázati számokra, hanem a keresőkérdést leíró komplex jelzetekre lehetne keresni, amelyeken a rendszer szintén elvégezné a fenti analízist majd az így kapott reprezentációt hasonlítaná össze a korábban elmentettekkel.

Ha a jelzetelemzés elég kifinomult, egy ilyen keresés könnyen hangolható akár az ETO-jelzetek kapcsolatainak elemzése, akár a kereső egyéni igényei alapján, hiszen a jelzetelekről minden adat a rendelkezésünkre áll.

A fenti rendszer három fő részből tevődik össze:

- a jelzetek elemzését végző program, mely a feldolgozást követően a memóriában előállít egy fa gráfot, amely tartalmazza a jelzet elemeit és azok kapcsolódásait;
- az adatbázis, amiben eltároljuk az így kapott hierarchikus reprezentációkat;
- legvégül a kereső program, amely a keresőkérdésként kapott jelzet elemzését követően annak eredményét az adatbázisban tárolt fákkal összevetve állítja össze a találati listát.

A továbbiakban egy jelzetelemző és reprezentáló rendszer logikai alapjainak a kidolgozására és szoftveres támogatására tesztek kísérletet.

Az ETO rendszere által felvetett kérdések

A fent felvázolt kereső rendszer központi eleme a jelzetértelmező modul. Ennek elkészítése szükséges előfeltétele a későbbi lépéseknek.

A szoftvert a lehető legtöbb esetre fel kell készíteni, hogy megfelelően értelmezze és kezelje a különböző alosztásokat.

Műveletek

Az ETO általánosan közös alosztásai két nagy csoportba sorolhatók: az elsőt a műveleti jelek, a másodikat az általánosan használható segéd-táblázatok alkotják.

Az általánosan közös alosztásokon kívül a nem önálló speciális alosztások között is található néhány, ami korlátozott hatókörű műveletként viselkedik.

Összekötés

Az összekötés a logikai „vagy”-nak, illetve a fogalmi kategóriák uniójának felel meg: több fogalom egymás melletti előfordulását reprezentálja. Ennek megfelelően az így alkotott jelzetek több részre bontva is rögzíthetők, a részjelzeteket külön jelzetként visszakereshetővé tevé⁵. Általában is ez a megoldás terjedt el.

A keresés szempontjából ez teljesen megfelelő. De ha az ETO-ra nemcsak könyvtári kereső, hanem ismeretrepresentáló eszközként is tekintünk, helyesebb lenne valamennyi műveletet megőrizni. Ahogy már korábban említettem, az egy dokumentumhoz több jelzet rögzítésének gyakorlatát több dolog is motiválhatja: a szinonimák által okozott adatvesztés elkerülése, egy jelzet permutált alakjainak rögzítése vagy az összekötés felbontása.

Ha egy szöveg bibliográfiai rekordjára tekintve több ETO-jelzetet látunk, nem egyértelmű, miért lettek mind rögzítve, ezért hiányos a dokumentumról szolgáltatott információ, nem tudjuk pontosan, hogy milyen formában tartalmazza a felsorolt témákat.

Ebből a szempontból nem tökéletes megoldás a jelzetek felbontása, és ha a rendszer megfelelően tudná kezelni a műveleteket, akkor nem is lenne rá szükség.

Kiterjesztés

A kiterjesztés több, azonos hierarchiaszinten, közvetlenül egymás mellett álló fogalom összekötésének a rövidített kifejezése. Más szóval, ETO-számok egy intervallumát jelenti. Pontosabban – mivel a kiterjesztéshez kapcsolódó alosztások nem tartozhatnak az intervallum egy-egy eleméhez, csak annak egészéhez – in-

kább egy szögletes zárójelben lévő több elemű összekötésről beszélhetünk.

