Piros Attila

A gondolat a szimbólum mögött

Az ETO-jelzetek automatikus interpretálásáról és reprezentációjáról

Az Egyetemes Tizedes Osztályozás (ETO) és más analitikus-szintetikus és fazettás osztályozási rendszerek számos lehetőséget nyújtanak a tárgykörök szintaktikai relációk segítségével történő kifejezésére. Az ilyen esetekben a relevancia eldöntése az információkeresés során szükségessé teszi a jelentés lehető legpontosabb megállapítását a jelzet alapján, melynek alapvető feltétele a prekoordinált jelzetek struktúrájának elemzése. Központi kérdés, hogy ez a jelentésfelismerés mennyiben támogatható automatikus eszközökkel, főleg azért, mert számos bibliográfiai forrás tartalmaz ilyen ETO-jelzeteket egyszerű szövegként eltárolva, ezek alapján pedig hatékony információkeresés csak nagyon körülményesen végezhető. Cikkemben az ETO-jelzetek platformfüggetlen, automatikusan feldolgozható, a teljes szintaktikai struktúrát megtartó formátumban való reprezentációjára irányuló kutatásom jelenlegi állását ismertetem. Az ETO-jelzeteket közvetlenül a tervezett formátumra átalakító algoritmus és szoftver is fejlesztési stádiumban van, továbbá a kutatás céljai között szerepel olyan konverziós algoritmusok kidolgozása és elérhetővé tétele, amelyek más szoftverek által is könnyen feldolgozhatók. Az írás elsősorban a kutatás elmúlt két évének eredményeit, illetve a jövőbeli terveket és irányokat igyekszik bemutatni.[1]

Tárgyszavak: *Egyetemes Tizedes Osztályozás; ETO-jelzet; elemzés*

Bevezetés

Híres művükben *Charles Kay Ogden* és *Ivor Armstrong Richards* a lingvisztikai szimbólumok és az általuk reprezentált objektumok kapcsolatát egy háromszögként modellezték, melynek csúcsaiban az objektum vagy referens, a róla alkotott gondolat és annak jelölése (szimbólum) állnak. [3]

Bibliográfiai metaadatok esetén, melyekben a tárgy az Egyetemes Tizedes Osztályozás (ETO) vagy valamely más osztályozási rendszer segítségével van kifejezve, az objektum a leírás tárgyát jelenti: valamely dokumentumot, szöveget, képet, műalkotást vagy bármilyen más indexelt objektumot. A gondolat az objektum fő tárgyköreinek rezüméje, egy indexelő által egyszerű állítások formájában kifejezve. Legvégül a szimbólum ezeknek az állításoknak egy osztályozási rendszer egyszerű vagy prekoordinált jelzeteire lefordított formája.

Az ETO és más analitikus-szintetikus osztályozási rendszerek a bonyolultabb tárgyköröket jelzetek építésével fejezik ki. Ilyen esetekben a hatékony relevanciadöntések, illetve az objektumok pontos beazonosítása szükségessé teszik a gondolat minél precízebb meghatározását a szimbólum alapján. Az információkeresésben központi kérdés az ilyen jellegű döntések automatikus eszközökkel való támogatása. Ehhez egyebek mellett a jelzetek struktúrájának az analízisére van szükség.

A cikkben bemutatott kutatás célja egy olyan új, platformfüggetlen formátum kifejlesztése, amely az ETO-jelzetek teljes szintaktikai struktúráját leírja, így támogatva azok további automatikus feldolgozását. A kutatás céljai között szerepel egy olyan algoritmus kidolgozása és megvalósítása is, amely az ETO-jelzeteket közvetlenül az említett formátumra alakítja, illetve olyan konverziós metódusok kialakítása és online elérhetővé tétele, melyek képesek őket további, más szoftverek által olvasható formátumokba átalakítani.

[1] A cikk a „Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar" 2017, London (UK), 14-15 September. Würzburg : Ergon Verlag, 2017. Nemzetközi ETO Szeminárium 2017 konferenciakötetben megjelent cikk magyar nyelvű verziója. [1] A konferenciáról a *Könyv, Könyvtár, Könyvtáros 2017* novemberi száma közölt részletes beszámolót. [2]

A további kutatási tervek elsősorban a kidolgozott megoldások lehetséges felhasználási területeire fókuszálnak.

Az ETO mint analitikus-szintetikus osztályozási rendszer

Azokat a fogalmakat, amelyeket alaptárgyak fazettákkal való pontosításával nyerünk, − *Ranganathan* után − összetett tárgyaknak nevezzük. [4] Az ETO-ban kiemelt jelentősége van azoknak a fazettakombinációknak, melyek általánosan közös alosztásokként (a hely, idő, formai megjelenés, nyelv, anyag, személyi vonatkozások, tulajdonságok, relációk stb. jelölésére) jelennek meg. Egyébként a fazetták, mint olyan attribútumok, amelyek tipikusan egy adott osztályon belül fordulnak elő, a leggyakrabban korlátozottan közös alosztásokkal vannak kifejezve. [5]

A komplex tárgykörök azok a fogalmak, melyekben kettő vagy több tárgy valamilyen köztük lévő kapcsolat alapján állnak össze egységes egésszé. [4] Az ETO-ban komplex tárgyköröket az egyszerű viszonyítás (:), sorrendrögzítés (::)[2] és csoportosítás használatával alkothatunk.

Az agglomerált alap tárgykör fogalmát *Neelameghan* vezette be azokra a fogalmakra, amelyek „egységeket nagyobb halmazokba vonnak össze a részek kohéziója nélkül”. [6] Az ETO-ban az összeadás (+) és a kiterjesztés (/) alkalmazható az agglomeráció két, Neelameghan által meghatározott fajtájának a kifejezésére.

