https://doi.org/10.3311/tmt.13260

Wikidata, a többnyelvű közösségi közreműködésen alapuló tudásgráf – 1. rész

Egy szakirodalmi és strukturális áttekintés

Molnár Bence

A kétrészes tanulmánysorozat a Wikidata felépítését, szemantikus weben betöltött szerepét és lehetséges alkalmazási módozatait vizsgálja nemzetközi példák bemutatása és szakirodalmi áttekintése által. A szándék kettős: magyar nyelven íródott szakirodalom hiányában átfogó és teljes áttekintést nyújtani a Wikidatáról, összefoglalva a tudásgráffal kapcsolatos legfontosabb kutatási eredményeket, másrészről bemutatni a Wikidata integrálásának megvalósult jógyakorlatait és lehetséges alkalmazásait, különös tekintettel a közgyűjteményi szereplők szempontjából. Az 1. részben feltárásra kerülnek mindazon kezdeti fejlesztési célok és lépések, amelyek az adatbázis ma ismert formájához vezettek. Az adatok RDF tripletként történő tárolásához szükséges szemantikus struktúra meghatározását követően bemutatásra kerül a Wikidata közösségi közreműködésen alapuló modellje és a többnyelvű adatok kezelésének mikéntje.

besorolási adat, besorolási adatok egységesítése, gépi dokumentumleírás, többnyelvű kapcsolt nyílt adatok, szemantikus web, tudásgráf, Wikidata

1. Bevezetés

A Wikidata egy közösségi közreműködésen alapuló többnyelvű tudásgráf. Szemantikus adatbázisként a világ entitásait egy-egy elem (item) azonosítja, amelyek az emberi tudás számos spektrumát – a tudománytól kezdve a kultúrán át a művészetekig – lefedik. A Wikidata összehasonlítható, megbízható és könnyen hozzáférhető adatokat tartalmaz, amelyeket a projekt önkéntes közreműködői közösen szerkesztenek és fejlesztenek. A Wikipédia társprojektjeként és adatszolgáltatójaként a Wikidata is hozzájárul az „open knowledge” (nyitott tudás) mozgalom céljaként megfogalmazott szabad, bárki számára egyenlő és hozzáférhető tudás megteremtéséhez. Dokumentum-orientált adatbázisként minden elem egy „Q” prefixet követő egyedi azonosítóval rendelkezik, így például az Eötvös Loránd Tudományegyetemet a „Q390287” [1], míg Magyarországot a „Q28” azonosítójú elem reprezentálja [2]. A tanulmány lezárásának időpontjáig (2023. augusztus 18.) a Wikidata több mint 1,96 milliárd szerkesztést és 106,14 millió tartalmi lapot [3] (elemet) tartalmazott [4].

A Wikidata hazai publikációk formájában szinte alig került még feldolgozásra, noha a témával kapcsolatos kutatások döntő részét – a későbbiekben ismertetett tényezők miatt – európai személyek és intézmények publikálják [5]. A szerző előzménynek tekinthető tanulmánya „A Wikidata és a Nemzeti Névtér kapcsolódási lehetőségei” címmel 2021-ben jelent meg a Könyvtári Figyelő szakfolyóiratban. Az akkor még béta állapotban lévő (Magyar) Nemzeti Névtér és a Wikidata lehetséges integrációjának aspektusait vizsgáló írás hazai és nemzetközi trendek bemutatásával kívánta katalizálni a formálódó együttműködést [6].

A tanulmánysorozat alapos szakirodalmi áttekintést és a téma legfontosabb műveinek és hivatkozásainak szintetizálását igyekszik nyújtani a laikus és szakavatott olvasónak egyaránt, abban a reményben, hogy ezzel elősegíti a Wikidata alkalmazásának elterjedését a magyarországi könyvtári közbeszédben. A sorozat azon problémafelvetésre keresi a választ, hogy a szemantikus web napról-napra szövevényesebb hálózatában milyen szerepet is tölt be a Wikidata, a mindenki által szabadon szerkeszt-

Molnár, B. *Wikidata, a többnyelvű közösségi közreműködésen alapuló tudásgráf – 1. rész: Egy szakirodalmi és strukturális áttekintés*, Tudományos és Műszaki Tájékoztatás, 70(3), p. 316–333, 2023. https://doi.org/10.3311/tmt.13260

hető és (újra)felhasználható többnyelvű, közösségi közreműködésen alapuló tudásgráf. Az első rész átfogó és teljes áttekintést kíván nyújtani a Wikidatáról, mint tudásgráfról és a projekttel kapcsolatos legfontosabb kutatási eredményekről. A Wikidata történeti áttekintése során feltárásra kerülnek a kezdeti fejlesztési célok és lépések, amelyek a weboldal ma ismert formájához vezettek. Az adatok Resource Description Framework (Erőforrás Leíró Keretrendszer, RDF) tripletként történő tárolásához szükséges szemantikus struktúra áttekintéseként definiálásra kerülnek az adatbázist felépítő elemek, tulajdonságok, értékek, minősítők, forráshivatkozások és azonosítók.

2. Történeti áttekintés

A Wikidata 2012. október 29-én indult. A projekt fejlesztésének nagyját a Wikipédiát is üzemeltető amerikai Wikimedia Foundation (WMF) németországi társszervezete, a Wikimedia Deutschland végezte. 2006 és 2012 között ezzel a Wikidata volt a WMF egyetlen újonnan indított projektje [7]. A fejlesztést a Microsoft társalapítója, *Paul Allen* által megálmodott Allen Institute for AI, a Gordon and Betty Moore Foundation (az Intel társalapítójának és feleségének alapítványa), valamint a Google összesen 1,3 millió euróval támogatta [8].

A Wikidata termékmenedzseri pozícióját betöltő, német származású informatikus *Lydia Pintscher* vezetése alatt a fejlesztők egyik kitűzött célja az volt, hogy javítsák a különböző nyelvű Wikipédiákban és más projektekben szereplő adatok minőségét azáltal, hogy átjárhatóvá és felhasználhatóvá teszik azokat valamennyi projekt között. Ezen elv a fejlesztés különböző fázisában más és más megközelítést jelentett. Az első lépés keretében a Wikidata az akkor több mint 280 nyelven elérhető Wikipédia centralizált csomópontjává vált, lehetővé téve, hogy az egy adott témában született szócikkek „interwiki" és „interlanguage" (wiki- és nyelvközi) hivatkozásai egy helyen kerüljenek tárolásra [9]. A Wikidatát megelőzően ugyanis a Wikipédia különböző nyelven született szócikkeinek egymásra mutató hivatkozásait manuálisan kellett karbantartania a szerkesztőközösségnek, ami – figyelembe véve, hogy a Wikidata indulásakor 56,4 millió szócikk létezett több mint 280 nyelven [10] – belátható módon hamar meghaladta a humánerőforrás-kapacitás limitjét. A gyakorlatban ez azért sem volt kivitelezhető, mert egy entitás különféle nyelveken írt szócikkeinek hivatkozásai mindig az adott nyelvű projekt szócikkének törzsszövegében tárolódtak; amennyiben egy Wikipédia-szerkesztő rendszeresen manuálisan nem ellenőrizte, hogy az összes felsorolt nyelvi hivatkozás működik (nem mutat még nem létező vagy már törölt helyre) és a lista valóban teljes (az entitás valamennyi nyelven létező szócikkét tartalmazza), a szerkesztőközösség és az olvasóközönség nem lehetett biztos benne, hogy minden lehetséges hivatkozás feltárásra került. Könnyen belátható tehát, hogy az akkori helyzet súlyos információvesztéssel járt, irreális terhet róva a szócikkeket író Wikipédia-szerkesztőkre. A megoldás logikus módon egy olyan központi adatbázis létrehozása volt, ahol a Wikipédiák közötti nyelvközi hivatkozásokon végzett szerkesztések közvetlenül megjelennek valamennyi Wikipédia-projekt szócikkében, így tehát egy frissen létrehozott szócikk „bekapcsolását" elég egyszer elvégezni. Az *1. ábra* a nyelv- és wikiközi hivatkozások Wikidata előtti állapotát mutatja: balra szerkesztői nézetben látható, hogy a különböző nyelvű Wikipédiákat az egységes erőforrás-helymeghatározójukban (Uniform Resource Locator, URL) is szereplő – ISO 639 szabványon alapuló – nyelvkódok azonosítják, majd egy kettőspont után következik az adott entitásról szóló szócikk címe (input), míg jobbra az olvasók nyelvközi navigálását segítő lista (output).

Az első fázisban kitűzött cél – egy központi repozitórium létrehozása a Wikimedia Foundation projektjeinek nyelv- és wikiközi hivatkozásainak tárolására – 2013. január 14-én lépett a megvalósulás útjára, amikor a magyar Wikipédia az első olyan projektté vált, ahol ezen információk már közvetlenül a Wikidatából töltődtek be [11]. Az utolsó Wikipédia 2013. március 6-án, míg a Wikimedia Commons 2013. szeptember 23-án állt át a Wikidata-szolgáltatta hivatkozások használatára [12, 13]. Az átállás természetesen a milliós szócikkszám okán nem zajlott teljesen zökkenőmentesen: mivel kikötés, hogy egy Wikidata-elem projektenként csak egy hivatkozással rendelkezhet (tehát például a Q42-es elem nem mutathat egyszerre a magyar Wikipédia „Douglas Adams" és „Budapest" szócikkére), több entitás szócikkének linkelése akadályba ütközött. Az akkor és a mai napig is előkerülő prob-

1. ábra Nyelvközi hivatkozások tárolása és reprezentációja a Wikidata bevezetését megelőzően: bal oldalt az angol Wikipédia „Wikidata” című szócikkének nyelvközi hivatkozásai láthatóak szerkesztői nézetben, míg jobbra a felhasználók által látott eredmény. Szerző: Gareth Edwards, CC-BY-SA 3.0 és GNU 1.2 licenc alatt; Forrás: https://commons.wikimedia.org/wiki/File:Interlanguage_links_prior_to_Wikidata.png

lémák vezetésére „Interwiki conflicts” (wikiközi hivatkozások konfliktusa) néven dedikált lapot hozott létre a szerkesztőközösség. Itt tapasztalt szerkesztők segítségét lehet kérni az azonos témát lefedő elemek összevonásához: például amikor két vagy több olyan elem is létezik, amely látszólag ugyanahhoz az entitáshoz vagy Wikimédia-oldalhoz tartozik, illetve amikor egy Wikipédia-szócikk egy fogalom tekintetében részletesebb, mint egy másik nyelvű, és így két Wikidata-elem készült hozzájuk [14].