Az elsődleges probléma vele kapcsolatban az, hogy az ETO-jelzetekre tárgyszavakként vagy mondatokként tekintő rendszerek nem tudják kezelni a rejtett, az intervallumok belsejében lévő jelzeteket. Ezért az ETO nyolcvanas évek eleji nagy átalakításakor még az is felmerült, hogy töröljék ezt a műveletet, és a szakértők jelentős része a mellőzésüket fogalmazta meg irányelv-ként⁵. A legújabb magyar nyelvű ETO-kiadás is inkább a számok egyesével történő rögzítését ajánlja, amennyiben minden számot szeretnénk visszakereshetővé tenni.⁸

Mindezt csak azért, mert a szoftverek nincsenek felkészítve egy probléma kezelésére. Sajnos ez példa arra, hogy az ETO szabályait és ajánlásait időnként a keresőprogramok hiányosságai miatt módosítják, korlátozva a lehetőségeit ahelyett, hogy a rendszereket írják meg úgy, hogy az ETO speciális jellegzetességeihez alkalmazkodjanak.

Ahogy fentebb írtam, a kiterjesztéshez kapcsolódó alosztások – beleértve a műveleteket, illetve az általános és a speciális alosztások osztályait – csak annak egészéhez tartozhatnak. Így a kiterjesztésre nem feltétlenül műveletként kell tekintenünk: úgy is megközelíthetjük a használatukat, hogy valamennyi egymáshoz kapcsolódó jelzetet intervallumként, egy-egy kezdő és záró táblázati számmal írunk le – amelyek persze lehetnek egyenlők.

Így kereséskor könnyedén meg tudjuk találni a keresett jelzetet tartalmazó intervallumokat és a vele megegyező táblázati számokat egyaránt.

Viszonyítás, sorrendrögzítés

A viszonyítás két fogalom kapcsolatát fejezi ki. Az ETO nem ad lehetőséget a tényezők viszonyának pontosítására, ezért ez a művelet a logikai „és”-nek, a két fogalmi kategória metszetének feleltethető meg. Ha viszont egy adott fogalom és a belőle viszonyítással kapott új fogalom kapcsolatát nézzük, az a logikai implikációnak felel meg (más szóval a metszet mindkét halmaznak részhalmaza).

A sorrendrögzítés egy olyan speciális viszonyítás, ami nemcsak a viszonyt, hanem annak irányát is rögzíti. Ebben az esetben a szoftvernek biztosítania kell a visszakereshetőséget mindkét tényező és a viszonyítás felől is, miközben a sorrendrögzítésre való kereséskor a tényezők sorrendjét is figyelembe kell vennie.

Csoportképzés

A csoportképzés (szögletes zárójel) a leginkább ellentmondásos művelet az ETO-n belül.

Eredeti funkciója a csoportosítás három vagy több jelzetelem összekapcsolása esetén. Emellett használható egy jelzetnek egy másikba való közbeékelésére, ha csak egy irányból szeretnénk biztosítani a visszakereshetőséget.⁵

Az ETO szabályainak különböző leírásaiban a szögletes zárójel használatának is eltérő meghatározásaival találkozhatunk.

A 2005. évi magyar nyelvű kiadás I. b) táblázatában például a következőt olvashatjuk:

„A csoportképzésére akkor lehet szükség, amikor a plusz jellel, törtvonallal vagy kettősponttal összekapcsolt ETO számokkal jelölt összetett tárgy (fogalom) mint egész kapcsolódik kettősponttal egy másik számhoz, vagy mint egészet módosítja egy általánosan vagy korlátozottan közös alosztás.”⁸

A korlátozottan közös alosztás csak abban az esetben értelmezhető, ha a zárójel több, művelettel összekapcsolt, a táblázatban egymáshoz közel található számot tartalmaz, az alosztás pedig mindegyikkel használható együtt. A leírás abból a szempontból is félrevezető, hogy első olvasásra úgy tűnik, mintha a csoportképzés csak a kettősponttal kapcsolódhatna egy másik jelzethez. Pedig abban az esetben, amikor „mint egészet” módosítja egy alosztás, az így kapott jelzet miért ne kapcsolódhatna összekötéssel egy másikhoz?