Az ETO analitikus-szintetikus osztályozási rendszer: a mély fogalmi hierarchia mellett nemcsak a tudásterületek fazettáinak a kifejtésére kínál számos megoldást, de a komplex és az agglomerált alaptárgykörök kifejezésére is. [7][8]

Az analitikus-szintetikus osztályozások esetében a szintaktikai relációk által hordozott információ jelentőségét át kell értékelnünk az online visszakeresés során. Nemcsak a facetták értelmezhetetlenek sok esetben a bázisosztályuk ismerete nélkül, de fontos különbséget kell tennünk az összetett, komplex és agglomerált tárgykörök között is. Emellett a reláció fázisának, akár a jelzetelemek hivatkozási sorrendjének a figyelembe vétele is szükséges lehet. [9] Ezek az adatok növelhetik a rendszer hatékonyságát, különösen a precízió szintjét a tárgyi böngészés és keresés során. Bár a hierarchikus struktúra ideális feltételeket teremt az inkluzív keresés számára [10], ehhez is szükséges a prekoordinált jelzetek elemeinek és relációinak pontos beazonosítása.

Az ETO-számok interpretációja

Az ETO komplex jellege és a fejlesztésének százéves története során felfedezhető inkonzisztencia megnehezíti az osztályozás gépesítését. Az utóbbi harminc év átszervezései az osztályozás teljesen fazettás formában történő átszervezését célozzák a fazettaanalízis elvének konzisztens alkalmazása alapján, olyan teljesen fazettás osztályozásokhoz hasonló módon, mint a BC2. [11][12][13][14][15] Az átszervezések másik mozgató elve az a feltételezés, hogy a táblázatok szisztematikusabb struktúrája következetesebb jelzetelést, ennek következtében jobb jelzetkezelést eredményez online környezetben is. [16]

A szakmai konszenzus szerint, az authority control kielégítő módja az osztályozás könyvtári OPAC-okban és tárgyi keresőkben való alkalmazásának: a 2015. évi ETO Szeminárium teljes egészében ezt a témát igyekezett körbejárni. [17] Ugyanakkor az authority control költséges eljárás, melyet a lehetőségek szerint automatizáltan is támogatni kell. Az is tény, hogy az authority control gyakran nem elérhető, ami az osztályozás alkalmazását akár teljesen ellehetetlenítheti. Végül, de nem utolsósorban az autorithy control alkalmazása bonyolult és nehézkes, ha az analitikus-szintetikus osztályozással kifejezett szintagmatikus fogalmak szintaktikai relációit is figyelembe szeretnénk venni [18], pedig az ilyen fogalmak kifejezhetősége az egyik legnagyobb előnye ezeknek a rendszereknek.

Az ETO-jelzetek automatikus felbontásának és az elemeik alapján történő indexépítésnek a lehetőségét az 1960-as években kezdték vizsgálni. [19] Az 1990-es években *Gerhard Riesthuis* fejlesztett és publikált olyan algoritmusokat és példaalkalmazásokat, melyek képesek voltak a jelzetrészek azonosítására. Riesthuis fő célja a jelzetek felbontásával a jelzetelemeknek az ETO mesterfájlból (UDC MRF) származó természetes nyelvi leírásokhoz való hozzárendelése, ezzel a természetes nyelvű visszakereshetőség biztosítása volt. [20][21] Eredményeit doktori disszertációjában összegezte, amely a mai napig a legátfogóbb munkának számít ebben a témakörben. [22]

Magyarországon *Mándy Gábor* végzett hasonló jellegű kutatásokat, illetve adott közre olyan (PHP nyelven írt) példaprogramokat, melyek képesek

egyes szintaktikai relációk és alosztások felismerésére. Az elképzelésének alapja egy olyan programcsomag kidolgozása, melynek algoritmusai lépésenként, egyfajta „szűrőként" viselkedve − minden eljárás az őt megelőző kimenetét kapva bemenetként − bontják részeire az ETO-jelzetet. Célja az ETO-jelzetek posztkoordinált használatának elősegítése volt azzal, hogy kész algoritmusokat nyújt a szoftverek fejlesztőinek. [23]

Az összetett, komplex és agglomerált tárgykörök részeinek felismerése lehetséges az MRF-ből származó jelölők alapján is, ha a jelzet tartalmazza ezeket. Ennek a megközelítésnek a másik előnye, hogy a jelölők a jelzetek szabályos rendezését is megkönnyítik. Hátránya, hogy feltételezi az MRF használatát a jelzetszerkesztés során.

Ahogy láthattuk, a korábbi kutatások elsősorban a prekoordinált számok elemeinek a felismerésére összpontosítottak: ugyanakkor ez a posztkoordinált szemlélet tovább fejleszthető, ha figyelembe vesszük a kompozícionalitás elvét [24], azaz azt, hogy a jelzetek jelentését az összekapcsolt ETO-számok és azok kapcsolatai együttesen határozzák meg.

Az ETO-jelzetek kontextust megőrző reprezentációja

Az ebben a fejezetben bemutatott kutatás célja az ETO-számok reprezentálása egy géppel olvasható, alkalmazásfüggetlen formátumban a szintaktikai relációk által kifejezett szemantika megőrzésével. A célok közé tartozik még egy olyan algoritmus kidolgozása, amely képes az ETO-jelzeteket közvetlenül lefordítani a tervezett formátumra, az MRF alkalmazása nélkül, továbbá egy, az algoritmust megvalósító online szolgáltatás létrehozása.

A formátum és a program online elérhető a http://piros.udc-interpeter.hu címen. Az ETO-jelzetek feldolgozhatók felhasználói vagy alkalmazásprogramozási interfészen keresztül, az XML-séma definíciója pedig letölthető és felhasználható a megfelelő licenc szerint.[3]

XML-séma definíció az ETO-jelzetek leírására

A legfontosabb követelmény a formátummal kapcsolatosan, hogy annak le kell írnia az ETO-jelzetek teljes szintaktikai struktúráját, megőrizve az összes releváns információt azok részeire, kapcsolódási módjaira, kifejezésben betöltött szerepére és sorrendjére vonatkozóan.

