AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024
33. Országos Informatikai Konferencia

2024. április 3–5.
Eszterházy Károly Katolikus Egyetem, Eger

AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024
33. Országos Informatikai Konferencia

2024. április 3–5.
Eszterházy Károly Katolikus Egyetem, Eger

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2024

HUN-REN
Magyar Kutatási Hálózat

NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória
Korrektúra: Danyi Melinda
Angol nyelvi lektor: Cseresnyés Dorottya

Networkshop 2024 konferencia előadásainak közleményei
Eszterházy Károly Egyetem, Eger
2024. április 3–5.

ISBN 978-615-82243-2-1
DOI: https://doi.org/10.31915/NWS.2024

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2024

Borítókép: freepik.com

TARTALOMJEGYZÉK

DOI: 10.31915/NWS.2024.3

NETWORKSHOP

Mesterséges intelligenciával támogatott adatgazdagítás a Nemzeti Levéltárban

Data enrichment supported by artificial intelligence in the Hungarian National Archives

Simon András
Magyar Nemzeti Levéltár, Digitális Szolgáltatásfejlesztési Osztály
ELTE ITDI Könyvtár és Információtudományi Tanszék, Doktorjelölt,
Monguz Információtechnológiai Kft. Ügyfélmenedzser
andras.simon@qulto.eu

Absztrakt

A Magyar Nemzeti Levéltár földrajzi névtere elsősorban a levéltári segédlet-adatbázisokban lévő földrajzi fogalmak gazdagítása céljából lett az elmúlt évek során létrehozva. A segédlet-adatbázisok levéltári dokumentumok leíró rekordjainak millióit tartalmazzák. A kezdetektől nyilvánvaló volt, hogy ekkora adattömegnél a kézi, intellektuális megfeleltetés lehetőségét mindenképpen ki kell egészíteni az automatikus megfeleltetés lehetőségével. Egyrészt az egyes levéltári dokumentumok leíró rekordjaiban a földrajzi nevek sokféle formában vannak rögzítve, másrészt ugyanaz a szó több hasonló nevű földrajzi fogalmat is jelölhet, így a dokumentumokban és a névtérben lévő névalakok hasonlósága alapján történő automatikus megfeleltetés önmagában nem elegendő. Az egyes névalakok valamilyen szintű hasonlóságán túl, a segédlet-adatbázisokban még sokféle egyéb információ áll rendelkezésre a névalak által jelölt földrajzi fogalom névtérben való azonosításához, melyek mind szerepet játszhatnak az azonosság megállapításakor. A sokféleképpen súlyozható szempontok együttes figyelembevétele mesterségesintelligencia-alkalmazás fejlesztését tette indokolttá. Az alkalmazás az azonosság valószínűségét (konfidenciaszint) százalékban fejezi ki. Több lépéses tesztelés során alakult ki az az algoritmus, mely már megbízható és a levéltár nyilvános felületein is megmutatható kapcsolatot állapít meg a névtérrekord és a levéltári rekord között.

Kulcsszavak: mesterséges intelligencia, adatgazdagítás, földrajzi tezaurusz, levéltári feldolgozás, levéltári segédlet-adatbázis

Abstract

The Geographical Namespace of the National Archives of Hungary has been created primarily to enrich the geographical terms in the finding aids of archival databases. These databases contain millions of descriptive records of archival documents. From the outset, it was clear that with such a large amount of data, intellectual matching had to be complemented by the possibility of automatic matching. On the one hand, geographical names are recorded in different forms in the descriptive records of individual archival documents, and on the other hand, the same string can denote several geographical concepts with similar names. So automatic matching based on the similarity between the names in the documents and the namespace does not seem to be sufficient. In addition to the similarity between the individual words, a wide range of other information is available in the finding aids to identify a geographical concept, and can play a role in matching the geographical terms of the finding aids and the namespace entities. The combination of these multiple weighting factors was considered necessary to develop an artificial

intelligence application. The application expresses the probability of identity (confidence level) as a percentage. The algorithm, which has been developed through several steps of testing, is now reliable and the results can be displayed on public interfaces too.

