https://doi.org/10.3311/tmt.13273

A névterek mint a hiteles tudás forrásai

A Nemzeti Levéltár földrajzi névtér projektjének bemutatása

Bánki Zsolt, Szatucsek Zoltán, Záros Zsolt

A Magyar Nemzeti Levéltárban a névtérépítés már több mint egy évtizedes múltra tekint vissza. Az analitikus kutatói szakrendszerek létrehozásának támogatására épül a személyek, testületek és földrajzi nevek névtere. E névterek megértek az újragondolásra, a koncepcionális és funkcionális korszerűsítésre, valamint tartalmi bővítésre.

Az egyes névterek összetettsége és felhasználási szükségessége alapján az a döntés született, hogy első lépésként a földrajzi névtér megújítása történik meg, és azt követi a személy- és a testületi névtér, úgy, hogy a földrajzi névtér projektben kidolgozott modellt visszük tovább a következő lépésekben is. Itt kellett kialakítani azokat az innovatív eljárásokat, azt a sajátos logikát, amely egyaránt tükrözi a levéltári szakmai elvárásokat és alkalmazza a legkorszerűbb technológiát.

névtér, levéltár, Nemzeti Levéltár, földrajzi névtér

A névterek mibenlétéről

A 2000-es évek kezdetétől az archívumi[1] feltárás hangsúlya végképp az adatbázisok irányába tolódott el, és a már évtizedek óta zajló számítógépes katalógus/nyilvántartás építés helyett a moduláris felépítésű relációs integrált gyűjteménykezelő és munkafolyamatmenedzsment-rendszerek, sőt a gráfalapú megoldások kezdtek elterjedni. Ezt a folyamatot erősítette a szemantikus web megjelenése, amely a géppel feldolgozott és az interneten publikált adatokat ún. jelölőkkel, illetve szintaxissal látja el, és ezáltal gépi algoritmusok által olvasható, felismerhető, jelentéssel bíró kifejezésekké alakítja őket. A jelölők által értelmezett és felépített hálózat azután a legtávolabbi forrásokat köti össze értelmes kereséssel, releváns találati halmazokká.

Ezen futurisztikusnak tűnő, de már a jelenben létező univerzumnak az alapja a dolgok, vagy másnéven entitások egyértelmű „önazonos” meghatározása, amelynek előfeltétele – egy egyedi és állandó azonosító karaktersor, amelyet a nemzetközi szakirodalom Persistent Identifier[2] néven határoz meg.

Mindezek okán az archívumi rendszerek – a kifejezést a legtágabb értelemben használva, vagyis szélesebb kört értve alatta, mint a klasszikus közgyűjteményeket – részint önmagukon belül, részint önálló, de közösen használt szolgáltatásokként olyan adatállományokat kezdtek építeni, amelyek az univerzum egy-egy típusos, halmazba rendezhető, vagyis közös jellemzőkkel bíró, de egyedi meghatározással (névvel) kifejezhető létezőit, individuumait rendezik meghatározó tulajdonságaik alapján számítógépes rendszerekbe. Ezeket az adatállományokat nevezzük angolszász szakkifejezés szerint *namespace*-nek[3], magyarul pedig névtérnek. A szemantikus web terminológiájában ezeket a tudásszervező rendszereket ontológia néven ismerik. A névtér alapvető célja, hogy

1 Az archívum fogalma alatt általában a kulturális örökséget őrző gyűjteményeket értve.

2 Wikipedia, Persistent identifier, Elérhető: https://en.wikipedia.org/wiki/Persistent_identifier (2023. 11. 17.)

3 Wikipedia, Namespace, Elérhető: https://en.wikipedia.org/wiki/Namespace

Bánki, Zs., Szatucsek, Z., Záros, Zs. *A névterek mint a hiteles tudás forrásai. A Nemzeti Levéltár földrajzi névtér projektjének bemutatása*, Tudományos és Műszaki Tájékoztatás, 70(4), p. 472–482, 2023. https://doi.org/10.3311/tmt.13273

az egyes entitásokat önmagukkal azonos, összetéveszthetetlen egyediségként regisztrálja, úgy, hogy az entitáshoz tartozó névváltozatokat is rögzítse. A névtérben tehát egy-egy entitás a jellemzők, névváltozatok és egy azonosító ID együttese.

Egy relációs adatbázis szintjén mindez egy adatbázisrekord metaadatainak, illetve azok kapcsolati hálójának összességeként jelenik meg.[4]

Miután az univerzum halmazképzési lehetőségeinek tárháza kimeríthetetlen, ezért elméletileg a névterek sokfélesége is az, de megmaradva a józan észnél, az archívumi felhasználás szempontjai alapján határozzuk meg a minket érintő legfontosabb névtértípusokat.

Az úgynevezett GLAM szektor[5] (Gallery, Library, Archives, Museum), mint a névterek egyik legjelentősebb intézményi felhasználói köre, leginkább a személy-, család-, testületi-, rendezvény-, földrajzi- és köznév típusú névtereket használja.

Összefoglalva, a névterek a létezés individuumainak halmazba/osztályba sorolásával létrejött, többnyire logikai, hierarchikus viszonyokat is kifejező klasszifikációs struktúrák, amelyek informatikai rendszerekben egyedi azonosításra és pontos visszakeresésre szolgálnak.

Hazai és nemzetközi példák, jó gyakorlatok

A nemzetközi gyakorlatban két tipikusnak tekinthető megoldás terjedt el.

1. Az intézmények egy része *saját gyűjteménykezelő rendszerein belül* fejleszt névteret. Ezek megnevezése közgyűjteményi ágtól és korszaktól függően változott, de lényegük a fenti definíció szerinti értelemben azonos maradt. Könyvtári viszonylatban nevezték besorolási nevek állományának, *authority file*-nak, vagy az analóg világig visszatekintve egységesített neveknek is. A köznévterek tekintetében ide tartoznak a legkülönfélébb tárgyszóállományok- és rendszerek, a legegyszerűbb tárgyszólistától a tezauruszokig és más, kontrollált természetes nyelvű információkereső szótárakig.
2. A névterek fejlődését tekintve ezt a korábbi szakaszt követően létrejöttek különböző nonprofit és forprofit vállalkozások keretében olyan *önálló névterek* is, amelyeket gyűjteménytől függetlenül építenek, és a felhasználók a szolgáltatást elérve használnak fel saját rendszereikben. A felhasználó intézmények szakmai preferenciáik alapján választhatnak névteret, és a szolgáltató üzleti modelljének megfelelően alkalmazhatják azokat. E felhasználási körbe tartozhat a teljeskörű nyílt, ingyenes felhasználás, a letölthetőség biztosítása (pl. GeoNames), a fizetős elérés, illetve az egyes entitások egyedi honosítása (ez utóbbi eljárást választotta korábban a Magyar Nemzeti Galéria a Getty Research Institute Art & Architecture Thesaurus alkalmazásával).