A második követelmény, hogy szabványos formátumnak kell lennie, mely feldolgozható más alkalmazások által, illetve könnyen átalakítható egyéb formátumokká.

Az ETO speciális tulajdonságait és a fenti követelményeket is figyelembe véve az XML-formátum megfelelőnek tűnt a kutatás céljaira. A választott szabvány legfőbb előnye a flexibilitás, a széles körű támogatottság és az XML-séma definíció (XSD) készítésének lehetősége.

Az ETO-számok reprezentálásának alapelvei

Az ETO alosztási szimbólumainak precedencia sorrendje azok koncepcionális definíciójából adódik. Az összetett tárgykörök összevonhatók egy viszonyítás (:) segítségével kifejezett komplex, az összetett és komplex tárgykörök pedig egy összekötéssel (+) kifejezett agglomerált tárgykörré. A kiterjesztés (/) felhasználható egymás melletti számok összekötésére, azaz intervallumok alkotására. Ezek az intervallumok, akárcsak a csoportképzéssel kifejezett tárgykörök éppúgy pontosíthatók fazettákkal, mint a táblázati számok.

A fenti precedenciasorrend minden ETO-szám esetén meghatároz egy fát, amelyben a különböző típusú tárgykörök különböző szinteken állnak. Például az 515.1+514.12 jelzetben az összekötés az első szinten van reprezentálva, a viszonyítás pedig a másodikon.

A fa legalsó szintjei az alapfogalmakat tartalmazzák, amelyek egy főtáblázati számból (intervallumból, szintézisből vagy csoportképzésből, esetleg egy önálló általánosan közös alosztásból) állnak, esetleg egy vagy több alosztással pontosítva.

Az általánosan, és egyes esetekben a korlátozottan közös alosztások is tartalmazhatnak további alosztási jeleket és számokat, amelyeket szintén kezelni kell. A levelek mindig táblázati számokat vagy intervallumokat jelölnek.

Egy másik előnye ennek a megközelítésnek, hogy az így létrehozott reprezentáció a fazetták fókuszát és bázisosztályát egyaránt tartalmazza anélkül, hogy szétválasztaná őket.[4] Például a 27'475.5-23 („szentírásokon alapuló szentbeszédek") jelzetben a bázisosztály (27) és a fazetta fókuszai (-475.5 és -23) jól felismerhető és reprodukálható módon vannak elmentve, függetlenül a fazettasorrendtől és a jelzet esetleges egyéb elemeitől. Így a

fazették érdemi zaj és információveszteség nélkül visszakereshetők.

A sémadefiníció

Minden fa leírható egy XML-lel. Az XML lehetséges elemei meghatározhatók egy XML-sémadefinícióval, amely így definiál egy az ETO-számokat leíró nyelvet.

Az XSD komplex típusai a fa ágait és leveleit határozzák meg. Az osztályok (táblázati számok) és az intervallumok olyan komplex típusok, amelyeknek két attribútuma tartalmazza az intervallumot kezdő és (opcionális) záró számot. A táblázati számokat leíró egyszerű típusok validációs célokat szolgálnak.

A következő példa egy bonyolult ETO-szám XML-reprezentációját mutatja be:

```
<ns:udc_concept
  xsi:schemaLocation="http://piros.udc-
interpreter.hu/#xsd udc.xsd"   xmlns:ns="http://piros.udc-
interpreter.hu/#xsd"   xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance"    udc_edition="2017"
  notation="[515.1+514:517]-32(02.025.2)=161.1">
  <ns:description xml:lang="EN">
    Topológia és az analízis és geometria kapcsolata (könyv orosz nyelven, illusztrációkkal)
  </ns:description>
  <ns:main_concept>
    <ns:main_table_subgrouping>
      <ns:main_table_addition>
        <ns:main_concept order="1">
          <ns:main_table_number number1="515.1"/>
        </ns:main_concept>
        <ns:main_table_relation order="2">
          <ns:main_concept order="1">
            <ns:main_table_number number1="514"/>
          </ns:main_concept>
          <ns:main_concept order="2">
            <ns:main_table_number number1="517"/>
          </ns:main_concept>
        </ns:main_table_relation>
      </ns:main_table_addition>
      <ns:special_auxiliary xsi:type="ns:special_auxiliary_hyphen" order="1">
       <ns:special_auxiliary_number
            xsi:type="ns:special_auxiliary_hyphen_number" number1="-32"/>
        </ns:special_auxiliary>
    </ns:main_table_subgrouping>
    <ns:common_auxiliary_independent xsi:type=”ns:common_auxiliary_of_form” order="1">
      <ns:common_auxiliary_of_form_number number1="(02)">
        <ns:special_auxiliary xsi:type="ns:special_auxiliary_pointnought" order="1">
        <ns:special_
            auxiliary_number xsi:type="ns:special_auxiliary_pointnought_number" number1=".025.2"/>
        </ns:special_auxiliary>
      </ns:common_auxiliary_of_form_number>
    </ns:common_auxiliary_independent>
    <ns:common_auxiliary_independent xsi:type="ns:common_auxiliary_of_language" order="2">
      <ns:common_auxiliary_of_language_number number1="=161.1"/>
    </ns:common_auxiliary_independent>
  </ns:main_concept>
</ns:udc_concept>
```

Interpreter az ETO-jelzetek felbontására és XML-formátumba alakítására

Az XML-formátum megtervezését követően a következő lépés egy olyan interpreter program tervezése és megírása volt, amely képes az ETO-számokat az új formátumban leírni.