Keywords: artifical intelligence, data enrichment, geographical tesaurus, cataloguing in an archive, finding aids in archives

A különféle szakterületeken létrejött névterek általában személyneveket, testület neveket és földrajzi neveket tartalmaznak, emellett vannak olyan névterek, melyekben események vagy más fogalmak találhatók.[1] A Magyar Nemzeti Levéltárban jelenleg a földrajzi névtér kiépítésével foglalkozunk. A levéltári adatbázis földrajzi névállománya a levéltár saját földrajzinév-tezaurusza és három külső adatforrásból származó földrajzinév-állomány összetöltésével jött létre. Valamennyi adatforrást földrajzinév-rekordok alkotják, melyek egy kitüntetett névalakból, azok névváltozataiból, a fogalomhoz tartozó egyéb információkból, (földrajzi hely jellege, koordináták stb.) és a fogalmak közötti adatkapcsolatokból állnak. Belső szóhasználatunkban a névváltozatokat „term"-eknek, az egyes adatforrásokból származó névrekordokat pedig „subject"-eknek nevezzük. A földrajzi név adatbázis névalakjainak névtérbe rendezése során az egyes adatforrásokból származó földrajzi fogalmak egymással összekapcsolásra kerültek és névtérrekordokat (saját szóhasználatunkkal „entity / entitás") képeznek. Egy földrajzi névtérrekordhoz tehát egy vagy több adatforrásból származó földrajzi fogalom, azokhoz pedig, egy vagy több névalak tartozhat. A különféle adatforrásokból származó földrajzi fogalmak névtér entitásokká való összekapcsolását „entitás egyesítésnek" nevezzük.

A névtér kiépítése után megtörténhetett a levéltári segédletrekordok adatgazdagítása. Ennek keretében a névtérben található földrajzi fogalmak lettek összekapcsolva a segédletállományok leíró rekordjaiban található földrajzi nevekkel. Mind a földrajzi nevek egymással történő összekapcsolása, mind a földrajzi fogalmak a segédletrekordokban található földrajzi nevekkel történő összekapcsolása során mesterséges intelligencia elvei szerint működő algoritmusok alkalmazásba vételére került sor.

A levéltári adatbázisban összegyűjtött földrajzinév-állomány forrásául négy adatbázis lett kiválasztva:

- Geotaurusz: A földrajzi nevek az OSZK megbízásából összeállíttatott tezaurusza.
- MNL GEO: A nemzeti levéltár saját – folyamatosan bővülő – földrajzi név állománya, (alapjául a Geotaurusz 2011-es verziója szolgált).
- Engel Pál-féle állomány: A középkori Magyarország Engel Pál által szerkesztett digitális atlasza.
- A GeoNames[2] földrajzi adatbázis teljes állománya.

1 Ungváry, R. „A névtér mint kulturális szükséglet", Tudományos és Műszaki Tájékoztatás, 59(8), o. 320–326, 2012.

2 GeoNames (https://www.geonames.org/)

A névelemek adatbázisonkénti számszerű megoszlását az 1. sz. táblázat tartalmazza:

Adatforrás neve	Számosság	Adatforrás azonosítója
GeoNames	12237573	75027
MNL GEO	70849	75028
Geotaurusz	109008	75030
Engel Pál-féle állomány	24148	75031

1. táblázat: a földrajzi adatbázisba került névelemek száma adatforrásonként

A levéltárban lokális földrajzinév-állománynak az MNL GEO adatállományát tekintjük. Új földrajzi nevet csak ebbe szabad felvenni. A többi adatforrásból egyszeri betöltéssel létrejött állományban módosítás már nem történik, csak adatgazdagítás céljából jöhetnek létre kapcsolatok az egyes adatelemek között. Az egyes földrajzi névalakok preferált (preferred), vagy névváltozat (variant) státuszúak lehetnek. Egy földrajzi fogalmat egy preferált státuszú név és adott esetben egy vagy több névváltozat státuszú névalakrekord reprezentálhat egy-egy adatforráson belül.

A preferált és névváltozat státuszú földrajzi fogalmak arányát, illetve az egyes névalakokhoz tartozó névváltozatok számát a második, illetve a harmadik számú táblázat szemlélteti:

Preferált	12423616
Névváltozat	6916116

2. táblázat: Preferált és névváltozat státuszú földrajzi fogalmak aránya a levéltári adatbázisban

Névváltozatok száma	Egy névváltozat sem	Egy névváltozat	Két névváltozat	Három névváltozat	Négy névváltozat	Öt vagy annál több névváltozat
Számosság:	8254374	2982361	556750	337674	124583	167875

3. táblázat: A preferált névalakokhoz tartozó névváltozatok száma

Látható, hogy nagyszámú földrajzi fogalomhoz csak egy – preferált – névalak tartozik, és rendre csökken a száma azoknak a névalakoknak melyekhez több névváltozat van hozzákapcsolva az adatbázisban.

A földrajzi fogalom minőségét (település, hegység, folyó, épület stb.) minden földrajzi fogalom esetében egy kötelezően kitöltött adatmező tartalmazza, mely az eredeti adatforrásból érkezett. Mivel a rekordok óriási többsége a GeoNames-ből származik, ezért a fogalmat minősítő mező adattartalma is a legtöbb esetben angol nyelvű. Az azonos tartalmú angol és magyar nyelvi változatokat az adatbázisban a névtérrekordok létrehozása végett összekapcsoltuk.