Magyarországon a gyűjteménykezelő rendszerekbe beépülő névterek tekintetében a legnagyobb gyűjteménnyel rendelkező intézmények jeleskedtek, vagyis az Országos Széchényi Könyvtár, az MTA Könyvtár és Információs Központ, a Fővárosi Szabó Ervin Könyvtár, a Magyar Nemzeti Levéltár, valamint az egyetemi- és a kiemelkedő szakkönyvtárak.

A múzeumok döntően más utat jártak be, részint a gyűjteménykezelő rendszerek kései alkalmazása miatt, részint azért, mert előnyben részesítették a külső névterek alkalmazását. Múzeumi területen egyedül a Petőfi Irodalmi Múzeum épített tudatosan névteret saját gyűjteménykezelő rendszerében, és azt közkincsként publikálta az intézményi és magánfelhasználók számára. Ez a publikus, nyomtatott életrajzi forrásokból épített személynévtér ráadásul a gyűjteményfeltárástól függetlenül, önálló biográfiai tevékenységként épül mind a mai napig.

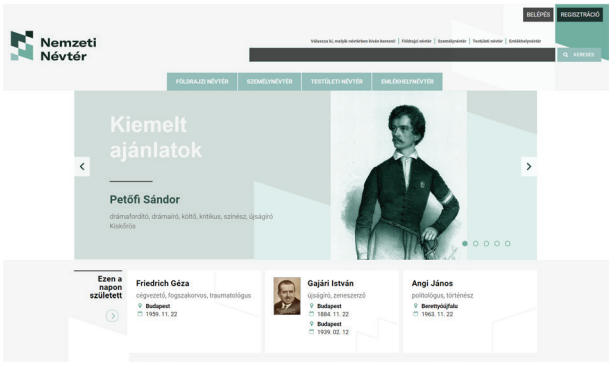
Mindenképpen meg kell említeni az „intézményfüggetlen” megvalósítás két jelentős példáját. A *Németh Tibor* celldömölki könyvtáros által több évtizede hivatástudatból épített Magyar Életrajzi Kalauzt,[6] amely fizetős szolgáltatásként érhető el; illetve a viszontagságos sorsa ellenére is a legjelentősebb hazai névtérépítő vállalkozásként nevezhető, a Nemzeti Névtér[7] létrehozására fordított jelentős erőfeszítéseket.

4 PIM, Digitális Bölcsészeti Központ, Dokumentációk, Elérhető: https://pim.hu/hu/digitalis-bolcseszeti-kozpont/dokumentaciok

5 Wikipedia, GLAM (cultural heritage), Elérhető: https://en.wikipedia.org/wiki/GLAM_(cultural_heritage)

6 Magyar Életrajzi Kalauz, Elérhető: https://mabi.hu/

7 Nemzeti Névtér, Elérhető: https://magyarnemzetinevter.hu/



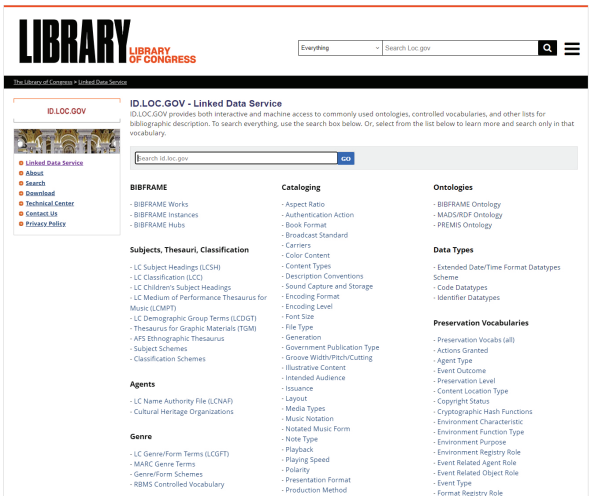
1. ábra Nemzeti Névtér

E sorba illik az Országos Széchényi Könyvtár égisze alatt, *Ungváry Rudolf* által épített legjelentősebb szabad felhasználású földrajzi- (Geotaurusz és Geohistaurusz) és köznévtér (Köztaurusz) is, amely a Magyar Országos Közös Katalógus (MOKKA) projektben lett publikálva.

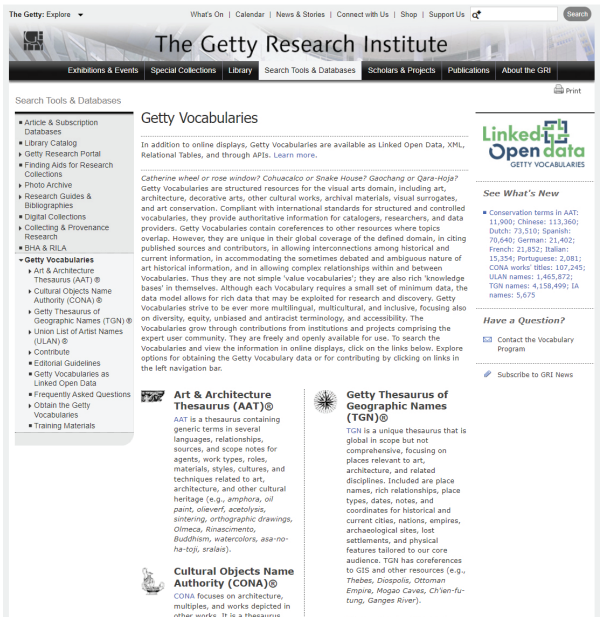
A Nemzeti Névtér (*1. ábra*), amelyet az OSZK Országos Könyvtári Rendszer projektjének keretében kezdtek fejleszteni, már alapkoncepciójában a teljes magyar archívumi közösség, sőt a legszélesebb felhasználói kör számára készült. Az idén elindított éles verzió bár egy publikus, ingyenes, korszerű, kollaboratív tartalomfejlesztésben működő szolgáltatás reményét nyújtja, még messze van egy országos központi szolgáltatástól elvárható színvonaltól (pl. az egyes entitások rekordjai nem tölthetőek le, illetve semmilyen szabványos formátumban sem jeleníthetőek meg a rekordok).