Az interpreterrel szemben támasztott legfontosabb követelmények a következők:

a) tekintettel kell lennie az ETO jelzetépítési szabályaira, megtartva a számok által hordozott minden információt (beleértve a részekre és azok teljes szintaktikai kontextusára vonatkozót is);
b) amennyire lehetséges, automatikusan kell feldolgoznia a számokat;
c) online elérhetőnek kell lennie felhasználók és programok számára egyaránt.

Az interpreter egy automata, amely az ETO-számok és alosztási szimbólumok formális nyelvét fogadja el. Az inputjai az ETO-szám és az ETO kiadási éve, amely alapján készült, az outputja a szám XML-reprezentációja, vagy egy hibaüzenet a probléma leírásával, amennyiben azt nem lehet feldolgozni, vagy nem felel meg a megadott ETO-verzió szabályainak.[5]

A kimeneti formátumok

Bár az XML szabványos, automatikusan feldolgozható formátum, közvetlenül nem támogatja az ETO használatának minden formáját. Ezért logikusnak és szükségesnek tűnik olyan konverziós metódusok elérhetővé tétele, amelyek a szoftverek számára könnyebben felhasználható formátumokat produkálnak.

Az XML-formátum mellett az interpreter HTML formában is képes megjeleníteni a legenerált fákat, illetve, ha szükséges, összeállítja a jelzetelemek listáját a kontextuális információ nélkül, JSON formátumban.[6]

Az alábbi példa egy ETO-jelzet elemeinek listáját mutatja, a szoftver által összeállított JSON sztring formájában:


```
{
    "concept":"378.007.1",
    "udc_edition":"1990",
    "pref_labels": {
        "pref_label_1": {"pref_label": "Főiskolák vezetése", "language": "HU"}
    },
    udc_numbers": {
        number_1": {
            "notation":"378","filing":"3T7T8C","uri":"http://udcdata.info/025169",
            "pref_labels": {
                "pref_label_1": {"language":"HU","pref_label":
                    "Felsőoktatás. Egyetemek. Főiskolák"}
            }
        },
        "number_2": {
            "notation": ".007.1", "filing":"P0T0T7T1C",
            "pref_labels":{
                "pref_label_1": {"language":"HU","pref_label":""}
            }
        }
    }
}
```


A legutóbbi kutatási eredmények

A fenti kutatási eredményeket és alapelveket részletesen kifejtettem a 2015. évi ETO Szemináriumon, illetve az Extensions and Corrections to the UDC utolsó dupla számában. [26][32] Ez a fejezet az azóta eltelt időszak eredményeit foglalja össze.

A Portugál Digitális Nemzeti Könyvtár – esettanulmány

A címben említett esettanulmányra 2015 végén került sor a The European Library (TEL) nyílt hozzáférésű adatbázisa alapján. A több mint 100 elérhető gyűjtemény közül a Portugál Digitális Nemzeti Könyvtárra esett a választás, elsősorban a köze-

pes méretű állomány, és az ETO-számokkal való nagymértékű lefedettsége miatt.[7]

Az adatbázis RDF/XML-formában való letöltését és az ETO-számok kinyerését követően a duplikációk törlésre kerültek. Az így nyert lista 13 741 különböző ETO-számot tartalmazott, melyek egy teszteralkalmazáson keresztül egyesével lettek feldolgozva. Ezen a módon a szolgáltatás néhány perc alatt a teljes listát feldolgozta.

A 13 741 szám közül 13 604-et sikerült hiba nélkül feldolgozni. A maradék 137 rekord megvizsgálása során két programhibára és öt olyan speciális osztályozási megoldásra derült fény, amelyeket a program még nem támogatott. A többi problémát gépelési hibák és szabálytalan indexelési megoldások okozták, vagy a jelzeteléskor használt és a feldolgozáskor megadott ETO-verziók különbsége. A feldolgozási hibák mellett az XML-validáció további gépelési hibákat és szabálytalan megoldásokat is felszínre hozott.

A teljesítmény tesztelése mellett az esettanulmány tapasztalatai kiváló visszajelzést jelentettek az adatformátum és a program további korrekciója és továbbfejlesztése számára.

Az XML-sémadefiníció új verziója

Az XML-sémadefiníció első verziója[8] nyomtatott és online ETO kiadások alapján készült.[9]

A teljes standard angol ETO-verziót tartalmazó, éves licenccel elérhető UDC Online [31] a kutatás későbbi fázisában lett bevonva. Ezt követően a formátumot újra kellett tervezni a legutóbbi kiadások alapján. A táblázatok online, felhasználóbarát interfészen keresztüli elérhetősége és a portál által nyújtott fejlett keresési és böngészési lehetőségek elősegítették a kivételes jelzetépítési megoldások megtalálását és a különböző verziók összehasonlítását, ezzel felgyorsítva a kutatást.

A UDC Online és az esettanulmányok, illetve a vonatkozó szakirodalom további áttekintése megfelelő alapot nyújtott a sémadefiníció és az interpreter hiányosságainak felismeréséhez és kijavításához.

Az említett korrekciók a sémadefiníció új verziójába kerültek bele, amely 2.1 verziószámon érhető el. Az új formátum áttekinthetőbb, jobban dokumentált és teoretikailag megalapozottabb, mint a megelőző volt. Ezen kívül megoldást nyújt számos olyan speciális és kivételes jelzetszerkesztési szabályra, melyeket a korábbi verziók nem kezeltek. A továbbiakban újabb módosításokra már nem lesz szükség, hacsak a táblázatok változásai ezt indokolttá nem teszik.