Az 1990. évi magyar kiadásban található leírás szerint még az alosztások sem szükségesek ehhez.¹⁴ Az ottani példa arra világít rá, hogy ha csoportképzést használunk, az összekötés a viszonyításra nézve disztributív. Eszerint a 331.31:[622+629] Munkaidő a bányászatban és kohászatban

viszonyítás az összekötést elemeire bontva is felírható

[331.31:622]+[331.31:629]

Itt a zárójeleket már akár el is hagyhatnánk.

Ha az eredeti szögletes zárójelben lévő összekötés három tényezős, akkor a felbontás után az összekötésnek is három tényezője lesz. Ha a viszonyítás első tagja is egy zárójelben lévő összekötés lenne, akkor már hat elemű összekötést kapnánk.

A klasszikus eljárás az alosztások esetén is hasonló, ugyanígy lehet felbontani a csoportképzést, minden eleméhez hozzákapcsolva valamennyi alosztást.

Így egy sor új jelzetet kapunk, amelyeket mind rögzíteni kell, illetve az egyes elemeiket visszakereshetővé is kell tenni. Ennek a rengeteg kifejezésnek a kezelése helyett célszerűbbnek tűnik a jelzet szerkezetét a csoportképzéssel és az eredeti kapcsolódásokkal együtt rögzíteni, mind indexeléskor mind kereséskor. Minden további megfeleltetés a keresőalgorithmus dolga.

Szintézis

A jelzetszintézis 1983-as revíziója elvben lehetővé tette az így összekapcsolt jelzetek felőli visszakereshetőséget is⁵. Ezzel a táblázat bizonyos helyein, pl. a 329.1/.6 (politikai pártok és mozgalmak) vagy 546 (szervetlen kémia) alosztályokban az aposztrófós speciális alosztás a sima viszonyításhoz hasonló, kommutatív, de korlátozott hatókörű műveletként viselkedik. Tehát a programnak fel kell ismernie annak elemeit, és biztosítani kell a visszakereshetőségüket.

Ugyanakkor arra is figyelni kell, hogy a művelet-jelleg mellett a szintetizált számok a fő táblázati számokhoz hasonlóan viselkednek: általánosan és korlátozottan közös alosztások kapcsolódhatnak hozzájuk, amelyek nem az egyes operandusok, hanem az összekapcsolással kapott teljes jelzet jelentését specifikálják.

Hierarchikus kapcsolatok

A fenti megközelítésben a hierarchikus kapcsolatok is a logikai implikáció megfelelői, tehát azt fejezik ki, hogy egyik kategória a másiknak

részhalmaza. Az már más kérdés, hogy az ETO szerinti leszámazás mennyire valós, azaz mennyire igaz a konkrét fogalmak esetén, hogy egy osztály továbbosztása annak tényleges fogalmi szűkítése.

Például a nullás vagy a hatos főosztály kis túlzással máshova be nem sorolható témák gyűjteménye. Ezért nehéz elképzelni, hogy valaki a „0” keresőkérdést használja azért, hogy a bibliográfiákat, az általános műveket és a programozási szakkönyveket egyetlen kereséssel visszkapja (vagy a „6” kérdést azért, mert az orvostudományi, technikai és mezőgazdasági művekre egyaránt kíváncsi).

Főosztálytól, illetve az alosztás típusától függően a második vagy a harmadik decimális számjegy az, amitől kezdődően már valósnak, akár a keresőszoftver által automatizálhatónak is tűnik a hierarchia (ennek pontos megállapítása azonban további kutatást igényelne). Ugyanakkor a UDC Summary adatainál is azt láthatjuk, hogy nem minden esetben használja automatikusan a decimális kapcsolatokat a fogalmak szűkítésekor. Például a 018 felső kapcsolata nem a 01, hanem a 017/019, az =84-é nem az =8, hanem az =84/=88 stb.⁹