A nyelv- és wikiközi hivatkozások integrációja mára szinte teljeskörűen megvalósult, a társprojektek újonnan létrehozott lapjai a szerkesztők által manuális vagy félautomata módon, bizonyos esetekben automatikusan bekötésre kerülnek a Wikidata adatbázisába. Ezen (kereszt)hivatkozások centralizált tárolása azonban nem csak azért hasznos, mert ezáltal a felhasználók kényelmesen navigálhatnak egy adott entitás különböző nyelven írt szócikkei között, hanem azért is, mert áttekintéssel szolgálhatnak például a Wikipédia nyelvváltozatainak tartalmáról. Az *1. táblázatban* példaképp vett projektek tartalmára vonatkozóan több érdekes következtetést is levonhatunk csupán a nyelv- és wikiközi hivatkozások adataiból. A nyers adatokban felfedezhetők például a nyelvek és nyelvváltozatok kulturális és politikai kapcsolatai: míg a projektek jellemzően az angol Wikipédiára hivatkoznak a legtöbbször (elemszámukat tekintve azzal állnak a legnagyobb átfedésben), addig a posztszovjet régiók nyelveinek esetében megfigyelhető az orosz Wikipédia felülreprezentáltsága. Távolabb tekintve, az indoárja urdu nyelv – amely Pakisztán és India egyes államainak hivatalos nyelve, előbbiben egyben lingua franca is – műszaki és irodalmi szókincse nagyban perzsa–arab eredetű [15], így talán nem meglepő, hogy az urdu Wikipédia második és harmadik legtöbbet hivatkozott projektje

1. táblázat Különböző Wikipédia-projektek nyelv- és wikiközi hivatkozásainak statisztikái a 2023. augusztus 20-ai állapotoknak megfelelően. Elemnek tekintendő egy Wikipédia minden olyan lapja – szócikk, portál, sablon stb. –, amely Wikidata-elemhez van kapcsolva. A hivatkozások átlagos száma az elemek Wikidata-adatlapjain szereplő nyelv- és wikiközi hivatkozások átlagát, míg az árva elemek a más projektre nem linkelő (csupán egy hivatkozást tartalmazó) elemek százalékát adja meg. A legtöbbet hivatkozott projektek sorrendje azok elemeinek metszethalmaza alapján kerül felállításra. A szerző saját gyűjtése és szerkesztése; Forrás: https://www.wikidata.org/wiki/User:Pasleim/Connectivity

Wikipédia-projekt	Elemek száma	Hivatkozások átlagos száma	Árva elemek százaléka	Legtöbbet hivatkozott három projekt		
angol	9 418 985	4,6	29,9%	frwiki	commons	arwiki
magyar	625 158	21,5	14,7%	enwiki	frwiki	itwiki
örmény	383 579	23,9	13,7%	ruwiki	enwiki	ukwiki
nyugat-örmény	14 747	55,5	12,4%	hywiki	ruwiki	enwiki
urdu	694 452	13,8	6,1%	enwiki	arwiki	fawiki

épp az arab és a perzsa Wikipédia. Hasonló a helyzet a nyelvváltozatok terén is: a nyugat-örmény nyelv például a legtöbbet az örmény és az orosz, majd ezt követően az angol Wikipédiára hivatkozik. A Wikidata több mint 616 ezer olyan elemet tartalmaz, ami egy, a magyar Wikipédiára mutató hivatkozással is rendelkezik. Ezen elemek átlagosan 21,4 hivatkozással rendelkeznek: körülbelül 91 ezerre tehető azon elemek száma, ahol a magyar Wikipédia az egyetlen hivatkozás az elemen, tehát az adott entitásról más projekten még nem született szócikk (feltételezhetően a téma kultúrkör- és nyelvspecifikussága okán) [16].

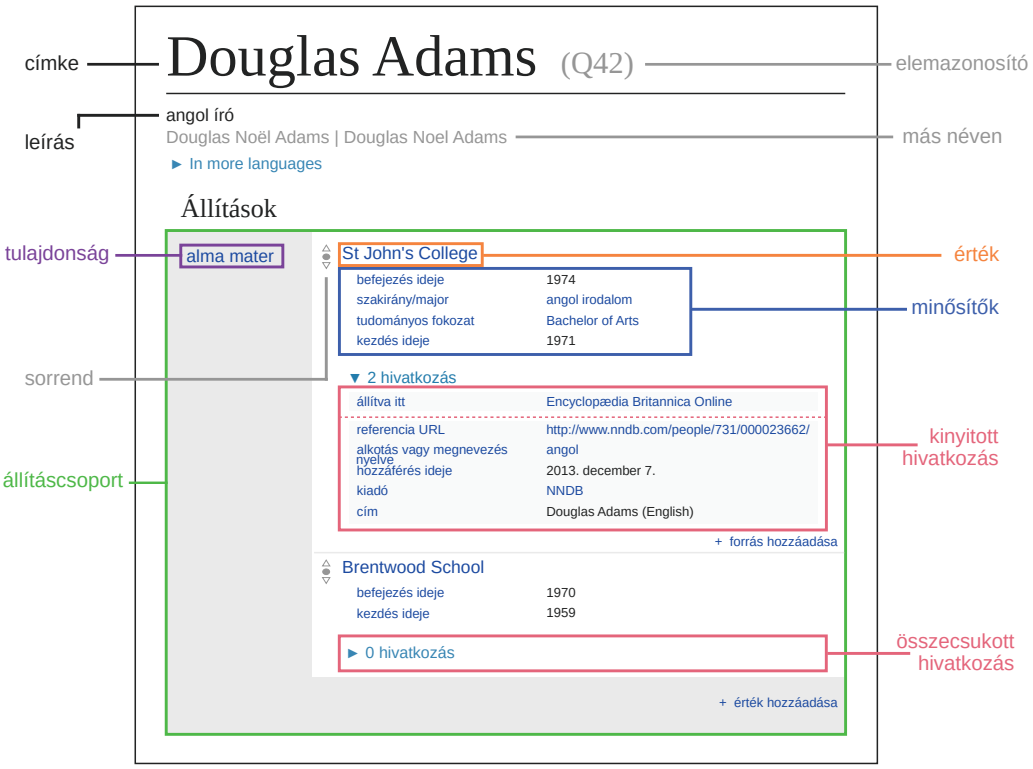
A második fázis célja egy központi adattér létrehozása volt, ahol az úgynevezett „infobox" sablonok információi egy helyen tárolódhatnak, hozzáférést biztosítva azokhoz valamennyi Wikipédia számára. Definícióját tekintve az infobox egy olyan strukturált dokumentum (jellemzően fizikai vagy digitális tábla/táblázat), amely egy adott témával kapcsolatos információk egy részhalmazának összegyűjtésére és bemutatására szolgál attribútum–érték párok segítségével [17]. A Wikipédia esetében a szócikkek bevezetőiben – az adott nyelv írásrendszerének megfelelően, jellemzően jobb oldalt – megtalálható infoboxok a téma rövid összefoglalására szolgálnak, életrajzi szócikkek esetében könnyen elérhetővé téve az olvasóknak például az adott személy születési idejét és életkorát, de az általa viselt közhivatal vagy beosztás információt is (tisztség neve, mettől meddig, elődje és utódja, stb.). Mindezen adatok az infobox sablonparamétereinek értékeként tárolódnak [18]. Azáltal, hogy ezeket az adatokat nem különféle projekteken elszórt módon, hanem egy centralizált térben tároljuk, a nyelvközi hivatkozásokhoz hasonlóan könnyebb karbantarthatósággal és az információk jobb ellenőrizhetőségével jár. Mindemellett, ha egy cikk alapvető adatait egy strukturált formájú infoboxban összefoglaljuk és elérhetővé tesszük, lehetővé válik azok számítógépbarát módon történő feltárása, megnyitva ezzel az utat az infoboxokban tárolt állítások ontológiák általi automatizált módon RDF tripletté történő alakítása előtt [17].

A második fázis keretében 2013. február 4-én váltak elérhetővé az állítások a Wikidatán. Innentől kezdve a Wikidata már nem csak nyelv- és wikiközi hivatkozások, hanem tulajdonságokhoz társított értékekből felépülő állítások tárolására is képes. Kezdetben egy tulajdonság értéke csakis egy Wikidata-elem vagy egy Wikimedia Commonsban tárolt médiafájl lehetett, ez az adattípus-megkötés azonban hamar megszűnt, lehetővé téve string (karakterlánc), koordináta és dátum adattípus megadását is [19]. Az állítások megjelenésével a Wikidata teljes értékű tudásgráffá vált, amely túllépve kezdeti – társprojektjeire korlátozódó – klientúráját rövid idő alatt megkerülhetetlen szereplője lett a szemantikus adatbázisok piacának, tartalmával behálózva internetes mindennapjaink valamennyi aspektusát.

3. Strukturális felépítése

Az entitásokra vonatkozó információk a szemantikai szabványokkal összhangban állításokként tárolódnak. A következőkben *Douglas Adams* (Q42) [20] angol író példáján szemléltetem egy Wikidata-elem felépítését (a vizuális reprezentáció a *2. ábrán*).

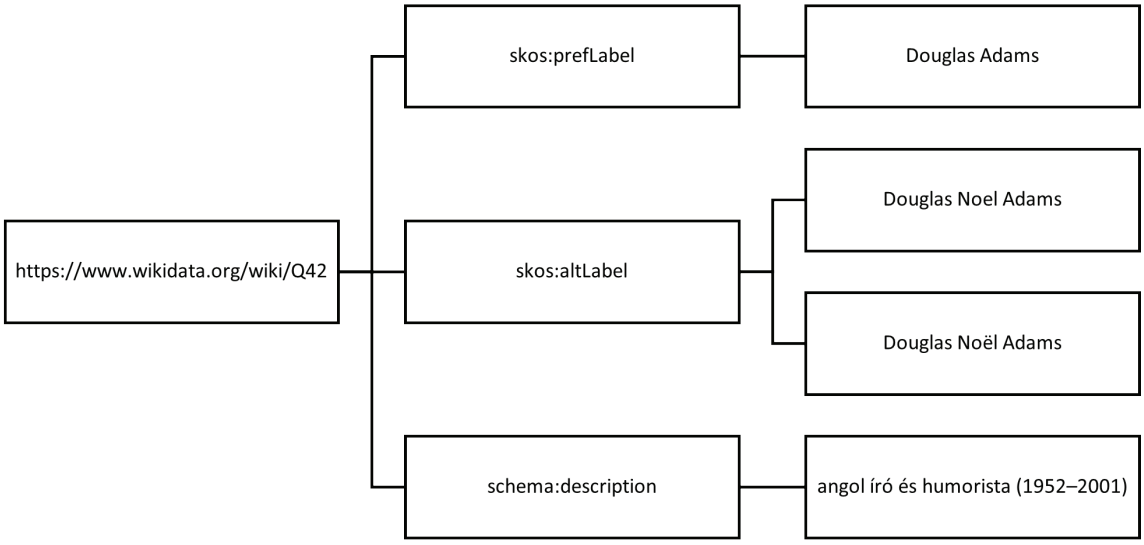
Egy elem visszakeresését és azonosítását nyelvi szempontból egy elsődleges címke (label), egy leírás (description), valamint az alternatív címkék (alias) teszik lehetővé. A címke az elem által leírt entitás neve, lehetőség szerint olyan alakban, ami a leginkább közismert/használt. Kereséskor a Wikidata a keresőkifejezésünkhöz leginkább illeszkedő címkéjű találatokat kínálja fel. Egy elemnek egy nyelven csak egy címkéje lehet. Az egyértelműsítést szolgálja az a szabály, hogy egy adott nyelven ugyan szerepelhet több azonos címkéjű elem az adatbázisban – a homonímiák mellett például gyakoriak az azonos nevű személyek is –, azonban a megkülönböztethetőség érdekében a leírások tartalmukat tekintve nem egyezhetnek meg [21]. A Q42-es elem esetében a magyar nyelvű címke Douglas Adams. A leírás elsődleges feladata az entitás legfontosabb ismérveinek összefoglalása egy mondatban, ezáltal az egyértelműsítés az azonos és hasonló címkéjű Wikidata-elemek között. Azonos nevű települések esetében a leírás segít beazonosítani, hogy Cambridge esetén az angliai (Q350) [22] vagy a massachusettsi (Q49111) [23] településről van szó. A leírások azonban nem definíciók (bár egyes esetekben olyan minőségűek is lehetnek), az entitást tartalmilag az állítások hivatottak reprezentálni [24]. A *2. ábrán* a Q42-es elem magyar nyelvű leírása „angol író", míg a *3. (az elem legfrissebb állapotát tükröző) ábrán*