A sémadefiníció legfontosabb változásai

A legfontosabb módosítás, hogy a korlátozottan közös alosztások az általánosan közösekkel azonos módon vannak kezelve, így a legspeciálisabb jelzetépítési szabályok is kezelhetővé váltak. Általánosságban a korlátozottan közös alosztások az alábbi formában írhatók le:

```
<xsd:complexType name="special_auxiliary">
  <xsd:complexContent>
    <xsd:extension base="udc:special_auxiliary_root">
      <xsd:sequence>
        <xsd:element name="special_auxiliary_number" type="udc:special_auxiliary_number"/>
      </xsd:sequence>
    </xsd:extension>
  </xsd:complexContent>
</xsd:complexType>
```

Egy másik fontos változás, hogy valamennyi jelzetelem hivatkozási sorrendje eltárolható. Ez az információ nemcsak a sorrendképzés vagy a jelzetszintézis esetén lehet fontos, de szükséges lehet az eredeti jelzetek reprodukálásához, illetve olyan esetekben, amikor a sorrend befolyásolja a jelzet jelentését. [9]

Néhány helyen szükség volt a táblázati számokat leíró validációs szabályok módosítására is. Például az időalosztások (1g táblázat) alatti dátumokat és az időintervallumokat a következő egyszerű típus írja le:

```
<xsd:simpleType name="common_auxiliary_of_time_number_string">
  <xsd:restriction base="xsd:string">
    <xsd:pattern value="'\.\.\.'"/>
    <xsd:pattern value="'(-|\+)?[0-2]\d{0,3}'">
    </xsd:pattern>
    <xsd:pattern value="'(-|\+)?([0-2]\d{3}(\.\d{2}(\.\d{2}(\.\d{2}(\.\d{2}(\.\d{2})?)?)?)?)?)'">
    </xsd:pattern>
    <xsd:pattern value="'[3-9]((\.)?\d{1,4})(\.\d{1,4})*'"/>
  </xsd:restriction>
</xsd:simpleType>
```

A kutatás további, megoldandó kérdéseket is napvilágra hozott, például:

- Intervallumok támogatása a külső forrásból származó (* szimbólummal bevezetett) alosztásokon belül (1h táblázatok).
- Földi területek meghatározása a kvadránsok segítségével [(161/164) osztályok].
- Térbeli méretek, dimenziók [(18) alatti osztályok].
- Fordítások a nyelvi általánosan közös alosztásokban (=03.1/.9, illetve =030.1/.9 alatti fazetták).
- Korlátozottan közös alosztások a dialektusok, helyi és regionális nyelvek, változatok és tájnyelvek kifejezésére (=…'276/'282).
- Az etnikai általánosan közös alosztásokban a reláció [(=1:...)] gyakran használt megoldás a Portugál Digitális Nemzeti Könyvtárban. Mivel a korábbi ETO-verziókban ugyanez a művelet ponttal volt jelölve [(=1.4/9)], az XSD-nek és az interpreternek is kezelnie kell mindkét központozási jelet.

A sémadefiníció második verziójának kiadására (2.0) a fentiekhez hasonló megoldások miatt volt szükség.[10]

A legutóbbi (2.1) verzió egy elméleti alapú módosítást tartalmaz. Az 1.0 és 2.0 verziók a főtáblázati számokat az alapfogalom attribútumaiként kezelte, az alosztásokat pedig annak elemeiként. Ez a megoldás inkább Ranganathan "kép-fal" elvének [4] felel meg, és nem annak, ahogy a jelenlegi ETO a közös alosztásokat kezeli.[11] Jelenleg a független általánosan közös alosztások a főtáblázati számokkal azonos szinten álló elemek, melyek az összetett fogalom bármely pontján állhatnak, de akár főtáblázati szám nélkül, önálló jelentéssel is szerepelhetnek. [8]

Ennek jobban megfelel az a megoldás, ha a főtáblázati számokra (esetleg intervallumokra, szintézisekre vagy csoportképzésekre) a fogalom elemeiként tekintünk, melyek az esetleges független közös alosztásokkal azonos szinten állnak. Ezt az alábbi komplex típus fejezi ki:

```
<xsd:complexType name="main_concept">
  <xsd:sequence>
    <xsd:choice minOccurs="0" maxOccurs="1">
      <xsd:element name="main_table_number" type="udc:main_table_number"/>
      <xsd:element name="main_table_synthesis" type="udc:main_table_synthesis"/>
       <xsd:element name="main_table_subgrouping"
            type="udc:main_table_subgrouping" minOccurs="1" maxOccurs="1"/>
    </xsd:choice>
    <xsd:element name="common_auxiliary_independent"
    type="udc:common_auxiliary_independent" minOccurs="0" maxOccurs="unbounded">
     </xsd:element>
  </xsd:sequence>
  <xsd:attribute name="order" type="xsd:int" use="optional"/>
</xsd:complexType>
```

A szoftver evolúciója

A szoftver módosítása az új XML-verzió követése érdekében

A sémadefiníció változásaiból következik, hogy az interpretert is módosítani kell azok követése érdekében. Ez a feladat elsőbbséget kell, hogy élvezzen minden további fejlesztés előtt.

További output-formátumok

Az XML és a KWOC mellett további géppel olvasható formátumok tervezése, illetve implementálása is folyamatban van.

A MARC formátumcsaládban két jelentős formátum létezik, amelyek határozottan osztályozási rekordok leírására és adatcseréjére lettek létrehozva. A MARC21 (korábban USMARC) osztályozási formátum [35] elsősorban DDC- és LCC-jelzetek számára lett létrehozva és nem kezelik az ETO speciális jelzetképzési szabályait.

A UNIMARC formátum tervezése valamivel később kezdődött, a MARC21 tapasztalatai alapján, de kimondottan az ETO-jelzetek kezelésének céljával. A formátum tervezése azonban félbeszakadt, és jelenleg is befejezetlen formájában érhető el az IFLA honlapján [36], bár továbbfejlesztésére több javaslat is született [37]. Mivel a UNIMARC formátum, különösen a javasolt változtatások után, az ETO-számok kezelésére leginkább alkalmas ETO-formátum, indokolt az interpreternek is támogatnia. Ezért ennek a kimeneti formátumnak a tervezése is elkezdődött és megvalósítása is folyamatban van.