Az adatbázisban lévő földrajzinév-állomány elemei hierarchikus kapcsolatban vannak egymással.

Az egyes adatkapcsolatok előfordulásának számosságát a 4. sz. táblázat szemlélteti:

Számosság	**Leírás**	**MARC-kód**
235643	Része	551 $u
235622	Átfogóbb egésze	551 $t
73488	Kapcsolatban	egyéb
16661	Kiegészítő	451 $c
16661	Helyettesített kifejezés	451 $d
13452	Korábban, kiindulása	551 $a
13450	Később, irányul	551 $b
12084	Lásd más értelemben	551 $l
5439	Helyettesített kifejezés vagylagosan	451 $y
5439	Általános altárgyszó	451 $x
1135	Általánosabban	551 $g
1135	Fajtája	551 $h
38	Helyettesített együttesen további kifejezéssel	451 $w
38	Formai altárgyszó	451 $v

4. táblázat: Az adatbázisban lévő adatkapcsolatok előfordulása

Nem minden adatforrás tartalmazott adatkapcsolatokat. Az adatkapcsolatok előfordulását az 5. sz. táblázat mutatja meg adatforrásonként. Ugyancsak az 5. sz. táblázatban szemléltetem az adatkapcsolatok közül az adatgazdagítás szempontjából legnagyobb fontossággal bíró „Átfogóbb egésze" adatkapcsolat előfordulását.

Adatforrás	**Számosság**	**Ebből „átfogóbb egésze"-típusú adatkapcsolat**
MNL GEO	280867	67266
Geotaurusz	289625	105107
Engel	59818	24148

5. táblázat: Adatkapcsolatok előfordulásának számossága adatforrásonként

Az adathierarchia mellett az egyes névelemek egymással és a levéltári segédletrekordokban lévő földrajzi nevekkel történő összekapcsolása során igen fontosak a geokoordináták is. A földrajzi nevek geokoordinátákkal való ellátottságát a 6. sz. táblázat mutatja meg:

Geokoordinátával rendelkező rekordok száma	12349247
Geokoordinátával nem rendelkező rekordok száma	92331

6. táblázat: Geokoordinátákkal ellátott földrajzi fogalmak számossága

Mivel egy entitáshoz annyi fogalom (preferált státuszú névalak) tartozik ahány adatforrásból származik, elképzelhető, hogy egy névalak az egyik adatforrásban preferált, a másikban variáns státuszú. A két névalak közti azonosságot a névtérben az entitásrekord azonosítója definiálja.

Az entitás egyesítésnél figyelembe vett szempontok:

- A névalak teljes vagy részleges egyezése.
- Hossza (minél hosszabb a névalak az azonosság annál inkább valószínű).
- Egyedisége (minél többször fordul elő egy névalak különféle földrajzi fogalmak esetében az adatbázisban, az azonosság annál kevésbé valószínű).
- A földrajzi fogalom jellege, (hegy csak heggyel, település csak településsel stb. lehet azonos).
- Geokoordináták egyezése, illetve egymástól való távolsága.

Az entitásegyesítés folyamata:

- Az adatok betöltése az adatforrásokból a közös levéltári–földrajzi adatbázisba.
- Adattisztítás, a felesleges karakterek és többletjelentéssel nem bíró stopszavak (pl. megye, hegy, domb) kiszűrése a névalakokból.
- Földrajzi fogalmak szabályalapú összekapcsolása, amennyiben teljes és kizárólagos az egyezőség az összekapcsolni kívánt földrajzi fogalmak között.
- Mesterségesintelligencia-alapú összekapcsolás, mely az előző felsorolásban szereplő adatelemek összevetésére épülő gépi tanulási modellt használ. Ezt az összekapcsolást első lépésben egy, az alkalmazás tervezői által összeállított minta (tanítóhalmaz) összeállítása és az alkalmazásba való betöltése előzte meg. A létrejött kapcsolatok esetében annak demonstrálására, hogy a gépi modell alapján a párt képező két földrajzi fogalom azonossága mennyire áll kétségen felül, egy konfidenciaszintnek nevezett százalékszámot használunk.
- Manuális validálás, (azoknak az eseteknek az emberi ellenőrzése, amikor a gépi tanulási modellre épülő összekapcsolás bizonytalan eredménnyel tért vissza) melynek eredményét figyelembe véve további entitásegyesítési lépések elvégzésre került sor.[3]

Az entitásegyesítés során entitáscsoportok jöttek létre, melyeket az egyes adatforrásokból származó, az entitásegyesítés során azonosnak talált földrajzi fogalmak alkotnak. Bizonyos szempontok és az adatforrások között megállapított hierarchia alapján, a földrajzi fogalmakat tartalmazó egyik adatforrás mindig preferáltként lett megállapítva.