2. ábra A Wikidata adatmodelljének grafikai reprezentálása Douglas Adams példáján keresztül. Szerző: Charlie Kritschmar (WMDE), Samat, CC0 1.0 licenc alatt; Forrás: https://commons.wikimedia.org/wiki/File:Datamodel_in_Wikidata_hu.svg

„angol író és humorista (1952–2001)". Az alternatív nevek vagy címkék az elsődleges címkékhez hasonlóan kereshetők és nyelvspecifikusak, azonban a címkékkel ellentétben számuk nem korlátozott, egy elem rendelkezhet több tucat alternatív névvel is, meglétük nélkül is teljesnek tekinthető egy elem. Leggyakrabban a születéskori nevek (példánk esetében „Douglas Noel Adams" és „Douglas Noël Adams"), becenevek és művésznevek, a biológiában fajok tudományos nevei, valamint a rövidítések és a betűszavak fordulnak elő [25]. A nyílt szabványokat fejlesztő World Wide Web Consortium (W3C) ajánlásainak megfelelően a címkék jelölése az „rdfs:label" állítmánnyal (predicate) történik. A tárgy (object) nyelvét egy „xml:lang="NYELVKÓD"" tag jelöli, ahol a nyelvkód a korábban már említett ISO 639 szabványon alapul. A Q390287-es elem magyar nyelvű címkéjét tartalmazó egyszerű kijelentő mondat (triplet) az alábbi módon néz ki: „<rdfs:label xml:lang="hu">Eötvös Loránd Tudományegyetem</rdfs:label>" [1]. Példánkhoz visszatérve, a Q42-es elem magyar nyelvű elsődleges és alternatív címkéje, valamint leírása RDF-formátumban a *3. ábrán* látható tulajdonságok által tárolódik.

Az entitásra vonatkozó információk állítások (statement) formájában, tulajdonságok (property) értékeiként (value) tárolódnak: a Wikidata ezen alapvető struktúrája (elem/item – tulajdonság/property – érték/value) nagyban hasonlít egy RDF tripletre. A tulajdonságok az elemekhez hasonlóan egyedi – „P" prefixumú – azonosítóval rendelkeznek. Egy tulajdonság három típusú értéket vehet fel: az érték lehet ismeretlen, egyedi vagy hiányzó. Leggyakrabban az egyedi értékek fordulnak elő, amelyek lehetnek kvantitatív értékek (például népességi adatok vagy dátumok) vagy Wikidata-elemek. Utóbbi útján jönnek létre az elemek között a hivatkozások. Ismeretlen értékkel rendelkezhet egy tulajdonság, ha például egy személy halálozási helye nem ismert, vagy ha az adott entitásnak még nincs eleme a Wikidatán. Végül bizonyos esetekben szükséges lehet a hiány puszta jelzése is: mivel *I. Erzsébet* angol királynőnek (Q7207) [26] nem született gyermeke, így a gyermek (P40) [27] tulajdonság esetében az érték hiányának indikálása több információt és teljesebb kontextust hordoz, mintha egyáltalán nem rögzítettünk volna semmit [28].



3. ábra A Q42-es azonosítójú Wikidata-elem magyar nyelvű elsődleges és alternatív címkéinek, valamint leírásának RDF-alapú reprezentációjának strukturált ábrája. A szerző saját gyűjtése és szerkesztése; Forrás: https://www.wikidata.org/wiki/Special:EntityData/Q42.rdf

A Wikidata valamennyi állítása a projekt közel 10,900 tulajdonságának értékeként szerepel az adatbázisban [29]. A Wikidata tulajdonságainak számos téma szerint (manuálisan) válogatott és adattípus szerint (automatikusan) készített listája is a szerkesztők rendelkezésére áll [30, 31]. A *2. táblázat* a személyek – így például Douglas Adams – leírásához leggyakrabban használt tulajdonságokat tartalmazza. A tulajdonságok egy közösségi megbeszélés folyamán születnek, amelynek keretében a közösség előbb véleményezi a létrehozandó tulajdonság javasolt paramétereit (nevét, adatelemeinek formátumát reguláris kifejezésként, formázó URL-jét stb.) és relevanciáját a projekt szempontjából, majd konszenzusos döntés keretében engedélyezi vagy utasítja el annak létrehozását [32]. A tulajdonságok bizonyos esetekben törlésre is kerülhetnek, jellemzően azonban ilyenkor adatutódlás keretében egy másik tulajdonság megörökli a közösség által törlésre kijelölt tulajdonság értékeit (például olyankor, amikor egy tulajdonságot elavultsága okán a közösség lecserél

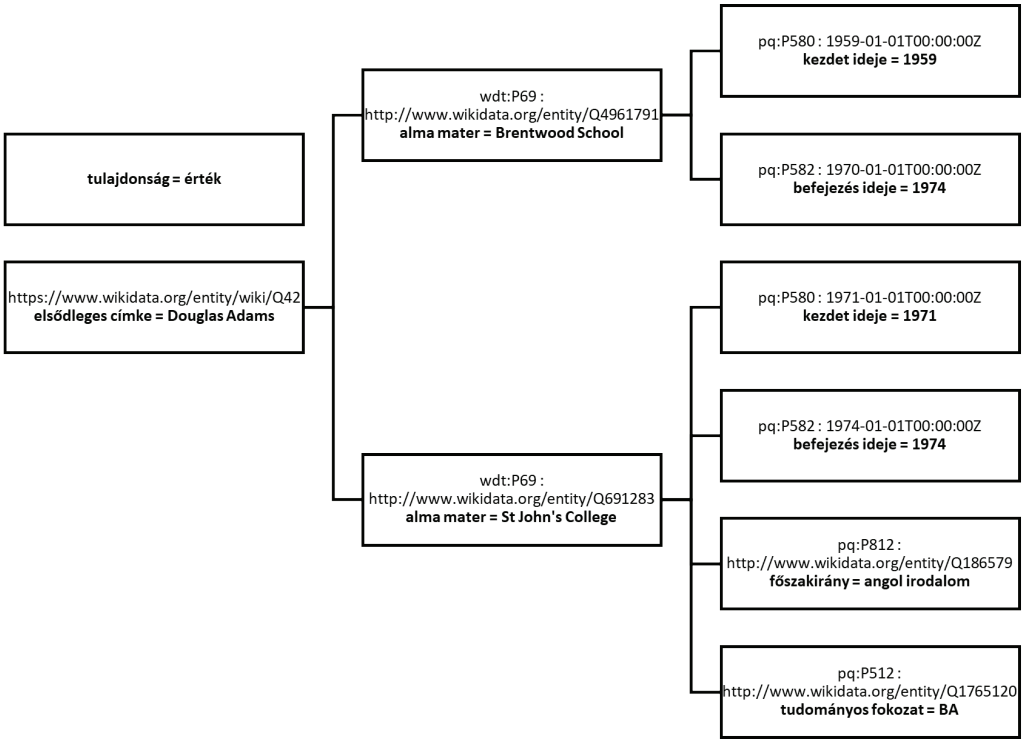
2. táblázat A személyeket leíró (osztály, amelynek példánya [P31] = ember [Q5] állítással rendelkező) Wikidata-elemeken szereplő tíz leggyakrabban előforduló tulajdonság magyar neve, azonosítója, tárolt értékének leírása és előfordulási gyakorisága a vizsgált sokaságon. A szerző saját gyűjtése és szerkesztése; Forrás: SQL-lekérdezés [36]

Tulajdonság	**Tárolt érték**	**Előfordulási gyakoriság**
osztály, amelynek példánya (P31)	ember (Q5)	100%
nem (P21)	a személy nemi identitása vagy társadalmi neme	79,5%
foglalkozás (P106)	szakma és/vagy munkahelyen végzett foglalkozás	72,2%
utónév (P735)	keresztnév	65,0%
születési idő (P569)	az alany születési dátuma	56,5%
állampolgárság (P27)	azon ország(ok), amely(ek)nek állampolgára az illető	41,6%
családnév (P734)	vezetéknév	40,0%
születési hely (P19)	a születés legspecifikusabb helye (pl. ország helyett település, város helyett kórház)	30,7%
halálozási idő (P570)	a halál bekövetkeztének dátuma	28,8%
VIAF-azonosító (P214)	a Nemzetközi Virtuális Katalógustár azonosítója	25,3%

egy másikra) [33]. Amikor 2023 februárjában a béta verziószámot elhagyva bemutatkozott a Magyar Nemzeti Névtér (MNN) [34], a Wikidatában addig szereplő – P6988-as azonosítójú tulajdonsághoz tartozó – mintegy 9000 személyazonosító linkromlás áldozatává vált, mivel az MNN-ben történt változtatások miatt az adatbázis entitásaihoz új azonosítók kerültek hozzárendelésre. A Wikidata szerkesztőközössége a probléma észlelését követően az MNN korábbi tulajdonságait elavult rangúvá állította (ezzel géppel olvasható módon is jelezve a tulajdonság használatának ellenjavallatát), és új tulajdonságok létrehozásáról döntött [35].

Az állítások továbbá minősítőkkel (qualifiers) és forráshivatkozásokkal (references) is elláthatók. Hasonlóképp az állításokhoz, a minősítők és a forráshivatkozások is legalább egy tulajdonság és egy érték párjából állnak. Előbbiek lehetővé teszik az állítások kiterjesztését és kontextusba helyezését azon túl, amit egy szimpla tulajdonság–értékpár ki tud fejezni. Mivel egy Wikidata-elem tulajdonsága akár több értékkel is rendelkezhet, a minősítők teremtik meg az értékek közötti disztingválás lehetőségét: így például a települések népességadatainál az adatfelvétel időpontja és módja teszi lehetővé az idő múlásával növekvő információ értelmezését [37]. Példánk esetében az, hogy Douglas Adams diplomáját a St John's College angliai felsőfokú intézményben szerezte, a tanulmányokra vonatkozó tulajdonság értékeként tárolódik, míg annak részletei – 1971-től angol irodalmat tanult, és 1974-ben szerzett BA diplomát – minősítőként kapcsolódnak az értékhez (lásd *4. ábra*).

A forráshivatkozások megléte a Wikidata-elemek állításainak ellenőrizhetőségét szolgálja. Hasonlóképpen egy Wikipédia-szócikkhez vagy egy tudományos publikációhoz, az információk visszakereshetősége és ellenőrizhetősége érdekében szükséges feltüntetni a felhasznált forrásokat. Ezek lehetnek könyvek, tudományos közlemények, papíralapú és online újságcikkek, weboldalak, jogszabályok és dokumentációk, de a Wikidata számos egyéb formájú forrás kezelésére is képes. A források azért is kulcsszereplői a Wikidata gyarapodásának, mert az információk jellemzően nagy mennyiségben, automatizált módon kerülnek az adatbázisba, tehát ezek egyben az adatok származási helyét is jelzik [38, 28]. Felépítésüket tekintve



4. ábra A Q42-es azonosítójú Wikidata-elem „alma mater” (P69) tulajdonságának és minősítőinek értékeinek RDF-alapú reprezentációjának strukturált ábrája. A szerző saját gyűjtése és szerkesztése;
Forrás: https://www.wikidata.org/wiki/Special:EntityData/Q42.rdf

a korábban már említettek szerint a forráshivatkozások is tulajdonság–értékpárokból állnak. Példánk esetében Douglas Adams felsőfokú tanulmányait két forrás is alátámasztja: míg az első egy konkrét – Wikidata-elemmel is rendelkező – műre, az Encyclopædia Britannica online kiadására mutat, addig a második egy weboldalra, aminek bibliográfiai adatait – cím, kiadó, URL, hozzáférés ideje, nyelv – egy-egy tulajdonság–értékpár tárolja.