Ha a nemzetközi példákat tekintjük, akkor gyűjtemény által saját rendszerben működtetett névterek közül világelsők a Library of Congress szótárai,[8] (*2. ábra*) és meg kell említeni néhány olyan szolgáltatást is, amelyek névtérépítés tekintetében rendkívül jelentősek, és nem (köz)gyűjteményi eredetűek.

A Getty Research Institute (*3. ábra*) által vállalkozásként épített ontológiák a névtérépítés kiemelkedő produktumai és rendkívül elterjedtek az archívumépítéssel foglakozó intézmények körében[9] is. Földrajzi nevektől, személyneveken át a művészeti diszciplína fogalmait tartalmazó tezauruszokig tart a Getty névtérépítési kompetenci-



2. ábra A Library of Congress szótárai



3. ábra A Getty Research Institute

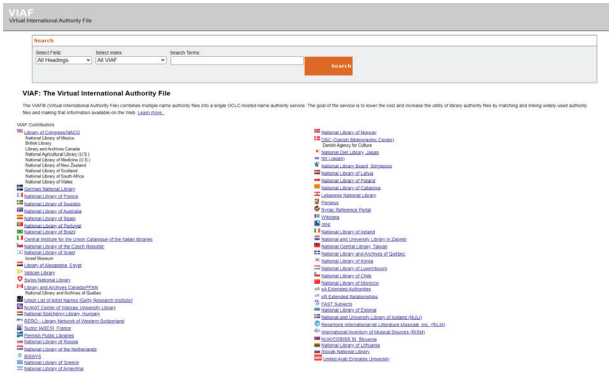
ája. Felhasználásuk az online kereséstől, a gyűjteményi rendszerekbe való betöltésen át az egyes rekordok honosításáig terjedhet.

A publikus földrajzi név szolgáltatások között talán a legnagyobb a GeoNames Team által 1990 óta épített, szabad felhasználású, több mint 12.000.000 földrajzi nevet tartalmazó névtér.

„A VIAF (Virutal International Authority File, Nemzetközi Virtuális Besorolási Állomány) (*4. ábra*) egy, az OCLC (Online Computer Library Center, Online Számítógépes Könyvtári Központ) által üzemeltetett, nyilvános katalógus rendszer. Célja a különböző nyelvterületek eltérő katalogizálási szabályai alapján kialakított névanyagok egész

8 Library of Congress , Elérhető: https://id.loc.gov/

9 Getty Research Institute, Eelérhető: https://www.getty.edu/research/tools/vocabularies/index.html



4. ábra Virutal International Authority File, Nemzetközi Virtuális Besorolási Állomány

világon elérhető virtuális katalógusba rendezése. A VIAF az országonként, nyelvenként, hagyományokként eltérő névvariánsokat, és személyekhez tartozó adattartalmakat kapcsolja össze."[10]

Névtér a Magyar Nemzeti Levéltárban

A Nemzeti Levéltárban a névtérépítés már több mint egy évtizedes múltra tekinthet vissza. Az analitikus kutatói szakrendszerek, - levéltári terminológia szerint - segédletek[11] létrehozásának támogatására ORACLE adatbázisban épül a személyek, testületek és földrajzi nevek névtere. E névterek, bár a mai napig ellátják alapvető funkciójukat – vagyis a segédletekben tükrözött egyes metaadatok mellé névtér ID-kat is rendelnek, és visszakeresési pontokként működnek az Adatbázisok online-on – megértek az újragondolásra, koncepcionális és funkcionális korszerűsítésre, valamint tartalmi bővítésre.

Ezért született meg egy új névtér fejlesztési projekt szándéka a Nemzeti Levéltár Informatikai és Innovációs Igazgatóságán. A tervezéskor az összes névtér szegmens megújítása belekerült a projekt látókörébe, de a feladat bonyolultsága miatt pontos munkamenetet kellett megállapítani.

Az egyes névterek összetettsége és felhasználási szükségessége alapján az a döntés született, hogy első lépésként a földrajzi névtér megújítása történik meg, és azt követi a személy- és a testületi névtér, úgy, hogy a földrajzi névtér projektben kidolgozott modellt visszük tovább a következő lépésekben is. E tekintetben a földrajzi névtér projekt a többi pilotjának is tekinthető. Itt kellett kialakítani azokat az innovatív eljárásokat, azt a sajátos logikát, amely egyaránt tükrözi a levéltári szakmai elvárásokat és alkalmazza a legkorszerűbb technológiát.

A tanulmány további részében e modellépítést és a földrajzi névtéren elvégzett fejlesztő munkát mutatom be.

Földrajzi névtér projekt – a koncepció

Az MNL földrajzi névtere 2011-ben jött létre a Geotaurusz akkori verziójának migrációjával. A földrajzi névtérben az azóta eltelt időben mind új entitások felvételére, mind adatgazdagításra is sor került. A nagyságrendileg 70.000 földrajzi nevet tartalmazó állomány tehát nem teljesen azonos a 2011-essel. Az adatállomány legnagyobb hiányossága az volt, hogy töredékesen tartalmazott koordinátákat, és a meglévők egy részét is hibásnak találtuk.

Fontos szempont a projekt tervezéséhez, hogy a névteret létrehozó szakemberek az Oracle adatbázisba egy, már létező nemzetközi adatmodellt implementáltak, és a migrálandó adatokat ennek az adatstruktúrának feleltették meg. A modell forrása a Getty Research Intstitute névterekre kialakított és ingyenesen nyilvánosságra hozott terméke[12]. Ez a modell rendkívül komplexen, de egyúttal flexibilisen kezeli a metaadatokat, így például lehetséges volt az ISO 2788-86 nemzetközi és az MSZ 3418-87 magyar tezaurusz-szabványoknak megfelelően épített Geotauruszt megfeleltetni a Getty adatmodellnek. Ezt a megoldást – legjobb tudomásom szerint – Magyarországon nem alkalmazzák, így az MNL úttörő szerepet tölt be a honosításban.