Az egyes entitáscsoportok preferált adatforrás szerinti megoszlását a 7. sz. táblázat tartalmazza.

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

Adatforrás	Számosság
GeoNames	12199289
MNL GEO	3312
Geotaurusz	109006
Engel	7609
Összesen	**12319216**

7. táblázat: Entitások megoszlása preferált adatforrás szerint

Az entitáscsoportok túlnyomó része csak egy adatforrásban található meg. Ez leginkább a hatalmas tömegű külföldi névre mondható el, melyek a GeoNames adatbázisból kerültek a levéltár földrajzinév-adatbázisába. Az entitáscsoportokhoz tartozó földrajzi fogalmak számosságát (mely egytől négyig terjedhet) a 8. sz. táblázat szemlélteti.

Fogalmak száma:	Számosság:
1 fogalom	12260472
2 fogalom	23521
3 fogalom	24768
4 fogalom	10455
Összesen	**12319216**

8. táblázat: Az egyes entitáscsoportokhoz tartozó földrajzi fogalmak száma a csoporthoz tartozó fogalmak számossága alapján

A segédletdúsítás alapelvei

- A levéltári segédletadatbázis-rekordok (Ezek logikailag megfelelnek a könyvtári katalógusok bibliográfiai rekordjainak) rendszerint tartalmaznak földrajzi adatokat, melyek a levéltári adatbázisok adatszerkezetének megfelelően különféle adattáblák különféle mezőiben találhatók.
- A földrajzi névtérben azokkal a névalakokkal kerülnek megfeleltetésre, amelyek a levéltári dokumentumban, és az azt leíró segédletrekordban is megtalálhatóak. A névalakot a segédletrekordban egyéb adatok, (fölérendelt földrajzi egység, a földrajzi fogalom fajtája) egészíthetik ki.
- A segédletdúsítás során a névtérben szereplő minden földrajzi nevet, vagyis az összes, az adatbázisban fellelhető névváltozatot fel lehet használni.
- A megfelelő névtérelem kiválasztását preferencia-sorrend (Geotaurusz, MNL GEO, GeoNames, Engel) támogatja. A nyilvános felületen történő megjelenítéskor is ebben a sorrendben jelennek meg a találatok.

A segédletdúsítás folyamata módszertani és műszaki tekintetben megegyezik az entitás egyesítésével. Itt is sor kerül tanítóhalmaz kialakítására, validáló halmazok képzésére, validálásra, a párok azonosságának százalékszámmal kifejezett valószínűségének (konfidenciaszintnek) a megállapítására, és a párosítás újabb és újabb paraméterekkel történő egyre jobb eredménnyel (több és magasabb konfidenciaszintű találattal) kecsegtető lefuttatására. A segédletdúsítás eredményét a 9. sz. táblázat szemlélteti:

Entitás preferált adatforrása	Összes segédlet rekordban párra talált entitás	Összes pár	Átlagos konfidenciaszint
Engel	16454	3568799	0,512206
GeoNames	11068	3651867	0,547161
Geotaurusz	12522	1024033	0,407073
MNL GEO	11068	987856	0,40377

9. táblázat: Párra talált névtérrekordok (entitások) és segédletrekordokkal képződött párjaik száma és a párok átlagos konfidencia szintje az entitás preferált adatforrása szerint

A mesterséges intelligenciát alkalmazó algoritmusok alapján felépített szoftver használata során megállapíthattuk, hogy ilyen komplex szempontrendszer alapján megbízható szabályalapú összekapcsolási eljárást nem tudtunk volna kidolgozni, mert az egyes szempontokat:

- névalak hasonlósága,
- előfordulásának gyakorisága és a névalak hossza,
- a geokoordináták megléte és egymástól való távolsága,
- a földrajzi fogalmak jellege

egymással csak önkényesen tudtuk volna emberi intelligenciával összemérni.

A névtér építése során az előálló legnagyobb nehézséget mindamellett az emberi ésszel felfoghatatlan és még kiválasztott szegmenseiben is áttekinthetetlen óriási adattömeg, illetve a segédletrekordokban levő nem elegendő információ jelentette. Az óriási adattömeg vizsgálata kezdetekben nem történhetett meg hipotézisek nélkül, ezek azonban nem kellő körültekintés esetében jelentősen befolyásolhatják a végeredményt. Ezért a mesterséges intelligenciára épülő alkalmazás elkészítése és hatékony, – a kezdeti tanítóhalmaz és a több lépésben történő kézi validálás alapú modellépítés támogatása mindenképpen jó választásnak bizonyult.