Az állítások egy specifikus csoportja az azonosítók (identifier), melyek azonosítják az adott entitást egy külső adatbázisban vagy egységesített besorolási rendszerben. A Wikidata is kínál ilyen azonosítókat: a korábban már említett „Q" és „P" prefixumú elem- és tulajdonságazonosítók ugyanis nem csak a tudásgráfban töltenek be fontos szemantikai szerepet, hanem egyben permanens egységes forrásazonosítókként (Uniform Resource Identifier, URI) is szolgálnak: az entitások azonosítói globálisan egyedi URI-ként is használhatók a http://www.wikidata.org/entity/AZONOSÍTÓ elérési út „AZONOSÍTÓ" helyére behelyettesítve. Mindez könyvtári szempontból azért is fontos, mert a Wikidata tartalma ezáltal válik automatizált módon hozzáférhetővé, ugyanis az URI-k alkalmazása segíti az authoritykontrollt, az adatok RDF-alapú konverzióját, a külső adatok felhasználása pedig lehetővé teszi továbbá a gyorsabb katalogizálást és a részletesebb authority-leírások készítését [39]. A Wikidata egyben különböző adatbázisok azonosítóinak központi csomópontjaként (hub) is működik, ugyanis nyelvi korlátok nélkül teszi lehetővé azok kereszthivatkozások általi összekapcsolását [6]. Az azonosítók értékei különfélék lehetnek: ezek jellemzően mutathatnak egy külső adatbázis vagy egységesített besorolási jegyzékszámra- – például a Nemzetközi Virtuális Katalógustár, a Virtual International Authority File (VIAF) [40] – vagy betűalapú – például a nyelveket jelölő ISO 639-3 szabvány – azonosítókra, de ezek kombinálásával létrehozott stringekre – például DOI-kra [41] – is.

Állítások gyakran többféle értékkel is rendelkezhetnek, így például II. Erzsébet brit királynő Wikidata-elemének gyermek (P40) tulajdonságánál mind a négy gyermeke értékként szerepel [42]. Hogy egy tulajdonság csak egy, vagy esetenként több értékkel is rendelkezhet, azt a tulajdonsághoz kapcsolódó kikötések határozzák meg: míg egy személy több gyermekkel és keresztnévvel is rendelkezhet, ezáltal pedig ezek tulajdonságai is több értéket vehetnek fel, addig adatbázisokban jellemzően egy entitáshoz csak egy azonosító tartozhat, így például a különböző azonosítók tulajdonságainak sajátossága az egyetlenérték-kikötés. Ahol egy tulajdonság különböző értékekkel rendelkezik, szükség lehet azok rangsorolására. Alapértelmezetten minden állításhoz normál rang kerül hozzárendelésre: a semleges címke jelzi, hogy az érték pontossága és valódisága nem került értékelésre.

Több értéket tartalmazó állítások esetében normál rangot kapnak azok az értékek, ahol nincs értelme azt jelezni, hogy az egyik helyesebb a másiknál (például II. Erzsébet gyermekei esetében), vagy azok korábban voltak érvényesek és relevánsak. A Wikidata-kompatibilis sablonok – így az infoboxok is – alapértelmezetten egy tulajdonság normál rangú értékét vagy értékeit fogja megjeleníteni, amennyiben nincs preferált érték. A preferált vagy előnyben részesített rangot a – tudományos vagy közösségi – konszenzust legjobban képviselő, érvényes és naprakész adatok kapják. A preferált rang ideális esetben olyan forráshivatkozásokkal és minősítőkkel rendelkező állításokhoz tartozik, amelyek érvényessége ezek által igazolható. Települések népességi adatainál például mindig a legfrissebb (időpont minősítővel rendelkező) értéknek kell preferált, míg a korábbi évek adatainak normál rangot kell kapniuk. Mivel a felhasználó feltételezhetően a legfrissebb értékre kíváncsi – hacsak nem deklarálja ennek ellenkezőjét –, a sablonok és a lekérdezések alapértelmezetten csak a preferált értékeket hívják meg.

Elavult rangot tudottan hibás, azonban (korábban) általánosan elfogadott állítások kaphatnak. Ezekről vagy ismert, hogy hibás mérési eljárások miatt hibát tartalmaznak, vagy olyan elavult ismereteket képviselnek, amiket valamikor helyesnek vélték. Elavult ranggal szerepelne például a mára tévesnek bizonyult geocentrikus világkép azon állítása, miszerint a Föld a világmindenség középpontja. A hibás állítások elavultként jelölése azok törlése helyett előnyös a tudásgráf adatminősége szempontjából, ugyanis lehetővé teszi a különféle elméletek és elképzelések fejlődésének strukturált módon történő ábrázolását, gazdagabb kontextust kínálva az adatok értelemzéséhez. Fontos

továbbá megjegyezni, hogy a Wikidata egy szekunder adatbázis, nem pedig a tények elsődleges forrása; az adatbázist felépítő (forrás)hivatkozások a tudomány jelen állásának megfelelő és ellentmondó álláspontokat egyaránt tükrözhetnek. A sablonok és lekérdezések alapesetben sosem használják az elavult rangú adatokat [43].

4. Poliglott közösség, többnyelvű adatok

A Wikidata – tartalmát és közösségét tekintve is – egy többnyelvű projekt [44]. A Wikidata többnyelvűségének fontossága elsősorban abban rejlik, hogy lehetővé teszi a világ különböző nyelvterületein élő emberek számára a közös tudásmegosztást és a kutatást. Mindez egybevág a világ különböző nyelvein elérhető Wikipédia céljával, azonban míg például a magyar nyelvű Wikipédia felületén a tudás disszeminálása kizárólag magyar nyelven történik (hasonlóképp az angol, örmény, urdu stb. projektekhez), addig a Wikidata képes egy projekt formájában összefogni a több száz nyelven elérhető információkat. Többnyelvű adatbázisként a tudás és információk különböző nyelveken és területeken is elérhetők, ami hozzájárul a tudás megosztásához és demokratizálódásához.

A többnyelvűség segíti továbbá a tudásmegosztás és a kutatás globális szintű bővítését is. Az adatok és információk megosztása a kevésbé privilegizált régiókkal lehetővé teszi a tudás elérését, összegyűjtését és feldolgozását a világ minden tájáról, megteremtve ezen területek és nyelvek becsatlakozásának lehetőségét a tudományos életbe, hozzájárulva ezáltal új tudományos felfedezések megszületéséhez. A Wikidata többnyelvűsége továbbá segíti a kis népességet lefedő nyelvek és nyelvváltozatok megőrzését is, mivel lehetővé teszi az adatok és információk nyelvi adatainak rögzítését, megóvva azokat az eltűnés veszélyétől, hozzájárulva ezáltal a világ nyelvi és kulturális sokszínűségének megőrzéséhez. A nyelvi egyenlőtlenségek egyik következménye például az is, hogy a Wikidatával kapcsolatos kutatások döntő hányada európai kutatókhoz és kutatóintézetekhez kapcsolódnak, miközben a világ többi része – beleértve a fenntartó Wikimedia Foundation székhelyét, az Amerikai Egyesült Államokat – alig aktív ezen a területen [45].

A strukturált adatok többnyelvűsége a szemantikus web egy fontos aspektusa, hiszen az humán felhasználók számára a címkék jelentik az adatokkal való interakció elsődleges formáját [46]. A többnyelvű adatok lehetséges felhasználási módjai sokszínűek, a felhasználók számára az információ megértéséhez természetes nyelven íródott címkék szükségesek. Egy jól működő és átfogó tudásgráf, amely adatai nyelvek sokaságán elérhető, kiváló alapul szolgálhat a emberi interakcióra tervezett alkalmazások – például csevegőrobotok vagy virtuális asszisztensek – számára [47]. Nem meglepő tehát, hogy a Wikidata az Apple és az Amazon virtuális asszisztensének is adatszolgáltatója – pontosabban fogalmazva mindkét cég alkalmazza a Wikidata szabadon felhasználható adatait a felhasználók által a virtuális asszisztenseknek feltett kérdések megválaszolásához [48].

A Wikidata adatainak többnyelvűségét a korábban már említett elsődleges és alternatív címkék valamint leírások teremtik meg. A projekt közreműködői ezek manuális vagy automatizált létrehozásával és szerkesztésével teszik lehetővé a felhasználók számára az információk természetes nyelven való megértését. A nyelvi információk felvitele mára jellemzően automatizált módon zajlik, így ha egy frissen létrehozott Wikipédia-szócikket bekapcsolunk az adott entitás Wikidata-eleméhez, de az elem elsődleges címkéjét már nem töltjük ki a szócikk nyelvén, akkor ezt idővel pótolja helyettünk a szócikk címének beimportálásával egy automatizált szerkesztések elvégzésére tervezett program (másnéven bot, lásd a későbbiekben). Noha az adatok automatizált módon történő importálása gyors és hatékony alternatívája a „kézzel” (manuális úton) történő felvitelnek, ez korántsem jelenti azt, hogy a továbbiakban nincs szükség az emberi közreműködésre, ugyanis ezen programok csak a Wikidata-elemek már meglévő állításaira támaszkodva képesek címkéket vagy leírásokat generálni, amelyek azonban nem mindig teljes körűek, illetve megtévesztőek is lehetnek az azokat nem kontextusában értelmező programkód számára (például a többszörös család- vagy utónevű személyek esetében az elsődleges címkéről generalizált módon nem eldönthető, hogy a szóban forgó illető neve milyen formában ismert a legszélesebb körben).

Egy 2017-es tanulmány [49] azt találta, hogy a Wikidata összes címkéjének majdnem fele 11 nyelvhez tartozik. A legnépszerűbb nyelv 11,04%-kal az angol volt, amit a holland 6,47, a francia 6,02, a német 5,08 és a spanyol 4,07 százalékkal követett. A nyelvek anyanyelvi beszélőinek száma és Wikidatában való reprezentáltsága között szembetűnő anomáliák figyelhetőek meg, legdominánsabban a kínai nyelv kapcsán. Miközben a kínai nyelv számos változatának van saját Wikipédiája (zh, zh-yue, zh-min-nan, cdo, wuu, hak, gan és zh-classical), a Kínai Népköztársaság cenzúrája és blokkolása miatt ezek csak kerülőúton – VPN- és proxyszolgáltatások segítségével – érhetők el az országban, így a szerkesztések jelentős hányada a világ hozzáférést nem korlátozó részeiről érkezik. Noha mindezek ellenére a legnagyobb kínai nyelvű Wikipédia több mint 1,3 millió szócikkel rendelkezik, a nyelvet beszélők száma ellenére alulreprezentált a Wikidatán, szemben például a holland nyelvvel, amely esetén fordítottan néz ki ez a helyzet.