A prekoordinált ETO-számok RDF-ként is reprezentálhatók. A tripletek meghatározhatók az XML alapján, az URI-k pedig a szabadon elérhető ETO osztályok, illetve alosztások alapján.[12] A cikk megjelenésekor az RDF-séma és -output fejlesztés alatt áll. A pontos formátum a tervek szerint a konverzióért felelős szoftverkomponenssel együtt lesz publikálva.

Elérhetőség RESTFul interfészen keresztül

A REST ("Representational State Transfer", Reprezentációs Adatátvitel) egy osztott hipermédia rendszerek számára létrehozott tervezési stílus. A REST olyan architekturális megszorításokat definiál, melyek maradéktalan megvalósítása biztosítja a komponensek interakciójának gyorsaságát és skálázhatóságát, az interfészek generikusságát, a komponensek független telepíthetőségét. [38]

A szolgáltatás jelenleg egyszerű HTTP hívásokon keresztül érhető el. A RESTFul stílusú átszervezés egy olyan standard interfészt nyújtana, amely még inkább megkönnyítené a rendszer alkalmazását más szoftverek számára. Ezért a tervek között szerepel a szolgáltatás átszervezése egy, a jelenleginél hatékonyabb és rugalmasabb architektúra szerint.

Általánosságban elmondható, hogy a jövőbeli fejlesztési terveknek fontos részét képezik az interpreter további funkcionális fejlesztései és a további konverziós eljárások.

A tesztkészlet

A kutatásnak már a kezdeti szakaszában szükségessé vált egy tesztkészlet felépítése a szoftver integritásának megőrzése, illetve az ETO szabályainak kellően alapos feltérképezése és analízise érdekében.

A tesztkészlet több, mint 700 tesztesetet tartalmaz, a tesztek céljai szerint csoportosítva. Vannak tesztek arra, hogy megtudjuk, hogy a különböző szabályok, illetve az alosztási jelek, kiterjesztés, csoportképzés, külső forrásból származó jelzetek és névalosztások stb. alkalmazásakor felmerülő speciális esetek megfelelően vannak-e kezelve.[13]

Minden teszteset egy ETO-számot tartalmaz, az összeállításához használt ETO-verzió évszámával, illetve az XML-t, amit a szám feldolgozása után az interpreternek produkálnia kell.

A tesztek manuálisan vagy automatikusan is felhasználhatók annak ellenőrzésére, hogy a feldolgozás eredménye megfelel-e az elvárásoknak.

Általában a teszteseteknek nem szükséges valós, jelentéssel bíró ETO-számokat tartalmazniuk, erre csak akkor van szükség, ha a központozási jelek önmagukban nem határozzák meg egyértelműen a jelzetelemek típusát és feladatát. A legtöbb jelzet könyvtári katalógusokból, osztályozási tankönyvekből és cikkekből lett összegyűjtve, illetve, amennyiben nem sikerült olyan példát találni, amelyikben a tesztelendő megoldás előfordult volna, határozottan a teszt céljára lett létrehozva.

A szoftver karbantartása mellett egy ilyen tesztkészlet hasznos példákat szolgáltat az XML-

formátum használatára, illetve lehetőséget ad az ETO-számok prekoordinációjára vonatkozó szabályok felülvizsgálatára és jelentésük megértésére.

Befejezés

A sémadefiníció 2.1-es verziójának kiadása és a program átírása a kutatás első fázisának lezárását jelenti, a további kutatások már annak eredményeire építve, a kimenetek lehetséges felhasználási módjaira, illetve az összegyűjtött tapasztalatok felhasználására fókuszálhatnak.

Az XML-formátum egyik legnagyobb előnye az ETO-számok egyszerű eltárolásával szemben, hogy ez a formátum áttekinthető és a különböző programozási nyelvek által jól támogatott. Ezért a jelzetek további automatikus elemzése és konverziója speciális algoritmusok és nagyobb programozási munka nélkül is elvégezhető. Ez által minden olyan módszer hatékonyabbá tehető, amely felhasználja az ETO-számok szintaktikai struktúrájára vonatkozó információt, beleértve a kvantitatív vizsgálatokat [41], a kompozit fogalmak hasonlóságának mérését [42] vagy fejlettebb inkluzív keresési és böngészési algoritmusok kidolgozását.

A generált XML intelligens osztályozási interfészek kialakításának is alapjául szolgálhat. A jelzetek megfelelő jelölőkkel való ellátása segítheti azok automatikus rendezését a szabályok szerint. A KWIC-indexek építése lehetővé teszi a jelzetek elemeik alapján történő böngészését, azok kontextusának figyelembe vételével. Lehetséges a jelzetelemek permutációja az ETO szabályainak megfelelően, amely jelentősen növelheti a jelzet hozzáférési pontjainak számát. A végső cél az lenne, hogy a böngészés során a jelzeteket átalakítva automatikusan olyan formában jelenítsük meg, amely megfelel a keresés feltételezett céljának és a kereső feltételezett kognitív státuszának.

A formátum és a program felhasználásán kívül a kutatás során felhalmozott tapasztalat lehetőséget nyújt annak megvizsgálására, hogy segítik az utóbbi évek revíziói az ETO-jelzetek könnyebb kezelését és az osztályozás használatát.

ꕥ ꕥ ꕥ ꕥ ꕥ

Köszönetnyilvánítás

Mindenekelőtt szeretnék köszönetet mondani a családomnak mindazért a támogatásért és segítségért, melyet a kutatómunkám során számomra biztosítottak és biztosítanak. Külön szeretném megköszönni témavezetőm Dr. Boda István értékes és konstruktív javaslatait valamint Dr. Aida Slavic szívből jövő támogatását, a rengeteg értékes és hasznos információt, melyet a rendelkezésemre bocsátott. Végül köszönettel tartozom Daniel Benediktssonnak és Jonathan Wildnak a cikk megírásához és a konferenciára való felkészüléshez nyújtott segítségükért.