Az nyilvánvalónak tűnt, hogy ezen az úton fogunk tovább haladni az új fejlesztés során is.

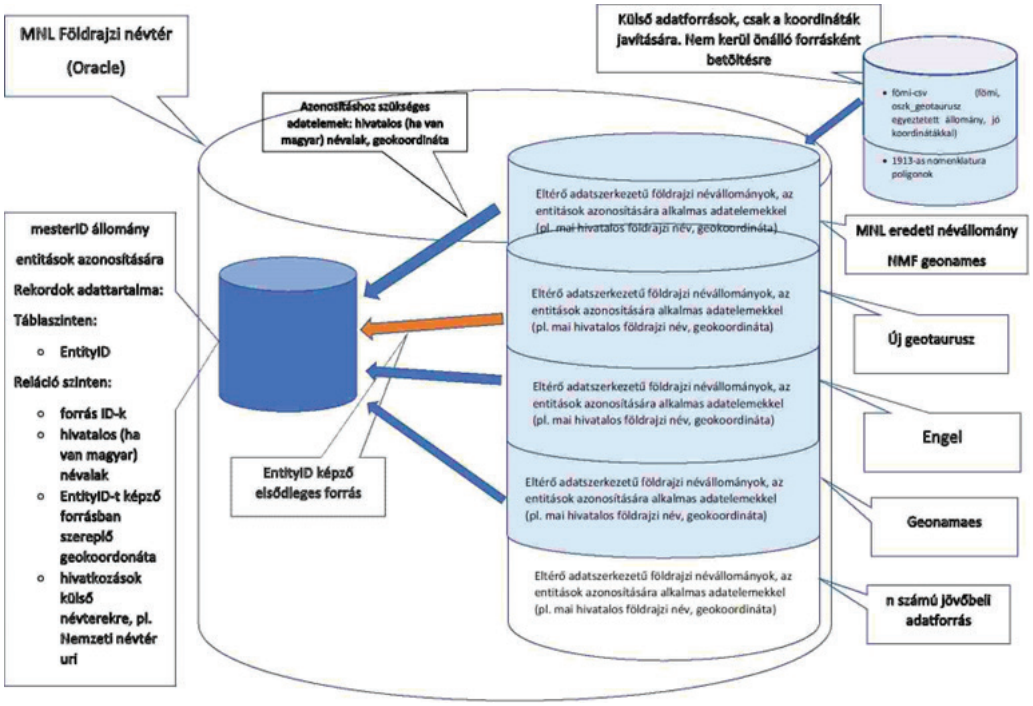
Alapos egyeztetések után a földrajzi névtér tekintetében a következő alapelveket fektettük le:

1. Az MNL földrajzi névtere heterogén, redundáns állományokból épül, amelyekben a földrajzi név entitásokat egy Entity_ID rekord azonosítja és egységesíti. (*5. ábra*)

10 Szakadát István: Névtér. (kézirat) 2018.

11 Magyar Nemzeti Levéltár, Levéltári szakkifejezések, Elérhető: https://mnl.gov.hu/mnl/ol/leveltari_szakkifejezesek#seg%C3%A9dlet

12 Getty Research Intstitute, Download Center, Elérhető: https://www.getty.edu/research/tools/vocabularies/obtain/download.html



5. ábra A földrajzi névtér modellje

2. A projekt első lépése a meglévő állomány adattisztítása, adatgazdagítása új, helyes koordinátákkal.
3. Az új földrajzi névtér tervezett adatállományai a következők:
   a. a jelenlegi MNL földrajzi névtér (MNL_GEO)
   b. a Geotaurusz aktuális állománya
   c. a középkori Magyarország digitális atlasza („Engel Pál" féle állomány)[13]
   d. a GeoNamesnek a Földre vonatkozó teljes állománya

Projektszakaszok

1. Input állományok (adatforrások) beszerzése.
2. Input állományok összevetése az MNL állománnyal.
3. Geokoordináták javítása, MNL állomány tisztítása.
4. Adatsémák megfeleltetése a Getty-Oracle adatsémának.
5. A Getty-Oracle adatséma felülvizsgálata, szükség esetén módosítása.
6. További adatforrások előkészítése betöltésre.
7. Adatforrások migrálása az Oracle adatbázisba.
8. Entity_ID (Mester_ID) rekordok létrehozása, adatforrások entitásainak megfeleltetéseivel.
9. A névtér ID-k összekötése a segédlet adatbázisokkal.
10. Publikus névtér *interface* fejlesztés.

A redundancia és a heterogenitás elve

Az általánosan elterjedt gyakorlat szerint a névterek alkotói megkeresik azt a szervező elvet, amely alapján kijelölik azt a névalakot, amely egy-egy entitást reprezentál. A könyvtári szabványosság ezt a formát nevezi egységesített vagy kitüntetett névalaknak, az eltérő formákat pedig utalóknak, vagy névváltozatnak. A felhasználás tekintetében az adott gyűjteményi forrás (pl. egy kézirat) leírásánál következetesen az egységesített alakot használja.

A levéltári gyakorlat számára azonban ez az eljárás nem megfelelő, hiszen elképzelhetetlen lenne, hogy egy középkori forrásban szereplő korabeli helységnév helyett a mai közigazgatási állapotnak megfelelő elnevezés kerüljön be egy segéd-

13 Az adatbázis a Magyar Királyság részletes térképét mutatja be a középkor végén, azaz az 1500-as évek körüli évtizedekben. Megtalálható rajta minden olyan település (város, mezőváros, falu, puszta, vár, kolostor – több mint 23 000), amely akkor létezett, és amelynek fekvése legalább hozzávetőlegesen megállapítható volt. A fontosabb vizeken és a megyehatárokon kívül feltünteti a jelentősebb királyi, egyházi és világi földbirtokok határait is (az 1498. évi állapot szerint). https://abtk.hu/hirek/1713-megujult-engel-pal-adatbazisa-a-kozepkori-magyarorszag-digitalis-atlasza

letbe. Azt az utat kell tehát járni, hogy a névtérbe kerülő összes földrajzi név – akár kitüntetett névalak, akár névváltozat – szükség szerint kiválasztható legyen.