A tanulmány szerzői rámutatnak továbbá még arra, hogy egy nyelv lefedettségéhez nem feltétlenül szükséges, hogy azt sokan beszéljék. A már említett holland mellett kiemelendő a svéd és a szebuano nyelv helyzete, ugyanis mindkettő azért végezhetett az előkelő 7. és 9. helyen a legtöbb címkével rendelkező nyelvek listáján (3,89 és 2,21 százalékkal), mert a svéd Sverker Johansson által írt Lsjbot elnevezésű automatizált program több mint 3 millió szócikket készített a svéd, és 5,3 milliót a szebuano nyelvű Wikipédián, amelyek címei a Wikidatába is importálva lettek. Belátható, hogy a Fülöp-szigeteken 15,9 millió [50] – ebből a 2010-es népszámlálási adatok szerint magát etnikailag is annak valló 9,1 millió anyanyelvi [51] –beszélővel rendelkező szebuano nyelv 2,21%-os részesedése a portugál nyelv 1,94%-os vagy a kínai nyelv 1,20%-os arányához képest nem egy aprócska nyelv aktív szerkesztőközösségének győzelme volt a világnyelvek felett, hanem az adatbányászatban és az automatizált szövegírásban rejlő lehetőségek egy lehetséges alkalmazhatósága [52]. Tény, hogy egy nyelv teljességre törekvő reprezentáltságához nem feltétlenül szükséges, hogy azt sokan beszéljék; megfelelő (jellemzően automatizált) eszközök kritikus szerepet tudnak játszani ennek sikeres elérésében. Azonban az is könnyen belátható, hogy az elemek és a tulajdonságok címkék útján történő lefordításával egyéni szinten is jelentős eredmények érhetők el, hiszen az az elem vagy tulajdonság attól kezdve az adott nyelven jelenik meg mindazon felhasználók számára, akik azon a nyelven hívják le a nyelvi adatokat, így például a legtöbbet használt (hivatkozott) elemek szerkesztésével rövid idő alatt látványos eredmény érhető el.

Ahogy az előzőekben láthattuk, általános probléma, hogy a nyelvek beszélőinek száma nem korrelál a Wikidata nyelvi adatainak megoszlása tekintetében elfoglalt helyével, ami részben a Wikipédia-szócikkek címeinek importálásából (minél több a szócikk, annál több a nyelvi címke az adott nyelven) és a két projekt közösségének átfedéséből fakadhat. A Wikidata szerkesztőközösségének nyelvtudását egy 2018-as tanulmány [53] kétféle megközelítésben is vizsgálta: elsőként a felhasználók bemutatkozására szolgáló „szerkesztői lapokon” (user page) található Bábel sablon [54] használatát vetették elemzés alá. Ez a sablon kettős funkcióval rendelkezik: egyrészről a külső szemlélők számára információval szolgál az adott szerkesztő nyelvtudásáról, valamint besorolja a bábeli kategóriafába [55] (lehetővé téve egy adott nyelven értő szerkesztők gyors megtalálását), másrészről a rendszer ezáltal az elemek nyelvi tulajdonságait (elsődleges címke, definíció, alternatív nevek) a Bábel sablonban megadott nyelveken kínálja fel a felhasználónak. Utóbbi átfedésben van a kutatók által szintén vizsgált „User Language Settings” beállítással [56], mely segítségével a felhasználói felület – alapértelmezetten angol – megjelenési nyelve módosítható. Ezen beállítás elemzésének első lépéseként az angol nyelv kizárásra került, így a vizsgált felhasználók nagyjából fele került csak elemzése. A leggyakrabban előforduló nyelvek a francia, a német, a spanyol és az orosz voltak, azonban az így vizsgált nyelvek megoszlása nem korrelált az adott nyelven végzett, az összes szerkesztés halmazán vett arányával.

A szerkesztők nyelvtudásának vizsgálatára alkalmasabb volt a Bábel sablon adatainak elemzése, ahol a felhasználó választása nem kizárólagos, hanem több választási lehetőség is rendelkezésére áll. A sablon által a felhasználó önbevallásos alapon egy hétfokozatú skálán értékelheti

nyelvtudását, ahol a nullás (0) érték a nyelv ismeretének teljes hiányát, míg az anyanyelvi tudást az N jelöli [57]. A 4120 felhasználót tartalmazó vizsgált mintában a leggyakrabban ismert nyelvek sorrendjében nem történt változás, azonban empirikus úton is bizonyosságot nyert az a feltételezés, hogy az angol nem csak a legnagyobb arányban megértett nyelv a szerkesztőközösségben, hanem a négy legelterjedtebb nyelv (angol, francia, német és spanyol) tekintetében a legkevesebben jelölték, hogy azt egyáltalán (0) vagy csak kicsit (1) értik. Ez azért is különösen fontos, mert a Wikidata közösségi megbeszéléseinek nyelve döntően az angol, így szükségszerű annak legalább alapszintű ismerete. Többnyelvű projektként azért is fontos a bilingvis és poliglott szerkesztők jelenléte, mert az angolul megszövegezett irányelvek és útmutatók lefordításával nyelvileg is inkluzívabbá válhat a közösség, a szerkesztés miként jével még csak ismerkedő kezdők pedig anyanyelvükön férhetnek hozzá a szükséges tudáshoz. A kutatás rámutatott még arra is, hogy noha a Wikidata többnyelvű közösségének tagjai jellemzően saját nyelveiken szerkesztenek, azokat azonban gyakran általuk nem ismert nyelvekre is kiterjesztik.

5. A közösségi közreműködésen alapuló modell

A Wikidata a Wikipédiához hasonlóan ingyenes, bárki által szabadon szerkeszthető és bővíthető projekt, amelynek tartalmát egy önkéntes szerkesztőkből álló közösség tartja karban. A két projekt közötti legnagyobb különbségek jellemzően strukturális eredetűek: míg a Wikipédia nyelvi alapon tagolódik különböző projektekre (enciklopédiákra), addig a Wikidata az információkat többnyelvű tudásgráfként egy projekt képében reprezentálja. Az adatok közösségi szerkeszthetőségét azok szabad licenc alatt történő közzététele teszi lehetővé: míg a magyar Wikipédia tartalmára bizonyos megkötéseket tartalmazó Creative Commons Nevezd meg! – Így add tovább! 4.0 és GFDL-1.2+ kettős licenc vonatkozik, addig a Wikidata strukturált adatai közkincsnek számítanak. Ezen licencek mindkét esetben garantálják a tartalom szabad (újra)felhasználhatóságát, ezáltal a közösségi közreműködésen alapuló modellt: jogi szempontból ez úgy valósul meg a Wikidata esetében, hogy a strukturált adatokat – elemeket, tulajdonságokat, lexémákat és sémákat – tartalmazó névterek Creative Commons 0 (Public Domain) licenc alatt kerülnek közzétételre, míg a fennmaradó névterek tartalmára a korábban már említett Így add tovább! 4.0 licenc vonatkozik, tehát a szerkesztők közreműködéseiket ezen licencek valamelyikén teszik közzé [58].

Típusát tekintve egy szerkesztő lehet ember vagy automatizált program (bot). A Wikidata filozófiája lehetővé teszi, hogy a projekthez bárki hozzájáruljon, ehhez pedig szerkesztői fiók sem feltétlenül szükséges, a legtöbb lap – hasonlóképpen a Wikipédiához – közvetlenül szerkeszthető is az olvasó által. Amennyiben a felhasználó nem kíván regisztrálni vagy bejelentkezni már meglévő fiókjába, akkor a weboldal mögött húzódó MediaWiki szoftver az IP-címéhez társítja a szerkesztéseket, ugyanis minden közreműködéshez felhasználóazonosítónak is kell tartoznia [59]. A Wikimedia Foundation által fenntartott projektek – például a Wikipédia, Wikikönyvek, Wikiforrás, Wikimedia Commons – felhasználói a központi azonosításnak hála (globális) fiókjukat valamennyi projekten, így a Wikidatán is használhatják, tehát nem szükséges külön regisztráció [60]. Ez az átjárhatóság jelentős könnyebbséget jelent például abban, hogy a Wikipédia-szócikkekről a Wikidatára átnavigálva a felület és az adatok címkéi rögtön a felhasználói fiók (a többnyelvűséget taglaló fejezetben leírt) beállításainak megfelelő nyelven jelennek meg, támogatva ezáltal az egynyelvű szerkesztők munkáját, valamint motiválva a többnyelvű szerkesztőket az elemen szereplő állítások még le nem fordított címkéinek pótlására.

A közreműködni kívánó felhasználók számára számos írott útmutató érhető el attól függően, hogy miképp szeretne hozzájárulni a projekthez, így például egy dedikált oldal szól az adataikat felajánlani vagy beimportálni kívánó intézmények, cégek és magánszemélyek számára. Az oldal laikusok számára is érthető nyelvezettel foglalja össze már megvalósult példák bemutatásával, hogy milyen előnyei vannak az (intézményi) adatfelajánlásnak, hogy hogyan zajlik mindez, és hogy miképp megvalósítható az adatok Wikidatába történő importálása [61]. Feltételezve, hogy egy közösségi közreműködésen alapuló projekt esetében a felhasználók elsősorban a szerkesztés miként jére

kíváncsiak, a Wikidata többek között olyan interaktív bemutatókat is kínál, amelyek által megismerhető a Wikidata működése, és általuk elsajátíthatók az adatok hozzáadásának fortélyai, akár szerkesztői fiók létrehozása nélkül is. A Wikidata-bemutatók két fő témakör szerint nagyjából öt perc alatt elvégezhető modulok által mutatják be a legszükségesebb tudnivalókat a felhasználóknak: az első témakör a Wikidata alapjainak elsajátítására, így például az elemek felépítésének, az állítások hozzáadásának és szerkesztésének, valamint a forráshivatkozások megismerésére helyezi a hangsúlyt. A második témakör már gyakran előforduló konkrét tevékenységeket mutat be, megtanítva a felhasználónak a Wikidata-elemek illusztráltságát növelő képek, valamint a földrajzi helyek feltártságát növelő koordináta és közigazgatási egység tulajdonságok hozzáadásának módját [62]. A kezdő szerkesztők eligazodását egy gyakran ismételt kérdéseket tartalmazó oldal is segíti, de a felmerülő nehézségeikkel nyelvspecifikus közösségi üzenőfalakon is fordulhatnak segítségért gyakorlott szerkesztőkhöz [63, 64]. Mindezen útmutatók szükséges (de nem elégséges) feltételei a kezdő szerkesztők megtartásának, amely kritikus egy közösségi közreműködésen alapuló – ezáltal lemorzsolódással különösen veszélyeztetett – projekt dinamikus növekedéséhez.