A részjelzetek azonosítása

A felismerés problémái

A jelzet részeinek automatikus felismerése nem olyan könnyű programozási feladat, mint amilyennek elsőre tűnik, mégpedig elsősorban az ETO szabályai miatt. Például vannak olyan karakterek, amik metanyelvi és nyelvi jelekként egyaránt megjelennek (például a .0-ás vagy a korábbi .00-ás alosztások esetén); a korábbi ETO-verziókban még szerepeltek a számjelzetes alosztások, amiket szinte lehetetlen beazonosítani; ugyanaz a karaktorsor a táblázat egyes helyein korlátozottan közös, máshol nem önálló általánosan közös alosztást vagy akár műveletet jelent; a földrajzi alosztásokban vannak sehol máshol elő nem forduló speciális alosztások, illetve jelzetelési lehetőségek stb.

Az ilyen jellegű problémákra most nem szeretnék részletesen kitérni, de azok jelentős része

fel van sorolva Mándy Gábor tanulmányában¹⁰. Szintén említésre méltók a szerző ezek alapján készült szoftveres jelzetelemző megoldásai, amelyek publikálva is lettek⁷, illetve online is elérhetők¹¹.

A jelzetváltozatok kezelése

A szabályok pontatlansága által okozott problémák mellett megemlítendő még az ETO jelzetváltozásaiból adódóak is, amelyekről *Prokné Palik Mária* beszél a tartalmi feltárás problémáiról szóló tanulmányában¹². Mint említi, ezzel a problémával a retrospektív adatbevitel kapcsán bármely könyvtár találkozhat. Mándy Gábor tanulmányához írt megjegyzésében a megoldást az utalók használatában látja, melyekkel a keresés során követhető lenne az, hogy az esetlegesen megszűnt számokat milyen újabbak váltották fel¹³. Ez azokban az esetekben megfelelő, amikor egyes osztályok a jelentés lényegi megváltozása nélkül kerülnek áthelyezésre a táblázat egy másik helyére vagy több, hasonló jelentésű jelzet lesz kiváltva eggyel, de nem elegendő akkor, ha egy terület teljes hierarchiája átszervezésre kerül új szempontok szerint, vagy amikor maguk a jelzetkészítési szabályok változnak.

A fentiek alapján a szoftvernek a jelzeteket mindig az azok elkészítésekor érvényben lévő szabályok szerint kell értelmeznie. A szabályok változása ezért azt jelenti, hogy a programnak újabb esetekre kell felkészülnie, ami újabb programozási munkával is jár. Ha a változás olyan jellegű, hogy az új szabályok szerint a régi jelzetek nem értelmezhetők, akkor az utalók is használhatatlanná válnak, ez pedig adatvesztéshez vezethet. Szintén problémát okoz, hogy a régebben rögzített jelzetek esetén nem mindig egyértelmű, melyik ETO-kiadás alapján lettek összeállítva.

Ezért, bár csábítónak tűnhet, hogy az osztályozási rendszert magát, illetve annak szabályait szervezzük át a számítógépesítés megkönnyítése érdekében, ez a programozók számára valójában inkább nehézséget jelent. Egy komplex, de jól definiált és lényegében változatlan rendszert könnyebben lehet kezelni, mint egy egyszerűbb-

nek tűnőt, ami nem egyértelmű és rendszeresen változik.

Az analógiák kezelése

Az ETO egyes osztályai között analógiás kapcsolat van. Például a nyelvi alosztások jelzeteiből származtathatjuk az egyes nyelvek nyelvtanát és irodalmát leíró számokat valamint az etnikai alosztásokat, illetve – a régi ETO-ban – a földrajzi alosztásból az egyes országok, kontinensek stb. földrajzát és történelmét. Az ilyen jellegű kapcsolatokról szintén Mándy Gábor tanulmányában olvashatunk részletesebben¹⁰.

A rendszer ezeket a kapcsolatokat is követheti a keresés során, a jelzet felbontása után. Ezt azonban érdemes opcionálissá tenni, a felhasználóra bízva, kíván-e élni ezzel a lehetőséggel.