E felismerés – a maga újszerűségével – megnyitotta a kaput abba az irányba, hogy redundánsan egymás mellé helyezzünk több földrajzi név állományt úgy, hogy azokat nem gyúrjuk össze egyetlen homogén masszává, hanem integritásukat és az eredeti kontextusokat minél teljesebben megőrizve nyújtsunk választási lehetőséget a felhasználóknak. A rendszer egyúttal nyitott marad további adatforrások befogadására is.

E megoldás nyilvánvalóan jelentős redundanciát eredményezett az egyes adatforrások között, de azokon belül megőriztük az egyediség elvét.

A forrásállományok hitelességének garantálása végett született meg az a döntés, hogy a jövőben, a feltáró munka közben csak az MNL eredeti névtérállományát fogjuk bővíteni, módosítani, adatgazdagítani, a külső, kompakt adatforrásokat validált egységükben őrizzük, azzal az opcióval, hogy időről időre érdemes egy update-tel frissíteni az állományt.

Nem mondhattunk le azonban arról a célról, hogy – bár az egyes források parallel léteznek – a névtérben mégis megteremtsük az entitások azonosságának kifejezését. E célt szolgálja az úgynevezett Entity_ID létrehozása, amely az összes névtérkomponensben logikai egységesítést hoz létre. E technikai csomópont mentén együtt láthatóak mindazok a névformák és kapcsolatok, amelyek fogalmilag összetartoznak.

„A földrajzi névtér létrehozása során a különböző adatforrásokban szereplő rekordok egyértelmű (elsődleges) azonosítóikon keresztül kerülnek összekapcsolásra az egyes adatforrások között. Ezen azonosításhoz egy újonnan létrehozott, úgynevezett Entity_ID-t használunk, mely egyértelmű entitásként azonosítja az egyes rekordokat, (csak olyan földrajzi neveket tekint egyezőnek, melyek ugyanazon földrajzi helyet írják le és azonos típusúak).”[14]

14 A Magyar Nemzeti Levéltár földrajzi névtér adatbázis rekordjainak összekapcsolása és levéltári segédletekkel való összerendelése – [Stratis Vezetői és Informatikai Tanácsadó Kft.:] Feladatleírás. Kézirat, 2022.

A meglévő állomány adattisztítása, adatgazdagítása új, helyes koordinátákkal

A meglévő állomány javításánál jelentős nehézséget jelentett, hogy a 2011-es input állománynak tekinthető Geotaurusz nem rendelkezik azonosító elemmel, illetve az aktuális, 2022-es verzió közvetlenül nem tartalmaz geokoordinátákat. Rendelkezésünkre állt azonban a Földmérési és Távérzékelési Intézet adatbázisa, amely 78.798 földrajzi hely pontos geokoordinátáját tartalmazta. Ezen adatbázis rekordazonosítói viszont be voltak építve az aktuális Geotauruszba. Így két lépcsőn keresztül mégis lehetséges volt a koordináták javítása, amennyiben sikerült magát a földrajzi helyet egyértelműen azonosítani az MNL rendszerében és a Geotauruszban. Ezt az műveletet a névalak azonossága alapján és GEO DISTANCE-el végeztük el. Tehát egy, a Föld sugarával megegyező gömb (tudjuk, hogy a Föld nem tökéletes gömb, ezért nem 100% pontos) palástjára vetítettük a koordinátákat és a két pont közötti legrövidebb térbeli távolságot vettük méterben és a következő eredményt értük el:

- 70.854 Subject
- 75.276 Term (névváltozat)
- 25.022 koordináta összesen
- 15.832 javított koordináta (63,58%)
- 9.190 javítatlan koordináta
- 45.832 koordináta nélküli subject

A számokat szemlélve – mielőtt gyors ítéletet mondanánk az eredményességet illetően – két szempontot vegyünk figyelembe. Az első az, hogy az azonosság megállapításánál nem tartottuk elegendőnek a névalak azonosságát, így a koordináta nélküli esetek kiestek a javítandók köréből. Ezt nem tekinthetjük veszteségnek, mivel a Geotaurusz koordinátákkal kiegészített aktuális verziója szerepelt a betöltendő új adatállományok között, és a Entity_ID képzés eredményétől vártuk el a régi és az új állomány azonosságainak megállapítását.

A másik szempont a 63,58% javítási arány, amely igen jónak mondható, hiszen abból a tapasztalatból indultunk ki, hogy a MNL állomány jelentős mértékben tartalmaz jó koordinátákat is. Megítélésünk szerint a projektszakasz megfelelő eredményt hozott.

Adatmapping és migráció – új adatforrások

A kiinduló, meglévő állomány adattisztítása után a következő lépésben az új adatforrásokkal folytattuk a munkát. Első lépésként pontos térképet kellett készíteni a rendelkezésünkre álló három állomány – két adatbázisból származó export, és egy szövegfájl – struktúrájáról és ezeket meg kellett feleltetni a Oracle-ben létező Getty modellnek.

Tapasztalatunk szerint a Getty adatmodell döntően alkalmas volt az input állományok fogadására, amit módosítani kellett, azt megengedte a rendszer rugalmas szerkezete.

A tanulmány megírásának pillanatában a tervezett állományok közül betöltöttük a GeoNames allCountry, a Geotaurusz és a középkori Magyarország digitális atlasza („Engel") állományait. Az adatbetöltés statisztikai adatai imponálók (*1. táblázat*).

Ezzel a robusztus névállománnyal a Magyar Nemzeti Levéltár Magyarország legnagyobb földrajzi névterét hozta létre, amely a történelmi Magyarország tekintetében rendkívül részletes és a teljes Földre vonatkozóan is szolgáltat neveket.

A hátralévő feladatokat már a létrehozott adatállományon kellett elvégezni, illetve a segédletadatbázisokkal való összekötéssel folytatódott.

Entity_ID rekordok létrehozása, adatforrások entitásainak megfeleltetéseivel

A projektszakasz elsődleges célja az egyes adatkészletekben megtalálható redundáns földrajzi nevek azonosságainak megállapítása, névváltozataikkal együtt.