A Wikipédiához hasonlóan, ahol a felhasználók egy marginális kisebbsége (egy százaléka) felelős a közreműködések döntő hányadáért, 9% csak szórványosan és a fennmaradó 90% pedig egyáltalán nem szerkeszt, a közreműködések egyenlőtlen megoszlása a Wikidata szerkesztőközössége esetében is megfigyelhető [65]. Egy, a regisztrált Wikidata-szerkesztők szerkesztési viselkedését elemző 2018-as longitudinális keresztmetszeti vizsgálat azt találta, hogy a régebb óta regisztrált szerkesztők (magasabb élettartalmú fiókok) rendszeresebben és állandóbban vesznek részt a projekt bővítésében, mint a frissebb regisztrációval rendelkező szerkesztők. Ugyanez figyelhető meg akkor, ha a szerkesztések számát vesszük alapul; a magasabb szerkesztésszámmal rendelkező fiókok többet és állandóbban szerkesztenek, míg az alacsonyabb szerkesztésszámú társaik. Kiemelendő, hogy a kutatásban vizsgált kezdeti – 2012 októbere és 2016 júliusa közötti – periódus időszakában a Wikidatának több ezer olyan szerkesztője volt, akik állandó jelleggel „végig szerkesztették" ezt a négy évet, ami egy önkéntes közreműködésre alapuló és annak belső motivációjára építő projekt esetében egy különösen értékes törzsbázist vagy keménymagot jelent. Kardinális azonban az utánpótlás kérdése: a Wikidata méretének és összetettségének növekedésével egyre nagyobb kihívást jelent a már aktív és újonc szerkesztőknek egyaránt a Wikidata szerkesztéséhez (edit), lekérdezéséhez (query) és vizualizálásához (visualise) szükséges eszközök és praktikák sokaságának elsajátítása, valamint a projekt azon aspektusainak megtalálása, amiben közreműködhetnek. A kutatás azt találta, hogy a Wikidata-szerkesztők jelenléte igen nagy részben tiszavirág-életű, közreműködésük és részvételük nem állandó, noha meglenne bennük a potenciál. A korai lemorzsolódás soktényezős változókkal magyarázható, így például a szabad tudás iránti elkötelezettség hiánya vagy például az újoncok számára nem eléggé kézenfekvő és intuitív felület [66].

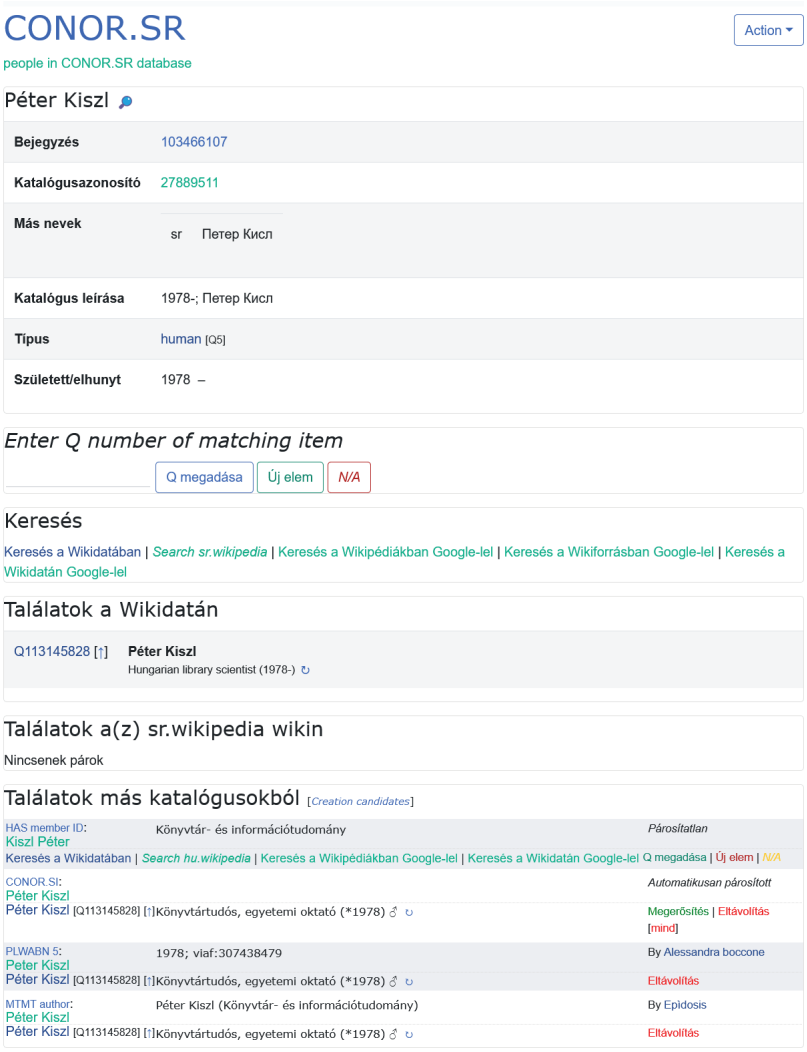
A Wikipédiához képest rövidebb időintervallumú szerkesztések a Wikidata természetéből fakadnak, hiszen míg előbbi esetében a fejezet- és szócikkíráson van a hangsúly, de a belső és forráshivatkozások, helyesírási hibák és elírások javítása aránylag rövid idő alatt abszolválhatók, addig a Wikidata tripletek képében tárolt strukturált adatainak szerkesztése messzemenőleg gyorsabb és gördülékenyebb. Hogy pontosan mennyire, az nagyban függ attól, hogy a szerkesztő milyen felületen kíván dolgozni. A legkézenfekvőbb a *2. ábrán* már korábban ismertetett webes felület, amely emberi használatra megfelelő – asztali és mobil eszközökre optimalizált – formában teszi lehetővé az adatok olvasását, valamint szerkesztését. Egy 2018-as tanulmány a Wikidata 2017. október elsejei pillanatképét vizsgálva arra az eredményre jutott, hogy a szerkesztők több mint 95%-a kizárólag ezt a felületet használta a szerkesztéshez [67].

A gördülékeny felhasználói élményt szolgálják továbbá a segédeszközök, amelyek az oldal testhezállóbb elrendezésétől kezdve általános szerkesztési műveleteken túl specifikus igények kielégítését szolgálják (pl. egy forráshivatkozás egyszerre több értékhez is történő beszúrása) [68]. Mivel a Wikidata több mint százmillió elemmel ren-

delkezik, elkerülhetetlen, hogy – az emberi figyelmetlenségből vagy az automatizált importálásból eredő okokból – már létező entitásoknak duplikált elemeik keletkezzenek. Ilyenkor az egységes forrásazonosítók (URI-k) permanensségének megőrzése érdekében ezek nem törlésre, hanem összevonásra kerülnek: ez egy többlépcsős folyamat, amelynek során előbb a két elemen megtalálható adatok az alacsonyabb azonosítószámú (korábban létrehozott) elem helyére kerülnek (ez lesz a célelem), majd a kiürült elem átirányítássá válik a célelemre, továbbítva az embereket és a programokat a teljesebb információhoz [69]. A segédeszközök döntő része valamennyi regisztrált szerkesztő számára elérhető, így az ezek és a nyelvi/bábeli beállítások nyújtotta előnyök okán is érdemes lehet bejelentkezett felhasználóként közreműködni a projektben. További előny, hogy egy fiókkal elérhetővé válnak az alapértelmezett szerkesztői felületnél egy fokkal felhasználóbarátabb és egyszerűbben használható félautomata eszközökhöz is: ezek átmenetet képeznek a kézzel és az automatizált módon történő szerkesztés között, lehetővé téve a magasabb rátájú, „kézzel" történő közreműködést. A „The Distributed Game" elnevezésű eszköz például egy játékmód keretében frissen létrehozott, Wikidata-elemmel még nem rendelkező Wikipédia-szócikkek összekapcsolására kínál felhasználóbarát módot, lehetséges elemeket felkínálva a szerkesztőnek, időt spórolva ezáltal a végső döntést meghozó embernek [70]. Hasonló eszköz még a „Mix'n'match", amely külső adatbázisok azonosítóinak Wikidata-elemekhez történő kézi párosítását könnyíti meg azáltal, hogy az adott katalógusban elérhető információk (például név, születési és halálozási dátum, típus) alapján hasonlóságot mutató Wikidata-elemeket kínál fel: amennyiben az adott entitásnak már létezik Wikidata-eleme, az akkor nagy valószínűséggel megjelenik a találatok között, a párosítás (az adott külső azonosító hozzáadása a Wikidata-elemhez) ekkor csupán már csak egy kattintás. Jelenleg is több olyan „Mix'n'match game" zajlik, ahol magyar adatbázisok párosítása zajlik, ilyen például a Magyar Tudományos Akadémia köztestületi tagjait tartalmazó Akadémiai Adattár, a Magyar Tudományos Művek Tára vagy például az Elektronikus Periodika Adatbázis [6].

A Mix'n'match gyakorlati működését az *5. ábra* illusztrálja, ahol a délszláv államok COBISS elnevezésű egységes könyvtári információs rendszerének CONOR.SR nemzetközi katalógusa látható. A CONOR.SR a COBISS szerb részadatbázisa, a képernyőképen látható játék ezek azonosítóinak párosítására szolgál. A bevezetőből a felhasználó számára kiderül Kiszl Péter katalógusazonosítója (27889511), latin és cirill írásrendszerű személynév főtétele (Кисл, Петер és Kiszl, Péter), valamint az ehhez kapcsolódó kiegészítő (kronologikus) adatok. Mivel Kiszl Péter (Q113145828) Wikidata-elemén [71] még nem szerepel a CONOR.SR-azonosító (P8851) tulajdonság, ezért a Mix'n'match párosítatlan azonosítóként felkínálja azt összekapcsolásra. A „Találatok más katalógusokból" szakaszban felsorolt további négy katalógus közül három – köztük a CONOR szlovén részadatbázisának – már meglévő párosítása szolgál megerősítő támpontként, hogy a „Találatok a Wikidatán" szakaszban felkínált elem valóban a keresett elem, amihez a fölfelé mutató nyílra (↑) kattintva hozzá is adható az azonosítószám.

A felhasználók különleges csoportját képezik a botok vagy robotok, amelyeket humán felhasználók fejlesztenek és tartanak karban elsősorban olyan feladatokra, ahol nagyszámú (általában emberi döntéshozatalt nem igénylő) szerkesztést kell elvégezni. A botok a humán szerkesztőkhöz hasonlóan egyedi felhasználói fiókkal rendelkeznek, beazonosíthatóságuk érdekében nevüket jellemzően „gazdájukról" kapják, kiegészülve egy „bot" címkével. A botjogosultsággal rendelkező felhasználói fiókok lényege, hogy általuk magas szerkesztésrátával nagy mennyiségű adatot lehessen beimportálni és szerkeszteni: mivel esetenként egy-egy feladat százezres szerkesztésszámmal is járhat, ezért a botok által végzett változtatások alapesetben elrejtésre kerülnek a szerkesztők elől. Egy bot legfontosabb többletjogosultságai közé tartozik a „bot" (automatikus folyamatként való kezelés) és az „apihighlimits" (nagyobb mennyiségű lekérdezés az API-n keresztül), valamint az „autopatrol" (szerkesztések automatikusan ellenőrzöttként való jelölése) jogok [72]. A jogokhoz azonban kötelezettségek is járnak: az etikus API-használat mellett arra is figyelni kell, hogy egy rosszul megválasztott magas szerkesztésrátával