Az ETO és a szemantikus web

A szemantikus web célja a weben található tartalmak szemantikai információk alapján történő összekötése (csoportosítása, egyértelművé és kereshetővé tétele stb.) a számítógépek által is követhetően. A tartalmak szemantikai jellemzőinek és az azok közötti kapcsolatok leírásának alapvető eszközei az ontológiák, melyek egy-egy tudásterület fogalmait és azok összefüggéseit írják le.

A World Wide Web Consortium kidolgozott egy ajánlást arra vonatkozóan is, hogyan lehetne a már meglévő tudásszervező rendszereket (tezauszok, osztályozási rendszerek, tárgyszólisták stb.) olyan formában leírni, hogy azok használhatóak legyenek a szemantikus web keretein belül. Ez az ajánlás a SKOS (Simple Knowledge Organization System) nevet kapta¹⁵.

Mivel az ETO egy része már 2009 óta hozzáférhető SKOS formátumban¹⁶, az egyszerű jelzetek egy része elvileg eddig is felhasználható volt a szemantikus web építéséhez.

Ez kiterjeszhető az összetett jelzetekre is, amennyiben megfelelően fel tudjuk bontani őket (mivel az osztályok mellett a műveletek is rendelkeznek saját azonosítóval).

Ez a lehetőség azonban nem áll fenn az ETO egészére, mivel csak a – reprezentációs és ok-

tatási célból létrehozott¹⁷ – UDC Summary által tartalmazott mintegy 2600 osztály érhető el a fenti módon (a Creative Commons Share Alike 3.0 licenc szerint).

A megoldás bemutatása példákon keresztül

Az alábbiakban a fentebb említett jelzetanalízis lehetséges eredményeire hozok – a 2005. évi magyar nyelvű nyomtatott ETO-kiadás⁸ segítségével összeállított – példákat.

Az ismertett reprezentációk az általam fejlesztett jelzetértelmező szoftver kimenetéről származnak, melyet abból a célból készítettem el, hogy demonstráljam vele a fent említett hierarchikus megközelítés lényegét, annak egy lehetséges megvalósítását, illetve azt, hogy valamennyi ETO-jelzet leírható ilyen módon.

A program bemenetként egy tetszőleges ETO-számot vár (a leírásával és az elkészítéséhez használt táblázat évszámával együtt), melyet további felhasználói beavatkozás nélkül dolgoz fel. A folyamat során jelzi az értelmezést akadályozó esetleges szintaktikai hibákat, és igyekszik figyelembe venni az ETO szabályainak olyan módosulásait is, mint a szempont szerinti általánosan közös, illetve a számvégződéses korlátozottan közös alosztások eltörlése, vagy a tulajdonságok szerinti általánosan közös alosztások bevezetése (az elmúlt húsz év fő változásainak felsorolása elérhető a UDC Consortium honlapján¹⁸).

Mivel a szoftver az ETO-jelzeteket kizárólag szintaktikai módszerekkel, a metanyelvi jelek azonosításával elemzi, semmilyen formában nem tárol ETO-számokat vagy táblázatokat.

A feldolgozás eredményeként a program a memóriában előállított fa struktúrát egy HTML-oldalon jeleníti meg. A fentebb vázolt rendszer megvalósításának második lépése ezeknek a struktúráknak az adatbázisba mentése.

1. példa

Az első jelzet a táblázatokban is szerepel a vegyületek indexelésének példájaként az 546 (szervetlen kémia) alatt⁸:

546.33'185-384.1 *Nátrium-dihidrogén-ortofoszfát (Na₂PO₄)*

A feldolgozás eredménye:

Jelzet: 546.33'185-384.1

ETO-verzió: 2005

leírás (magyar): Nátrium-dihidrogén-ortofoszfát (Na₂PO₄)

↳ *szintézis*

↳ *főtáblázati szám, jelzet: 546.33*

↳ *főtáblázati szám, jelzet: 546.185*

↳ *korlátozottan közös alosztás, jelzet: -384.1*

A fa gyökere tartalmazza a jelzetet, az ETO-verziót, a jelzet leírását és annak nyelvét. Leszármazottai tartalmazzák a feldolgozás tényleges eredményét.