Ehhez világosan kell látni, hogy a Getty modellben minden földrajzi név egy-egy TERM-öt képez. A TERM-ök vagy Preferred, vagy Variant státuszúak és a Pereferred TERM-ök a SUBJECT-ek. Egy SUBJECT-hez N számú TERM tartozhat (Vezérszó, utaló, Név, névváltozat stb.)

Az egyes azonos entitást reprezentáló SUBJECT-eket egy ENTITY ID fogja össze. Az ENTITY ID fejezi ki az adott földrajzi entitás egységét.

1. táblázat Az adatmigráció számszerű eredményei

Adatforrás	Földrajzi nevek	Entitások	Névváltozatok
GeoNames	19.172.396	12.237.573	6.934.823
Geotaurusz	135.414	109.047	26.367
„Engel"	56.084	24.148	31.936

Az egységesítés folyamatához szükséges volt az adatforrások közötti preferencia sorrend megállapítása. Az MNL által meghozott döntés értelmében a sorrend a következő: Geotaurusz, MNL_GEO, GeoNames, Engel.

A feladat végrehajtására célszoftvert fejlesztettünk külső partner bevonásával, amely szabályalapú és mesterséges intelligenciára épített eljárást alakított ki. Az azonosítást végző modell futtatása szükség esetén – például új adatforrás betöltésekor – tetszőleges alkalommal elvégezhető saját hatáskörben.

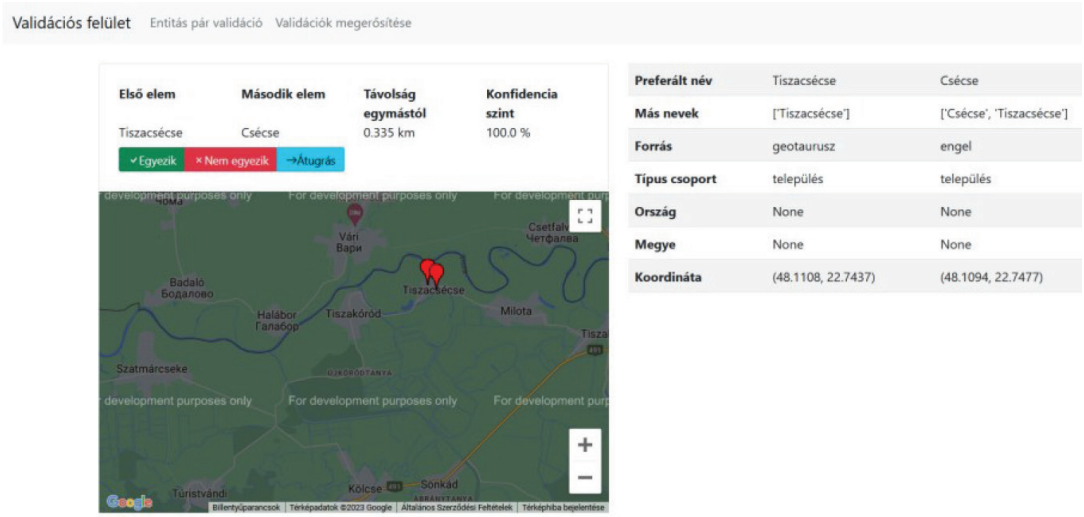
A MIREL (Mesterséges Intelligencia Reláció) névre hallgató alkalmazás fejlesztése kapcsán elvégeztük a teljes névtérhez kapcsolódóan az egyes földrajzi nevek tipizálását, amelyhez kiválóan fel tudtuk használni a Geotaurusz A/F relációját, és a GeoNames Feature Codes kategóriáit. Az eredményt a Getty modell PTYPE_GROUP táblájába töltöttük be.

Entitás azonosítás – teljes összekapcsolás menete

A MIREL-ben (*6. ábra*) végzett munkafolyamat lépései a következők:

1. Konfigurációs fájl kitöltése: szükséges információk megadása - milyen adatok állnak rendelkezésre és azok milyen oszlopnévvel szerepelnek.
2. PTYPE_GROUP mappelés: amennyiben elérhető a földrajzi fogalom típusnak a standardizált PTYPE_GROUP taxonómiához mappelése.
3. Adatok beolvasása és tisztítása: Term-ek tisztítása, felesleges karakterek és többletjelentéssel nem bíró stopszavak (pl. megye, hegy, domb) kiszűrése.
4. Szabály alapú összekapcsolás: amennyiben teljes és kizárólagos az egyezőség a bekapcsolni kívánt földrajzi hely adathalmaz és a meglévő entity adathalmaz között.
5. AI alapú összekapcsolás: karakterhasonlóságra és földrajzi távolságot felhasználó feature-ökre épülő gépi tanulási modellt használva.
6. Manuális validálás: azoknak az eseteknek az ellenőrzése, amikor a modell közepes konfidenciájú találatot adott vissza.

Az entitásegyesítés eredményeképpen 12.319.217 ENTITY rekord átkerült a névtérbe.



6. ábra A MIREL-ben végzett munkafolyamat lépései

Azt vizsgálva, hogy egy entitáshoz milyen arányban tartozik több adatforrásból földrajzi név, az alábbi eredményt kapjuk:

1. 4 elemet tartalmazó entitás 10.425 db.
2. 3 elemet tartalmazó entitás 24.790 db.
3. 2 elemet tartalmazó entitás 23.571 db.

Összesítve: a 12.319.217 ENTITY rekord közül 58.786-hoz tartozik egynél több SUBJECT rekord, és 163.212 SUBJECT tartozik olyan entitáshoz, amelynek több eleme van.

Figyelembe véve, hogy az egyes adatforrások döntően a történelmi Magyarország tekintetében mutatnak átfedést, ez az eredmény kiválónak tekinthető. A „szingli" entitások java része a GeoNames-ből, és világ más tájairól származik.

A névtér ID-k összekötése a segédlet adatbázisokkal

A projekt hasznosulásának egyik legizgalmasabb eredménye, hogy az https://adatbazisokonline.mnl.gov.hu/ -n folyamatosan növekvő számú segédlet adatbázist automatizált módon gazdagítjuk az új névtér elmeivel. Az adatrögzítéskor az egyes adatbázisok megfelelő tábláiban létrejöttek ugyan a földrajzi nevek leírásai, de ezek csak töredékesen lettek – intellektuálisan – összekötve a földrajzi névtérrel. A rögzített földrajzi nevek döntő többsége tehát csak stringként szerepelt a rendszerben.