5. ábra Képernyőkép a CONOR.SR-azonosítókat párosító Mix'n'match felületéről. Forrás: https://mix-n-match.toolforge.org/?#/entry/103466107

végzett feladat nehogy leterhelje a szervert, ezt megelőzendően mindig szükséges megadni egy „maxlag" (maximálisan engedélyezett késedelem) értéket [73, 74]. Mivel egy hibásan működő bot jelentős károkat képes okozni a projektnek – elsősorban a már említett feltűnésmentes működése által és a hibás szerkesztések időigényes visszaállítása miatt –, ezek működését egy hivatalos Wikidata-irányelv szabályozza, amelynek része a közösség általi engedélyeztetésük formalizált procedúrája is [75]. Az irányelv értelmében a botjog iránt folyamodó szerkesztőnek mindenekelőtt egyértelmű leírást kell adnia a futtatni kívánt feladat(ok)ról. A bot működéséhez használt programkódok megosztása ugyan nem kötelező, de a közösség bátorítja azok (szabad licenc alatt történő) nyilvánosságra hozatalát, lehetővé téve azok közösségi közreműködés általi karbantartását, egy-egy feladat tovább öröklését a projekt életéből kivonuló szerkesztőktől. A kérelem megírását követően egy 50 és 250 közötti próbaszerkesztés-sorozatot kell végeznie, amely alapján a közösség véleményezheti, végső soron pedig támogathatja vagy ellenezheti a botjog megadását. Mivel a próbaszerkesztéseket úgy kell elvégezni, hogy az adott fiók még nem rendelkezik botjoggal, így azok valamennyi szerkesztő számára megjelennek, növelve ezáltal az esetleges hibás és félreműködések észrevételének esélyét. A közösségi hozzájárulást követően egy bot mindaddig üzemelhet, amíg tulajdonosa le nem állítja (mert például a jövőben nem kívánja már karbantartani), valamennyi jóváhagyott fela-

datát el nem végzi (például egy adott, tovább már nem bővülő adatbázis elemeinek importálását) vagy a közösség vissza nem vonja engedélyét (például félreműködése miatt). Noha a botjoggal rendelkező fiókok aránya csupán az aktív – az elmúlt 30 napban legalább egy szerkesztéssel rendelkező – szerkesztők másfél százalékát teszi ki (2023. augusztus 18-án a Wikidata 23 394 aktív szerkesztője közül 360 rendelkezett botjogosultsággal) [76], a Wikidata szerkesztéseinek döntő részét botok végzik. A projekt indulását követő években a botok a szerkesztések majdnem 90%-áért voltak felelősek: ez a szám a wiki- és nyelvközi hivatkozások nagyszámú importálásának lezárultával sem csökkent számottevően, tehát a szerkesztések döntő többségéért jelenleg is automatizált szerkesztéseket végző programok felelősek [67].

A Wikidata százmilliós elemszáma és az adatok növekvő mélységű komplexitása predesztinálja, hogy a projekt humán – regisztrált és anonim – szerkesztői csak automatizált közreműködésre képes programok segítségével képesek lépést tartani a reprezentálni kívánt világ információinak változásaival és vice versa. Noha eme két szerkesztői csoport élesen elkülönül egymástól, mindkét oldal közreműködése egyenértékűen elengedhetetlen a Wikidata fennmaradásához és dinamikus bővüléséhez, együttműködésük törvényszerű. Tény, hogy számszerűleg a botok végzik a legtöbb szerkesztést, azonban a szerkesztések minőségének vizsgálatakor a Wikidata nem részesíti előnyben – ad hominem – egyik felet sem, a botok ugyanúgy felülírhatják az emberi szerkesztők közreműködéseit, mint fordítva, szemben például a kizárólag emberi befogadásra szánt szövegeket tartalmazó Wikipédiával. Az emberi szerkesztők és botok közötti interakció létfontosságú a Wikidata minőségének fenntartása érdekében. Az adatok bővítésének és pontosságának növelésében mindkét oldal egymásra van utalva, hiszen a jellemzően repetitív és ezért könnyen automatizálható feladatok esetében az állandó emberi kontroll az önkéntes projekt szűkös humánerőforrás-kapacitásának elpazarlását jelentené, amely elengedhetetlen például a Wikidata-tulajdonságok komplex rendszerének fenntartásához. Az emberi szerkesztők és a botok közötti nagyobb kooperáció és egyenlőbb feladatmegosztás a Wikidata adatminőségének javulásához vezethet. Érdekesség, hogy a szerkesztőközösség mérete azonban csak korlátozottan van pozitív hatással a projekt teljesítményére, szemben a Wikipédia kapcsán eddig megfigyelt trendekkel [77].

6. Összegzés

A Wikidata 2012-es indulásakor megfogalmazott kezdeti célkitűzések elérését követően a projekt hamar meghaladta a Wikimedia Foundation nyelv- és wikiközi hivatkozásait gyűjtő repozitórium szerepét, és rövid idő alatt a szemantikus web egyik legfontosabb adatbázisává nőtte ki magát. A tudásgráf adatainak gyorsléptékű növekedése és alkalmazásának széleskörű elterjedése ellenére azonban a Wikidatával kapcsolatos kutatások alacsony száma és egyenlőtlen eloszlása a kutatási téma alulreprezentáltságát és kiforratlanságát jelzi. A téma Magyarországon alig publikált, az ezzel foglalkozó kutatások, konferencia-előadások és workshopok szinte kivétel nélkül a – korábban a Petőfi Irodalmi Múzeumhoz, 2021 decemberétől az Országos Széchényi Könyvtárhoz tartozó – Digitális Bölcsészeti Központ munkatársaihoz kötődnek. A Wikidata átfogó, tudományos igényű és magyar nyelvű feldolgozását célozva a dolgozat részletesen ismertette a tudásgráf strukturális felépítését és legfontosabb ismérveit. Az információkat felépítő RDF tripletekhez szükséges – elem, állítás, tulajdonság és érték – alapfogalmakon kívül az adatok (újra)felhasználásához elengedhetetlen forráshivatkozások és azonosítók, valamint a visszakereséshez szükséges címkék és leírások is részletesen bemutatásra kerültek.

A Wikidata és társprojektjei által támogatott nyitott tudás mozgalom terjesztésének fontos aspektusa a többnyelvűség: az összesen több mint 320 nyelven elérhető 61 millió Wikipédia-szócikk adataival kiegészült Wikidata kiváló adatforrásaként szolgálhat a humán interakcióra tervezett alkalmazások, így például a mesterséges intelligenciával támogatott csevegőrobotok (pl. ChatGPT) és virtuális asszisztensek (pl. Amazon Alexa, Apple Siri) számára. A magyar nyelv szerencsés helyzetben van e tekintetben, ugyanis a Wikidata Európacentrikussága, a magyar nyelvű közreműködők felülreprezentáltsága a szerkesztőközösségben és a magyar Wikipédia–Wikidata szoros kooperá-

ciója előnyös kiindulási helyzetet teremt az integrációban gondolkodó hazai intézmények számára. Az adatbázis strukturált adatainak Wikipédia-szócikkek információiból, partnerségi adatcserékből, közgyűjteményi (adat)hozzájárulásokból, és nem utolsó sorban a közösségi közreműködésen alapuló modellből fakadó töretlen növekedése egyben a projekt népszerűségét indukálta.

A tanulmánysorozat következő részében bemutatásra kerülnek a Wikidata integrálásának megvalósult jógyakorlatai és lehetséges alkalmazásai, így például az, hogy a Wikidata miképp lehet képes kiváltani a közgyűjteményi szereplők hagyományos authoritykontrollját, lehetővé téve az intézményi adatsilókba zárt adatok nyílttá és szabadon elérhetővé tételét.

Irodalom és jegyzetek

[1] Wikidata, *Eötvös Loránd Tudományegyetem*, Elérhető: https://www.wikidata.org/wiki/Q390287 (Utolsó elérés: 2023. 08. 18.)

[2] Wikidata, *Magyarország*, Elérhető: https://www.wikidata.org/wiki/Q28 (Utolsó elérés: 2023. 08. 18.)

[3] Tartalmi lapnak számít minden olyan fő névtérben található elem, amely nem átirányító lap (redirect) és törlésre sem került. Átirányító lapok általában úgy keletkeznek, hogy egy már létező elem által reprezentált entitásról újabb elem, duplikátum kerül létrehozásra: ilyenkor a két elem összevonásra kerül, és a szokásjog szerint az alacsonyabb számú azonosítóval rendelkező elem reprezentálja tovább az entitást, amire a magasabb számú azonosítóval rendelkező elem átirányítja az olvasót. Egy elem törlésére csak akkor kerül sor, ha az nem felel meg a Wikidata nevezetességi kritériumainak. Mindezek következtében a tartalmi lapok számánál mindig kevesebb, mint a legújabb elem azonosítója.

[4] Wikidata, *Statisztika*, Elérhető: https://www.wikidata.org/wiki/Special:Statistics (Utolsó elérés: 2023. 08. 18.)

[5] Magyar Tudományos Művek Tára, *Keresőkifejezés: Wikidata*, Elérhető: https://m2.mtmt.hu/gui2/?mode=search&query=publication;labelOrMtid;eq;Wikidata (2023. augusztus 18-án a Magyar Tudományos Művek Tárában a „Wikidata" keresőszó összesen nyolc találatot eredményezett)

[6] Molnár, B. *A Wikidata és a Nemzeti Névtér kapcsolódási lehetőségei*, Könyvtári Figyelő, 67(1), p. 46–55, 2021.

[7] Roth, M. *The Wikipedia data revolution*, Diff, Elérhető: https://diff.wikimedia.org/2012/03/30/the-wikipedia-data-revolution/ (Utolsó elérés: 2023. 08. 18.)

[8] Dickinson, B. *Paul Allen Invests In A Massive Project To Make Wikipedia Better*, Insider, Elérhető: https://www.businessinsider.com/paul-allen-invests-in-wikidata-project-2012-3 (Utolsó elérés: 2023. 08. 18.)

[9] Leitch, T. *Wikipedia U: Knowledge, authority, and liberal education in the digital age*, JHU Press, 2014. ISBN: 9781421415505

[10] Wikimedia Statistics, *Pages to date*, Elérhető: https://stats.wikimedia.org/#/all-projects/content/pages-to-date/normal|table|all|page_type~content|monthly (Utolsó elérés: 2023. 08. 18.)

[11] Pintscher, L. *Erste Schritte von Wikidata in der ungarischen Wikipedia*, Wikimedia Deutschland Blog, Elérhető: https://blog.wikimedia.de/2013/01/14/erste-schritte-von-wikidata-in-der-ungarischen-wikipedia/ (Utolsó elérés: 2023. 08. 18.)

[12] Pluta, W. *Wikidata ist für alle Wikipedien da*, Golem.de, Elérhető: https://www.golem.de/news/onlineenzyklopaedie-wikidata-ist-fuer-alle-wikipedien-da-1304-98941.html (Utolsó elérés: 2023. 08. 18.)

[13] Wikimedia Commons, *Wikidata is here!*, Elérhető: https://commons.wikimedia.org/wiki/Commons:Village_pump/Archive/2013/10#Wikidata_is_here.21 (Utolsó elérés: 2023. 08. 18.)

[14] Wikidata, *Wikidata:Interwiki conflicts*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Interwiki_conflicts (Utolsó elérés: 2023. 08. 18.)

[15] Thirumalai, M.S., Mallikarjun, B., Singh Rangila, R. *Bringing Order to Linguistic Diversity: Language Planning in the British Raj*, Language in India, Elérhető: https://web.archive.org/web/20080526010825/http:/www.languageinindia.com/oct2001/punjab1.html (Utolsó elérés: 2023. 08. 18.)

[16] Wikidata, *User:Pasleim/Connectivity*, Elérhető: https://www.wikidata.org/wiki/User:Pasleim/Connectivity (Utolsó elérés: 2023. 08. 18.)

[17] Yu, L. *A developer's guide to the semantic Web*, Springer, 2011. ISBN: 978-3-642-15970-1

[18] Broughton, J. *Wikipedia: The Missing Manual*, O'Reilly Media, 2008. ISBN: 9780596553777

[19] Pintscher, L. *Erste Teile von Phase 2 von Wikidata in Betrieb*, Wikimedia Deutschland Blog, Elérhető: https://blog.wikimedia.de/2013/02/04/first-parts-of-phase-2-of-wikidata-going-live/ (Utolsó elérés: 2023. 08. 18.)