A jelzet egyetlen szintézist tartalmaz az 546.33 (Nátrium) és az 546.185 (foszfátok) főtáblázati számok között, melyhez egy korlátozottan közös alosztás tartozik. Tehát az előállított fa gyökeréhez is egy szintézis kapcsolódik, melynek leszármazottai az általa összekötött főtáblázati számok és a hozzá kapcsolódó korlátozottan közös alosztás.

Mivel a korlátozottan közös alosztás nem a két szám egyikéhez, hanem a szintézishez kapcsolódik, biztosított a visszakereshetőség az 546.33 és a 546.185 valamint hierarchikus őseik, illetve az 54-384.1 felől egyaránt.

2. példa

510.6+510.3(075.8)=161.1 *Matematikai logika és halmazelmélet, egyetemi jegyzet orosz nyelven*

A feldolgozás eredménye:

Jelzet: 510.6+510.22(075.8)=161.1

ETO-verzió: 2005

leírás (magyar): Matematikai logika és halmazelmélet, egyetemi tankönyv orosz nyelven

↳ *összekötés*

↳ *főtáblázati szám, jelzet: 510.6*

↳ *főtáblázati szám, jelzet: 510.22*

↳ *formai általánosan közös alosztás, jelzet: 075.8*

↳ *nyelvi általánosan közös alosztás, jelzet: =161.1*

A jelzet egy összekötésből és két általánosan közös alosztásból áll.

Látható, hogy a formai és nyelvi alosztások nem az összekötéshez vagy annak elemeihez, hanem a dokumentum egészéhez, tehát a fa gyökeréhez csatlakoznak.

3. példa

A következő példa egy kiterjesztést tartalmaz. Csoportosítási okokból a földrajzi alosztás a főtáblázati szám közepén található.

622(437.1)333/336-022.316 *A csehországi szén- és grafitbányászat fenntarthatósága*

Mivel a -02 (tulajdonságok szerinti) általánosan közös alosztások 1999-ben lettek bevezetve ¹⁸. ha korábbi évszámot választunk, a program hibáüzenet ír ki („A jelzet érvénytelen a megadott ETO verzióban”).

Ha 1999-et vagy annál későbbi évszámot választunk, a következő eredményt kapjuk:

Jelzet: 622(437.1)333/.336-022.316

ETO-verzió: 1999

leírás (magyar): A csehországi szén- és grafitbányászat fenntarthatósága

↳ *főtáblázati szám, jelzet 1 / jelzet 2:* 622.333 / 622.336

↳ *földrajzi általánosan közös alosztás, jelzet:* 437.1

↳ *tulajdonságok szerinti általánosan közös alosztás, jelzet:* -022.316

A főtáblázati szám tartalmazza a kiterjesztés elejét és végét, így a kereséskor a köztük lévő számok is elérhetőek lesznek. Azt is láthatjuk, hogy az általánosan közös alosztások az intervallumhoz kapcsolódnak annak elemei helyett.

4. példa

Az alábbi, kicsit bonyolultabb jelzet a 2013-as ETO-szeminárium egyik előadásán elhangzott példából kiindulva lett megszerkesztve a csoportképzés kezelésének bemutatására²⁰.

394.4:[929(439):329(439).17'11]

Hivatalos ünnepek a magyarországi nemzeti konzervatívizmushoz kapcsolódó személyiségek részvételével.

A feldolgozás eredménye:

Jelzet: 394.4:[929(439):329(439).17'11]

ETO-verzió: 2000

leírás (magyar): Hivatalos ünnepek a magyarországi nemzeti konzervatívizmushoz kapcsolódó személyiségek részvételével.