Az MNL már korábban is alkalmazott mesterséges intelligenciára építő adatgazdagítást a Szovjet táborok magyar foglyai adatbázis építésénél (*7. ábra*), ahol személy- és földrajzi nevek azonosításához használtuk a MIREL speciálisan erre az esetre feltanított verzióját, de itt a felismert (cirill betűs szövegből átírt) metaadatok nem kerültek névtérbe.

Erre az előzményre építettük a MIREL – az entitásegyesítés melletti – másik funkcióját, amely a segédletek földrajzi neveinek névtérelemekkel való azonosítását és összekötését végezi el.

Az adatgazdagítás kiindulópontja az a konfigurációs állomány, amelyben paraméterezni lehet a névtérelemek és az adatbázis string-ek összevethetőségét. Ezek között – egyebek mellett – a geolokáció (koordináták, poligonok) távolsági és bennfoglaló meghatározása, a névazonosság mértéke, a földrajzi hely tipológiája (PTYPE_GROUP) szerepel. A szoftver maximum három lehetséges javaslatot tesz, az azonosítás bizonyosságát jelző konfidencia értékkel jelölve.

7. ábra Szovjet táborok magyar foglyai adatbázis adatlapja

A tesztkörnyezetben elvégzett futtatásaink alapján a következő megállapításokat tettük:

1. 0,6 fölötti konfidencia értékkel bíró azonos névalakú elemek megbízhatóan azonos entitások.
2. 0,6 fölötti konfidencia értékű, nem azonos névalakú elemek döntő többsége azonos földrajzi helyet jelent, csak más írásképben. pl: Kismórichida - Mórichida (Kis-)
3. 0,5 és 0,6 között a megbízhatóság már tapasztalhatóan csökken, de az azonossági arány még publikálhatóvá teszi az eredményt.
4. A konfidenciaértéket tekintve a kritikus érték a 0,5. Ez alatt már jelentős tömegben eltérő párok találhatóak. Az alacsony konfidenciaértékkel bíró párokat sem kell azonban automatikusan elvetni, hanem együtt kell kezelni a szövegazonosságot (a segédletben szereplő földrajzi hely és a névtér elem között) és a konfidenciaérték megállapított küszöbértékét. Vagyis tételezzük fel, hogy a publikációs küszöbnek a 0,5-öt határozzuk meg, akkor egy olyan esetben, ahol a konfidenciaérték alacsonyabb, mint 0,5, de szövegazonosság áll fenn, a szövegazonosság alapján publikálhatónak tartjuk az eredményt.

A *2. táblázatban* egy MIREL-lel végzett adatgazdagítás eredményének részletét láthatjuk, ahol korábban intellektuális névtérazonosítást is végeztek. Figyelmesen szemlélve még azt is felfedezhetjük, hogy az új névtérben szerepel olyan földrajzi név, amely pontosabb azonosítást tett lehetővé a mesterséges intelligencia számára, mint a korábbi eljárás.

Miután az eljárás könnyen rutinná tehető, egyszerűen végezhető humáninformatikai feladat, hogy az eddig névtérhez nem kötött földrajzi nevek nagy sebességgel és tömegben kapcsolhatóak a földrajzi névtérhez.

A projektet 2024 elején egy új publikus felület létrehozása koronázza meg és zárja le, amely a névteret a felhasználók számára teljes komplexitásában szolgáltatja.

2. táblázat MIREL adatgazdagítási eredmény

SEGEDLET_NEV (feldolgozáskor rögzített string)	**SEGEDLET_NEVTER_TERM (feldolgozáskor intellektuálisan létrehozott névtér kapcsolat)**	**MIREL_TERM (AI-al megállapított névtér kapcsolat)**	**CONFIDENCE**
Smíchov	Prága	Smíchov	0,908605754
Budapest 13. kerület	Budapest 13. kerület	Budapest 13. kerület	0,899937868
Filatorigát	Filatoridűlő	Filatorigát	0,898101151
Máriaremete	Máriaremete	Máriaremete	0,894359827
Uherské Hradiště	Uherské Hradiště	Uherské Hradiště	0,891546011
Vérd	Vérd	Vérd	0,891546011
Tomislavgrad	Tomislavgrad	Tomislavgrad	0,891546011
Split	Split	Split	0,891546011
Tourcoing	Franciaország	Tourcoing	0,891546011
Diessenhofen	Svájc	Diessenhofen	0,891546011
Daugavpils	Lettország	Daugavpils	0,891546011
Nadvirna	Nadvirna	Nadvirna	0,891546011
Sinj	Sinj	Sinj	0,891546011
Speyer	Speyer	Speyer	0,891546011
Gloggnitz	Gloggnitz	Gloggnitz	0,891546011
Issy-les-Moulineaux	Issy-les-Moulineaux	Issy-les-Moulineaux	0,891546011
Drniš	Drniš	Drniš	0,891546011

A névtér hasznosulása

A befektetett munka volumenét tekintve fel kell tenni azt a nagyon gyakorlatias kérdést, hogy miként hasznosul a – szakzsargon által elég riasztóan – eredményterméknek nevezett rendszer. Mely pontokon segíti, támogatja a kutató, feltáró munkát? Valóban megkönnyíti-e az információkeresést, biztosabb, hitelesebb találati halmazokhoz jut-e a felhasználó?

Visszatérő aggály a névtérhasználattal szemben, hogy körülményesebbé, lassabbá teszi a feldolgozó munkát, hiszen minden egyes névtérelemet meg kell keresni, és beemelni a karbantartó űrlapra.

Meggyőző érvekkel kell alátámasztani a névtérhasználat szükségességét, ahhoz, hogy elfogadott, sőt nélkülözhetetlen legyen használata. Négy érvet javaslunk megfontolásra:

1. A névtér hiteles adatot szolgáltat;
2. Önálló belépési, kutatási pontként szolgál;
3. Beépül a gyűjteménykezelő rendszerekbe, a leírt adatokat részévé teszi a formalizált (akár távoli) kereséseknek;
4. Kapcsolatot hozhat létre más névterekkel, intézményekkel, így a keresési hatékonyság és a források köre meghatványozódhat.