[20] Wikidata, *Douglas Adams*, Elérhető: https://www.wikidata.org/wiki/Q42 (Utolsó elérés: 2023. 08. 18.)

[21] Wikidata, *Help:Label*, Elérhető: https://www.wikidata.org/wiki/Help:Label (Utolsó elérés: 2023. 08. 18.)

[22] Wikidata, *Cambridge*, Elérhető: https://www.wikidata.org/wiki/Q350 (Utolsó elérés: 2023. 08. 18.)

[23] Wikidata, *Cambridge*, Elérhető: https://www.wikidata.org/wiki/Q49111 (Utolsó elérés: 2023. 08. 18.)

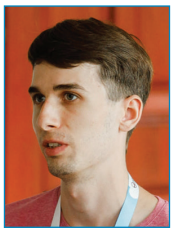
[24] Wikidata, *Help:Description,* Elérhető: https://www.wikidata.org/wiki/Help:Description (Utolsó elérés: 2023. 08. 18.)

[25] Wikidata, *Help:Aliases*, Elérhető: https://www.wikidata.org/wiki/Help:Aliases (Utolsó elérés: 2023. 08. 18.)

[26] Wikidata, *I. Erzsébet*, Elérhető: https://www.wikidata.org/wiki/Q7207 (Utolsó elérés: 2023. 08. 18.)
[27] Wikidata, *gyermek*, Elérhető: https://www.wikidata.org/wiki/Property:P40 (Utolsó elérés: 2023. 08. 18.)
[28] Wikidata, *Help:Statements*, Elérhető: https://www.wikidata.org/wiki/Help:Statements (Utolsó elérés: 2023. 08. 18.)
[29] Wikidata, *Help:Properties*, Elérhető: https://www.wikidata.org/wiki/Help:Properties (Utolsó elérés: 2023. 08. 18.)
[30] Wikidata, *Wikidata: List of properties*, Elérhető: https://www.wikidata.org/wiki/Wikidata:List_of_properties (Utolsó elérés: 2023. 08. 18.)
[31] Wikidata, *List of Properties*, Elérhető: https://www.wikidata.org/wiki/Special:ListProperties (Utolsó elérés: 2023. 08. 18.)
[32] Wikidata, *Wikidata:Property creation*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Property_creation (Utolsó elérés: 2023. 08. 18.)
[33] Wikidata, *Wikidata:Properties for deletion*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Properties_for_deletion (Utolsó elérés: 2023. 08. 18.)
[34] A Magyar Nemzet szerzője, *Elérhető a Magyar Nemzeti Névtér bővített változata*, Magyar Nemzet, Elérhető: https://magyarnemzet.hu/kultura/2023/02/elerheto-a-magyar-nemzeti-nevter-bovitett-valtozata (Utolsó elérés: 2023. 08. 18.)
[35] Wikidata, *Wikidata:Property proposal/Hungarian National Namespace person ID (new)*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Property_proposal/Hungarian_National_Namespace_person_ID_(new) (Utolsó elérés: 2023. 08. 18.)
[36] Wikimedia Cloud Virtual Private Server, *Quarry*, Elérhető: https://quarry.wmcloud.org/ (Utolsó elérés: 2023. 08. 18.) Vizsgált adatbázis: wikidatawiki_p. A lekérdezéshez használt kód: SELECT pid2, probability FROM wikidatawiki_p.wbs_propertypairs WHERE qid1=5 and pid1=31 and context='item' and count>100 order by probability desc
[37] Wikidata, *Help:Qualifiers*, Elérhető: https://www.wikidata.org/wiki/Help:Qualifiers (Utolsó elérés: 2023. 08. 18.)
[38] Wikidata, *Wikidata:Verifiability*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Verifiability (Utolsó elérés: 2023. 08. 18.)
[39] Serra, L. G., Schneider, J. A., Segundo, J. E. S. *Person Identifiers in MARC 21 Records in a Semantic Environment*, Cataloging & Classification Quarterly, 58(5), p. 505–519, 2020. https://doi.org/10.1080/01639374.2020.1771499
[40] Wikidata, *VIAF-azonosító*, Elérhető: https://www.wikidata.org/wiki/Property:P214 (Utolsó elérés: 2023. 08. 18.)
[41] Wikidata, *DOI*, Elérhető: https://www.wikidata.org/wiki/Property:P356 (Utolsó elérés: 2023. 08. 18.)
[42] Wikidata, *II. Erzsébet*, Elérhető: https://www.wikidata.org/wiki/Q9682 (Utolsó elérés: 2023. 08. 18.)
[43] Wikidata, *Help:Ranking*, Elérhető: https://www.wikidata.org/wiki/Help:Ranking (Utolsó elérés: 2023. 08. 18.)
[44] Wikidata, *Help:Multilingual*, Elérhető: https://www.wikidata.org/wiki/Help:Multilingual (Utolsó elérés: 2023. 08. 18.)
[45] Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. *A systematic literature review on Wikidata*, Data Technologies and Applications, 53(3), p. 250–268, 2019. https://doi.org/10.1108/DTA-12-2018-0110
[46] Ell, B., Vrandečić, D., Simperl, E. *Labels in the Web of Data*, In: The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I, Springer, Heidelberg–New York, 2011, p. 162–176. ISBN: 978-3-642-25073-6
[47] Vougiouklis, P., Hare, J., Simperl, E. *A neural network approach for knowledge-driven response generation*, In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Oszaka, 2016, p. 3370–3380. ISBN: 978-4-87974-702-0
[48] Simonite, T. *Inside the Alexa-Friendly World of Wikidata*, Wired, Elérhető: https://www.wired.com/story/inside-the-alexa-friendly-world-of-wikidata/ (Utolsó elérés: 2023. 08. 18.)
[49] Kaffee, L., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., Pintscher, L. *A glimpse into Babel: an analysis of multilinguality in Wikidata*, In: Proceedings of the 13th International Symposium on Open Collaboration, ACM, New York, 2017, p. 1–5. https://doi.org/10.1145/3125433.3125465
[50] Dreisbach, J. L., Demeterio III, F. P. A. *Intergenerational Language Preference Shift among Cebuanos on the Cebuano, Filipino, and English Languages*, LLT Journal, 23(2), p. 220–240, 2020. https://doi.org/10.24071/llt.2020.230203
[51] National Statistics Office, *2010 Census of Population and Housing*, Report No. 2A – Demographic and Housing Characteristics (Non-Sample Variables), National Statistics Office, 2013, ISSN: 0117-1453
[52] Lsjbot AB, *Home*, Elérhető: http://lsjbot.se/ (Utolsó elérés: 2023. 08. 18.)
[53] Kaffee, L. A., Simperl, E. *Analysis of editors' languages in wikidata*, In: Proceedings of the 14th International Symposium on Open Collaboration, Association for Computing Machinery, New York, 2018, p. 1–5. https://doi.org/10.1145/3233391.3233965
[54] Wikidata, *Wikidata:Userboxes*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Userboxes (Utolsó elérés: 2023. 08. 18.)
[55] Wikidata, *Category:Babel - User by language*, Elérhető: https://www.wikidata.org/wiki/Category:Babel_-_Users_by_language (Utolsó elérés: 2023. 08. 18.)
[56] Wikidata, *Help:Navigating Wikidata/User Options*, Elérhető: https://www.wikidata.org/wiki/Help:Navigating_Wikidata/User_Options (Utolsó elérés: 2023. 08. 18.)

[57] Wikipédia, *Wikipédia:Bábel*, Elérhető: https://hu.wikipedia.org/wiki/Wikip%C3%A9dia:B%C3%A1bel (Utolsó elérés: 2023. 08. 18.)
[58] Wikidata, *Help:About data*, Elérhető: https://www.wikidata.org/wiki/Help:About_data (Utolsó elérés: 2023. 08. 18.)
[59] Phabricator, *Add user preference to deactivate/delete user account*, Elérhető: https://phabricator.wikimedia.org/T34815 (Utolsó elérés: 2023. 08. 18.)
[60] Meta, *Help:Unified login*, Elérhető: https://meta.wikimedia.org/wiki/Help:Unified_login (Utolsó elérés: 2023. 08. 18.)
[61] Wikidata, *Wikidata:Data donation*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Data_donation (Utolsó elérés: 2023. 08. 18.)
[62] Wikidata, *Wikidata:Tours*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Tours (Utolsó elérés: 2023. 08. 18.)
[63] Wikidata, *Help:FAQ*, Elérhető: https://www.wikidata.org/wiki/Help:FAQ (Utolsó elérés: 2023. 08. 18.)
[64] Wikidata, *Template:ProjectChatLanguages*, Elérhető: https://www.wikidata.org/wiki/Template:ProjectChatLanguages (Utolsó elérés: 2023. 08. 18.)
[65] Yasseri, T., Sumi, R., Kertész, J. *Circadian patterns of wikipedia editorial activity: A demographic analysis*, PloS one, 7(1), e30091, 2012. https://doi.org/10.1371/journal.pone.0030091
[66] Sarasua, C., Checco, A., Demartini, G., Difallah, D., Feldman, M., Pintscher, L. *The evolution of power and standard Wikidata editors: comparing editing behavior over time to predict lifespan and volume of edits*, Computer Supported Cooperative Work, 28, p. 843–882, 2019. https://doi.org/10.1007/s10606-018-9344-y
[67] Piscopo, A. *Wikidata: A New Paradigm of Human-Bot Collaboration?*, arXiv, 2018. https://doi.org/10.48550/arXiv.1810.00931
[68] Wikidata, *Wikidata:Tools*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Tools (Utolsó elérés: 2023. 08. 18.)
[69] Wikidata, *Help:Merge*, Elérhető: https://www.wikidata.org/wiki/Help:Merge (Utolsó elérés: 2023. 08. 18.)
[70] Toolforge, *The Distributed Game*, Elérhető: https://wikidata-game.toolforge.org/distributed/ (Utolsó elérés: 2023. 08. 18.)
[71] Wikidata, *Kiszl Péter*, Elérhető: https://www.wikidata.org/wiki/Q113145828 (Utolsó elérés: 2023. 08. 18.)
[72] Wikidata, *User group rights*, Elérhető: https://www.wikidata.org/wiki/Special:ListGroupRights (Utolsó elérés: 2023. 08. 18.)
[73] MediaWiki, API:Etiquette, Elérhető: https://www.mediawiki.org/wiki/API:Etiquette (Utolsó elérés: 2023. 08. 18.)
[74] MediaWiki, *Manual:Maxlag parameter*, Elérhető: https://www.mediawiki.org/wiki/Manual:Maxlag_parameter (Utolsó elérés: 2023. 08. 18.)
[75] Wikidata, *Wikidata:Bots*, Elérhető: https://www.wikidata.org/wiki/Wikidata:Bots (Utolsó elérés: 2023. 08. 18.)
[76] Wikidata, *Statistics*, Elérhető: https://www.wikidata.org/wiki/Special:Statistics (Utolsó elérés: 2023. 08. 18.)
[77] Piscopo, A., Phethean, C., Simperl, E. *What makes a good collaborative knowledge graph: group composition and quality in wikidata*, In: Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I, Springer, Cham, 2017, p. 305–322. https://doi.org/10.1007/978-3-319-67217-5

Beérkezett: 2023. augusztus 25.



Molnár Bence
a PTE BTK Könyvtár- és Információtudományi Tanszékének egyetemi tanársegéde,
az Európai Unió Kiadóhivatalának munkatársa
ORCID: https://orcid.org/0009-0002-0274-7784
E-mail: molnar.bence@hotmail.com