↳ *viszonyítás*

↳ *főtáblázati szám, jelzet:* 394.4

↳ *csoportképzés*

↳ *viszonyítás*

↳ *főtáblázati szám, jelzet:* 929

↳ *földrajzi általánosan közös alosztás, jelzet:* 439

↳ *szintézis*

↳ *főtáblázati szám, jelzet:* 329.17

↳ *főtáblázati szám, jelzet:* 329.11

↳ *földrajzi alosztás, jelzet:* 439

Látható, hogy a jelzet egyetlen viszonyítást tartalmaz, amely egy főtáblázati számot és egy csoportképzést köt össze. A csoportképzésen belül egy újabb viszonyítás áll, melynek operandusai egy főtáblázati szám és egy szintézis, a hozzájuk kapcsolódó alosztásokkal.

5. példa

Az utolsó jelzet a sorrendrögzítésre, idő alosztásra és az A/Z névalosztásra egyaránt példát nyújt.

929.:78(430)"16/17" Bach A Bach-család

A feldolgozás eredménye:

Jelzet: 929.:78(430)"16/17" Bach

ETO-verzió: 2005

leírás (magyar): A Bach-család

↳ *sorrendrögzítés*

↳ *főtáblázati szám, jelzet:* 929, sorszám: 1

↳ *főtáblázati szám, jelzet:* 78, sorszám: 2

↳ *névalosztás, jelzet:* Bach

↳ *idő szerinti alosztás, jelzet 1 / jelzet 2:* 16 / 17

↳ *földrajzi alosztás, jelzet:* 430

A sorrendrögzítés miatt szükséges a főtáblázati számok sorrendjének az elmentése. A név- és időalosztás feldolgozása hasonló a földrajzi és formai alosztásokéhoz.

Befejezés

Barátné Hajdú Ágnes az ETO dinamizmusát tárgyalva kiemeli, hogy a könyvtárosok identitásának eleme az ETO (és a TO), mely a tudást, a gondolkodást, a visszakeresést, a könyvtár szellemiségét jelképezi¹⁷. Ugyanakkor jelentősége túlmutat a könyvtári információkezelésen.

Ungváry Rudolf arra mutat rá, hogy az ETO egy ontológia, „*a mai ontológiák őse, s egyben hatalmas tárháza a fogalmi tudásnak*”¹⁹. Valójában az ETO a fogalmi tudást, a sokrétűséget tekintve talán még többet is nyújt a mai ontológiáknál.

A mesterséges nyelv által biztosított nyelvfüggelenséget is figyelembe véve talán nem túlzás kijelenteni, hogy a megfelelő szoftverek megléte esetén az ETO továbbra is az egyik legjobban használható ismeretrepresentáló és -kereső eszköz lehetne. Ebbe az irányba lehet egy nagy lépés a fentiekhez hasonló elvek szerinti formális nyelvi értelmező- és keresőrendszerek kidolgozása. Ennek érdekében készítettem el a jelzértérmező szoftver prototípusát, amelyet reményeim szerint minél előbb hozzáférhetővé fogok tenni a könyvtáros közösség és az érdeklődő jövőbeli felhasználók számára.

Jegyzetek

1. SLAVIC, Aida: Use of the Universal Decimal Classification: A world-wide survey = *Journal of Documentation*, 64. vol. 2008. 2. no. 211–228. p.
2. SLAVIC, Aida: UDC libraries in the world – 2012 study = *Universal Decimal Classification Blog*, August 20 <http://universaldecimalclassification.blogspot.hu/2012/08/udc-libraries-in-world-2012-study.html> [2014. március 1.]
3. Collections indexed by UDC <http://udcc.org/index.php/site/page?view=collections> [2014. március 1.]
4. UDC Summary Linked Data: Common auxiliaries of place. Table 1e <http://udccdata.info/001951> [2014. március 1.]
5. BABICZKY Béla: Szemléletváltozás az ETO jelzetszerkesztésében = *Könyvtári Figyelő*, 31. évf. 1985. 1. sz. 17–27. p.
6. OROSZ Gábor: Az ETO alkalmazása a gépesített dokumentációs és könyvtári munkában. In: Babiczky Béla: Szöveggyűj-

temény az osztályozás és indexelés kérdéseinek tanulmányozásához. – Bp.: Tankönyvkiad., 1988. 241–242. p.