Megjegyzések

[1] A konferenciáról szóló részletes beszámoló a Könyv, Könyvtár, Könyvtáros 2017. októberi számában jelent meg. [1]

[2] Ha szükséges, fázisrelációt is képezhetünk, a kapcsolat fázisát a –042 (1k táblázat) alatti alosztásokkal jelölve.

[3] Az XML séma-definíció a Creative Commons Nevezd meg!-Ne add el!-Ne változtasd! 4.0 nemzetközi licenc szerint használható fel. [24]

[4] Az összetett ETO-jelzetek strukturális elemeinek részletes tárgyalása megtalálható Claudio Gnoli cikkében. [4]

[5] Bár célkitűzés, hogy az interpreter, amennyire lehetséges, szintaktikai alapon dolgozza fel a számokat, vannak esetek, amikor a jelzetelem táblázatokon belüli helye is meghatározza a szám struktúráját és elérési pontjait. Ilyen esetek például a szintézis vagy a 0/9 típusú párhuzamos alosztások. Az intervallumok zajmentes felbontása szintén igényli a táblázati számok ismeretét. [20]

[6] A JSON (JavaScript Object Notation) egy önleíró, könnyen érthető, nyelvfüggetlen adatcsereformátum. A JSON szerializáció során a program az ETO-szám által meghatározott objektumhierarchiát alakítja át a megfelelő sztringgé az alkalmazások közötti adatcsere számára.

[7] A The European Library (TEL), és a Portugál Digitális Nemzeti Könyvtár tesztelésre való felhasználásának ötletét Dr Nuno Freire vetette fel a 2015. évi ETO Szemináriumon. Bár a TEL portál karbantartása és frissítése 2016. december 31-én véget ért, az adatbázisok továbbra is elérhetők. [32]

[8] Az XML 1.0-ás verziója a 2015. évi ETO Szemináriumon lett bemutatva. [25]

[9] A felhasznált ETO-kiadások az 1990-es [25] és 2005-ös [26] magyar nyelvű nyomtatott kiadások, a BSI 2005-ben publikált nyomtatott szabvány ETO-kiadása [27], és a UDC Summary [28] voltak.

[10] A 2.0-ás verzió a Extensions and Corrections to the UDC-ben lett bemutatva. [31]

[11] Az előző megoldás az ETO tradícionális, prekoordinált szemléletét tükrözi, amely szerint az alosztások nem állhattak önmagukban és az alosztások sorrendjének a jelenleginél nagyobb jelentősége volt. Ez a megközelítés a nyolcvanas évek elején változott meg, elsősorban éppen a gépesítés megkönnyítésének segítése és a rendszer egységesítése céljával. [33]

[12] A teljes ETO SKOS/RDF formátumban való közzététele eredetileg a 2012. évi módosítások közzététele utánra volt tervezve [38], melyek a 2016-os Extensions & Correctionsben lettek publikálva. [40]

[13] A teljes tesztkészlet elérhető online a http://piros.udc-interpreter.hu#tests címen.

Irodalom

[1] PIROS Attila: The thought behind the symbol: about the automatic interpretation and representation of UDC numbers. = Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar 2017, London (UK), 14-15 September. Würzburg : Ergon Verlag, 2017. p. 203-218.

[2] PIROS Attila: A facettás osztályozás napjainkban: elmélet, technológia és a végfelhasználók. = Könyv, könyvtár, könyvtáros, 2017. (26. évf.) 11. sz. p. 22-30.

[3] OGDEN, C. K., RICHARDS, I. A.: The Meaning of Meaning. A Study of the Influence of Language Upon Thought and of the Science of Symbolism. 8th edition. New York: Harcourt, Brace & World. Inc., 1946. 363 p.

[4] RANGANATHAN, Siyali Ramamrita: Prolegomena to library classification. 3rd ed. New York, Asia Publishing House, 1967. 640 p. *http://hdl.handle.net/10150/106370* [2018. 01. 03]

[5] GNOLI, Claudio: Facets in UDC: a review of current situation. = Extensions and Corrections to the UDC. 33. The Hague : UDC Consortium, 2011. p. 19-36.

[6] BINWAL, J. C.: Modes of formation of subjects and their role in information retrieval. Dharwad, Karnatak University, 1988. 376 p. *http://shodhganga.inflibnet.ac.in/handle/10603/94558* [2018. 01. 03]

[7] BROUGHTON, Vanda: Essential Classification. Second Edition. London, Facet Publishing, 2015. 421 p. ISBN 978-1-78330-031-0

[8] Az Egyetemes Tizedes Osztályozás (ETO) alapelvei, revíziójának és kiadásának szabályai (FID 603). Budapest, OMIKK, 1983. 39 p. ISBN 963-592-247-7

[9] ROBINSON, Geoffrey: Citation Order in UDC. = Extensions and Corrections to the UDC. 25. The Hague : UDC Consortium, 2003. p. 19-27.

[10] SOERGEL, Dagobert: Indexing and retrieval performance: The logical evidence. = Journal of the American Society for Information Science, 1994. (45. évf.) 8. sz. p. 589-599. *http://www.dsoergel.com/cv/B46.html* [2018. 01. 03]

[11] McILWAINE, I. C.: The Universal Decimal Classification: some factors concerning its origins, development and influence. = Historical studies in information science. Medford, NJ : Information Today, 1998. p. 94-106.

[12] McILWAINE, I. C.: The new ecumenism: Exploration of a DDC/UDC view of religion. = Extensions and Corrections to the UDC. 28. The Hague : UDC Consortium, 2006. p. 9-16.