A hitelesség

Egy névtér értékét az határozza meg, hogy a szolgáltatott adatokat mennyire tekinthetjük hitelesnek. A döntő kérdés az, hogy mi, vagy ki garantálja, hogy hihető-e az, amit a névtér állít egy adott entitásról. A hitelesség garantálásának módját minden magára valamit is adó intézménynek, vállalkozásnak meg kell határoznia és ezt nyilvánosságra is kell hoznia.

A Petőfi Irodalmi Múzeum például úgy döntött, hogy személynévterét publikált, nyomtatott adatforrásokból építi és ezeket az adatforrásokat minden egyes entitásnál közli. Ez a döntés garantálja, hogy az adat ellenőrizhető, tartós forrásból származik (vs. az internet illékonysága). Ez az eljárás persze nem oldja meg az életrajzi lexikonok egymástól örökölt hibáinak a pontatlanságát, de elegendő biztonságot jelent, még az adateltérések kezelésére is.

Az egyik hitelesítő út tehát az adatforrások biztonsága. A másik, legalább ennyire fontos lehetőség, ha egy intézmény vállalja a felelősséget az általa felvett, gondozott és publikált adatokért. Tipikusan ilyenek a nemzeti közgyűjtemények adatbázisai. Bár ezek az intézmények sem tévedhetetlenek, de egy adatot hitelesnek tekinthetünk csupán azért, mert a Magyar Nemzeti Levéltár, a Magyar Nemzeti Múzeum vagy az Országos Széchényi Könyvtár közli. Ezt bizton elvárhatjuk ezektől az intézményektől.

Nyilvánvalóan a hitelesítő tényezők sorába tartozik a tudományos életben betöltött szerep is. Az információközvetítéssel foglalkozó intézeteknek is létezik szakmai validáló szerepe, amint azt az általunk is nagyra tartott Getty Research Intstitute fémjelzi.

Az információkeresés támogatása

Az adatok pontos rögzítésén alapuló adatfeltáró munka azt az elvárást támasztja az archivátorral szemben, hogy a dokumentumon szereplő alakot tükrözze a leírásban. Ezzel az eljárással azonban jelentős információs zaj keletkezik, mert az egyes entitásoknak, amelyek a valóságban azonosak, egymástól eltérő alakjai jönnek létre. Ezt a szóródást fokozza az a tényező is, hogy egyes entitások megnevezései és tartalma is megváltozhat az idők folyamán.

A névterek használata kiküszöböli a fenti problémát és hatványozottan pontosítja az információkeresés folyamatát. Fel tudjuk készíteni ugyanis a névterünket arra, hogy Bia és Torbágy esetén detektáljuk a két egykori település összeolvadását Biatorbágyban és tetszés szerint kereshessünk az egyes névalakokra, vagy akár az összes előfordulásra is. Ehhez hasonló esetek természetesen minden névtértípusban szép számmal előfordulnak, elegendő csak a személyek neveinek változataira utalnunk.

Ezt a funkcionalitást nem tudnánk elérni névtérhasználat nélkül.

Névtér azonosítás a munkafolyamatban

A pontos keresésnek nyilvánvaló előfeltétele, hogy az adott dokumentum leírásakor a rögzített metaadat azonosítva legyen egy névtérelemmel. Amennyiben ezt a lépést kihagyjuk a feltáró munkából, a fent leírt nyereséggel nem számolhatunk. Elengedhetetlen tehát, hiába lassítja a leíró tevékenységet, hogy az egyes adatbázisok létrehozásánál névtereket alkalmazzunk. Nem elhanyagolható

a feladat intellektuális része, mivel alkalomadtán az entitások azonosítása is kutatómunkát vagy alapos tárgyismeretet kíván. A pontos névtérazonosítás emeli a tudományos színvonalat és adatgazdagítással jár, mert a névtérben tárolt információk kiegészítik a dokumentumon szereplőket.

Kutatói hálózatok

A szemantikus web világában a világhálón publikált tudományos adatok részévé válhatnak azoknak a keresőszolgáltatásoknak, amelyek kiszabadítják ezeket domain függő környezetükből és információkereső hálózatok részévé tehetik őket, megsokszorozva ezáltal a felhasználók számát. Ezt a célt a névtérhasználat rendkívüli módon támogatja, amennyiben az entitások nemzetközi szabványos azonosítókkal ellátva lettek publikálva. Azok a korszerű technológiák, amelyek ezeket a szolgáltatásokat elérik és indexelik, képesek adatainkat távoli keresőkbe is bekapcsolni, és távoli találati halmazokba integrálni.

A keresési funkcionalitás mellett meg kell említeni a névterek kollaboratív építésének és felhasználásának lehetőségét is. Több intézmény együttműködésében épített és felhasznált névterekre nemzetközi példák már vannak, de hasznos lenne a hazai közgyűjteményi gyakorlatban is ebbe az irányba lépni, amint azt a Nemzeti Névtér tervezte is. Bár ezek a célok még nem valósultak meg, de a Nemzeti Levéltár névterénél kívánatos ezt az opciót is figyelembe venni.

Köszönetnyilvánítás

A tanulmány befejezése előtt a szerzők köszönetet mondanak a projektben résztvevő munkatársaiknak. Álljon itt azok névsora, akik elévülhetetlen érdemeket szereztek az MNL földrajzi névterének létrejöttében:

Simon András az MNL Informatikai és Innovációs Igazgatóságáról, *Havas Ádám* a Helion Kft-től, *Csizmadia József* és *Szalontai István* a Stratis Kft-től. Köszönet nekik!

Beérkezett: 2023. november 14.

Bánki Zsolt István
Magyar Nemzeti Levéltár
Informatikai és Innovációs Igazgatóság
Digitális Szolgáltatásfejlesztési Osztály
Osztályvezető

Szatucsek Zoltán
Magyar Nemzeti Levéltár
Informatikai és Innovációs Igazgatóság
Igazgató

Záros Zsolt
Magyar Nemzeti Levéltár
Informatikai és Innovációs Igazgatóság
Digitális Szolgáltatásfejlesztési Osztály
Vezető fejlesztő