7. MÁNDY Gábor: A posztkoordináció esélyei az ETO-ban = *Könyvtári Figyelő*, 23. (59.) évf. 2013. 1. sz. 65–84. p. <http://ki.oszk.hu/kf/2013/04/a-posztkoordinacio-eselyei-az-eto-ban/> [2014. március 1.]
8. Egyetemes Tizedes Osztályozás : UDC Publ. No. P057 / [szerk. és bev. Barátné Hajdu Ágnes] ; [közrem. Ackermanné Kelő Kamilla et al.] ; [közread. az] Országos Széchényi Könyvtár Könyvtári Intézet. – Budapest : KI, 2005.
9. UDC Summary <http://www.udcc.org/udccsummary/php/index.php> [2014. március 1.]
10. MÁNDY Gábor: Régi ETO, új ETO = *Könyvtári Figyelő*, 21. (57.) évf. 2011. 2. sz. 317–322. p. <http://ki.oszk.hu/kf/2011/07/regi-eto-uj-eto/> [2014. március 1.]
11. MÁNDY Gábor: Online mintaprogramok PHP nyelven <http://mandynamics.hostzi.com/php-oldal.html> [2014. március 1.]
12. PROKNÉ PALIK Mária: A tartalmi feltárás problémái online könyvtári katalógusokban = *Tudományos és műszaki tájékoztatás*, 52. évf. 2005. 11–12. sz. http://tmt.omikk.bme.hu/show_news.html?id=4243&issue_id=467 [2014. március 1.]
13. PROKNÉ PALIK Mária: Észrevételek Mándy Gábor tanulmányához = *Könyvtári Figyelő*, 21. (57.) évf. 2011. 2. sz. 317–322. p. <http://ki.oszk.hu/kf/2011/07/eszrevetelek-mandy-gabor-tanulmanyahoz/> [2014. március 1.]
14. Egyetemes Tizedes Osztályozás : FID Publ. No. 691 / [szerk. Babiczky Béla, Schneller Károly] ; [közread. az] OMIKK, OSZK–KMK; [közrem. Cséffalvy Olga et al.]. – Röv. kiad. – Bp. : OMIKK, OSZK–KMK, 1990–1991.
15. W3C: Introduction to SKOS <http://www.w3.org/2004/02/skos/intro> [2014. március 1.]
16. UDC Summary Linked Data <http://udccdata.info/> [2014. március 1.]
17. BARÁTNÉ HAJDU Ágnes: Az ismeretszervezés dinamizmus, különös tekintettel az Egyetemes Tizedes Osztályozásra = *Könyv és nevelés*, 13. évf. 2011. 4. sz. mell. http://olvasas.opkm.hu/portal/felso_menusor/konyv_es_neveles/az_ismeretszervezes_dinamizmus_a_kulonos_tekintettel_az_egyetemes_tizedes_osztalyozasra [2014. március 1.]
18. Major changes to the UDC 1993-2013 http://udcc.org/index.php/site/page?view=major_revisions [2014. március 1.]
19. UNGVÁRY Rudolf: Észrevételek egy ETO-ról szóló tanulmányhoz = *Könyvtári Figyelő*, 21. (57.) évf. 2011. 2. sz. 317–322. p. <http://ki.oszk.hu/kf/2011/07/eszrevetelek-mandy-gabor-tanulmanyahoz/#more-1914> [2014. március 1.]
20. SMIRAGLIA, Richard et. al.: UDC in Action http://www.udccs.com/seminar/2013/media/slides/rsmiraglia_udccseminar2013.pdf [16 dia] [2014. március 1.]

Beérkezett: 2014. március 3.