[13] McILWAINE, I. C., WILLIAMSON, Nancy: Medicine and the UDC: the process of restructuring Class 61. = Extensions and Corrections to the UDC. 30. The Hague : UDC Consortium, 2008. p. 9-16.

[14] GNOLI, Claudio: The UDC Philosophy Revision: First Report. = Extensions and Corrections to the UDC. 31. The Hague : UDC Consortium, 2009. p. 25-31. *http://hdl.handle.net/10150/200633* [2018. 01. 03]

[15] BROUGHTON, Vanda: Concepts and Terms in the Faceted Classification: the Case of UDC. = Knowledge Organization (KO), 2010. (37. évf.) 4. sz. p. 270-279.

[16] SLAVIC, Aida, DAVIES, Sylvie: Facet analysis in UDC: questions of structure, functionality and data formality. = Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar 2017, London (UK), 14-15 September. Würzburg : Ergon Verlag, 2017. p. 181-198.

[17] Classification & authority control: expanding resource discovery: proceedings of the International UDC Seminar 2015, 29-30 October 2015, Lisbon, Portugal. Würzburg, ERGON-Verlag, 2015. 248. p. ISBN 978-3-95650-124-1

[18] TARTAGLIA, S.: Authority Control and Subject Indexing Languages. = Cataloging & Classification Quarterly, 2004 (39. évf.) 1/2. sz. p. 365-377.

[19] RIGBY, Malcolm: Computers and the UDC. A decade of progress 1963−1973. (FID 523.). The Hague, FID, 1974. 108 p.

[20] RIESTHUIS, Gerhard J. A.: Decomposition of Complex UDC Notations. = Extensions and Corrections to the UDC. 19. The Hague : UDC Consortium, 1997. p. 13-19.

[21] RIESTHUIS, Gerhard J. A.: Searching with words : re-use of subject indexing. = Extensions and Corrections to the UDC. 21. The Hague : UDC Consortium, 1999. p. 24-32.

[22] RIESTHUIS, Gerhard J. A.: Zoeken met woorden : hergebruik van onderwerpsontsluiting. Amsterdam, University of Amsterdam, 1998. 186 p.

[23] MÁNDY Gábor: A posztkoordináció esélyei az ETO-ban. = Könyvtári figyelő, 2013. (59. évf.) 1. sz. p. 65−84. http://epa.oszk.hu/00100/00143/00086/pdf/EPA00143_konyvtari_figyelo_2013_1_065-083.pdf [2018. 01. 03]

[24] https://plato.stanford.edu/entries/compositionality [2018. 01. 03]

[25] https://creativecommons.org/licenses/by-nc-nd/4.0 [2018. 01. 03]

[26] PIROS Attila: Automatic interpretation of complex UDC numbers: towards support for library systems. = Classification & Authority Control: Expanding Resource Discovery : Proceedings of the International UDC Seminar 2015 29-30 October 2015 Lisbon, Portugal. Würzburg : Ergon Verlag, 2015. p. 177-193.

[27] Egyetemes tizedes osztályozás. Rövidített kiadás (FID Publ. No. 691). 1. kötet Táblázatok. Budapest, OSZK-KMK, 1990. 388 p. ISBN 963 593 109 3

[28] Egyetemes tizedes osztályozás (UDC Publ. No. P057). 1. kötet Táblázatok 1-2. rész. Budapest, OSZK KI, 2005. ISBN 963 201 609 2

[29] Universal Decimal Classification: standard edition: volume 1: systematic tables. London, British Standards Institution, 2005. 898 p. ISBN 0 580 45469 X

[30] http://www.udcc.org/udcsummary/php/index.php [2018. 01. 03]

[31] http://www.udc-hub.com [2018. 01. 03]

[32] PIROS Attila: New automatic interpreter for complex UDC numbers. = Extensions and Corrections to the UDC. 36-37 (2014-2015). The Hague : UDC Consortium, 2018. [megjelenés alatt]

[33] http://www.theeuropeanlibrary.org/tel4/access/data/opendata [2018. 01. 03]

[34] BABICZKY Béla: Szemléletváltozás az ETO jelzetszerkesztésében. = Könyvtári figyelő, 1985. (31. évf.) 1. sz. p. 17−27.

[35] http://www.loc.gov/marc/classification/eccdhome.html [2018. 01. 03]

[36] Concise UNIMARC Classification Format (2000). Concise Edition. The Hague, International Federation of Library Associations and Institutions (IFLA), 2000. 36 p.

http://softsbgp.free.fr/bibliotheque/Unimarc_Format_US.pdf [2018. 01. 03]

[37] SLAVIC, Aida: Faceted classification: management and use. = Axiomathes, 2008 (18. évf.) 2. sz. p. 257-271.

[38] FIELDING, Roy Thomas: Architectural Styles and the Design of Network-based Software Architectures. Irvine, University of California, 2000. 162 p. http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf [2018. 01. 03]

[39] http://universaldecimalclassification.blogspot.hu/2012/08/udc-as-linked-data.html [2018. 01. 03]

[40] Extensions and Corrections to the UDC. 34-35 (2012-2013). The Hague : UDC Consortium, 2016.

[41] SMIRAGLIA, Richard, SCARNHORST, Andrea, SALAH, Almila Akdag, GAO, Cheng: UDC in action. = Classification and visualization: interfaces to knowledge: proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands. Würzburg: Ergon Verlag, 2013. p. 259-272.

[42] LULA, Paweł, CIERASZEWSKA, Urszula: Similarity measurement between UDC classmarks and its application. = Faceted classification today: theory, technology and end users: proceedings of the International UDC Seminar 2017, London (UK), 14-15 September. Würzburg : Ergon Verlag, 2017. p. 219-240.

Beérkezett: 2018. I. 4-én.

Piros Attila
a Debreceni Egyetem Matematika- és Számítástudományok Doktori Iskolájának doktorjelöltje.
E-mail: *atilla.piros@gmail.com*