A kódexektől a magára ismerő szövegig

PANEK SÁNDOR
Fotó: Nagy Dóra

A Nemzeti Névtér (abcd.hu) projekt és az Országos Széchényi Könyvtár készülő országos platformja a Google-kereséseinket is érinti majd. A könyvtárosok azt várják: könnyebb lesz hiteles adatokat találni a weben.

Az Országos Széchényi Könyvtár Nemzeti Névtér projektje és készülő egységes országos könyvtári platformja, valamint a törekvés, hogy a webtérben a kapcsolatok rendszerezetten jöjjenek létre, éveken belül átalakítják a magyar könyvtári adatkincs megosztását. A könyvtárosok szerint izgalmas időszak következik, amelynek végeredménye az lesz, hogy az információk a virtuális térben is a valóságnak megfelelő kapcsolatokba kerülnek. A nehezen kereshető mély web aljáról előkerülnek a kapcsolódó adatok, összefüggéseikben találunk az információkra.

Kokas Károly és *Nagy Gyula*, az SZTE Klebelsberg Könyvtár igazgató-helyettesei egyetértenek: nagyot fordulhat a világ a könyvtáros szakmában.

Megvalósulhat-e az álom, hogy a könyvtárak adatkincse a legszélesebb körben, a weben legyen elérhető?

Nagy Gyula: – Legalábbis egyre közelebb kerülünk ehhez. A könyvtárak katalógusai most is elérhetők a neten, sőt, együtt is elérhetők, a Mokka nevű szolgáltatásban. Az Országos Széchényi Könyvtár új projektje ezt egy szinttel feljebb viszi, mert nemcsak az úgynevezett metaadatokat (például a bibliográfiai adatokat) fogja elérhetővé tenni, hanem egy összekapcsolt felhőben közös felhasználókezelést, közös könyvfeldolgozást, vagyis egy országos integrált platformot fog jelenteni.

Kokas Károly: – Az integrált könyvtári rendszereket úgy kell elképzelni, mint az egészségügyi pénztár receptrendszerét. Az orvosok receptet írtak, s a felhasználó az ország minden patikájában kiveheti a felírt gyógyszert. Egy-egy számítógép minden rendelőből a központi géppel kommunikál, s azt látja a patikus is, bárhol legyen is földrajzilag. Az OSZK-projektben is egyetlen országos rendszert telepítenének a felhőbe, ehhez pedig központilag csatlakoznának a könyvtárak.

Névterek: a magyar digitális tudás horgonyai

Mik is ezek a névterek? Az látszik a Nemzeti Névtér oldalán, hogy a magyar könyvtárak összesített adatkincse e nevek köré szerveződik. De hogyan lesz ez az adatbázis több egy újabb digitális lexikonnál?

Kokas Károly: – Az országos könyvtári platform adatai a névtereken keresztül kerülnek be a szemantikus webtérbe. A Nemzeti Névtér weboldalán is látszik: a magyar személy-, földrajzi és testületi nevek hitelesített és összefüggésbe helyezett adatokkal kerülnek fel. Ezek alkotják a névtereket. Annak mértékében, ahogyan sorra jönnek létre a hitelesített névterek mint fix keresési pontok, a Google-keresésünk is relevánsabb, több összefüggést kínáló találatokat nyújt majd. Nézzünk egy példát! A Wikipediában egy magyar író oldalán láthatók a művei; ha alapos a szócikkíró, akkor minden mű benne van, ha őrülten elhivatott, akkor lehet, hogy minden kiadás is. Mihelyt összekapcsoljuk az író Wikipedia-szócikkét a névterrel, átjárót kapunk az összes könyvtár katalógusaihoz. Innentől ugyanaz a „Babits Mihály” név mint fix pont kapcsolja össze a Wikipédiát a könyvtárak katalógusával. A könyvtárak által növekvő számban létrehozott névterek azután egyre több weboldalhoz kötődnek majd. A Wikipédián ez gyorsan megy, de idővel egy jobb színház műsoroldalán is a szerzőre kattintva ez a fix pont érhető el, ahol minden érvényes Babits Mihály-hivatkozás összekapcsolódik. Jelenleg még a sűrűn előforduló nevek esetében a találatok összekeverednek a Google-ban.

Nagy Gyula: – Akik a magyar névteret építik szemantikus webes alapon, számos meglévő adatfor-

rást használnak fel, köztük könyvtári katalógusokat, életrajzi lexikonokat. Ezeket az adatokat és összefüggéseiket könyvtárak, levéltárak, múzeumok hitelesítik. Napjainkban a könyvtárak, egyfajta modern "hiteles helyként" tekintve magukra, fontos küldetésüknek tartják az információk megbízhatóságának ellenőrzését. A félinformációktól, átverésektől, áltudományos tartalmaktól, "fake news-októl" hemzsegő világban erre a tevékenységre minden eddiginél nagyobb az igény.

Szintek a Szegedi Tudományegyetem Klebelsberg Könyvtárában. Fotó: Kokas Károly

Kokas Károly: – Egy másik példa: a washingtoni Kongresszusi Könyvtár nagy figyelmet fordít a szerzői profilokra. Ennek ellenére Jókai Mór névváltozataira eltérő találatokat kapunk, nem is beszélve az idegen nyelvi változatokról. Vagyis, szerzői szempontból a Jókai-névvariánsok még feldolgozatlanok. Ha a Kongresszusi Könyvtár adatai bekerülnek a névtérbe, összekapcsolódnak, és attól kezdve mindegyik találata összefüggésbe kerül. Amint a névtér elfogadott lesz a világban, egyre több szereplőnek lesz érdeke, hogy a neveket összekapcsolja. Ugyanakkor persze, a könyvtárak katalógusai eltérő szintűek. Míg a Klebelsberg Könyvtáré mély, egy vidéki kiskönyvtár katalógusában nem feltétlenül van meg, hogy egy könyv illusztrált-e, illetve van-e előszava.

Nagy Gyula: – A szemantikus weben akár történelmi figurákra is lehet keresni. Egy könyvtári katalógusban ez behatároltabb, hogy ne legyen túl bonyolult a keresés, de van könyvtári szoftver, ami már e keresés határait is feszegeti. Éppen szegedi műhely, a Monguz fejleszt egy könyvtári katalógus szoftvert, amely nemcsak az egyes könyveket próbálja leírni, hanem azokat a mű szintjén is. Az Egri csillagokra keresve például nemcsak egy listát kapok a könyvtárban előforduló kötetekről, hanem azok kategorizálva lesznek. Különválasztva például a fordítások, adaptációk, rövidítések, filmváltozatok.

„A tudás fog megtöbbszöröződni." Nagy Gyula és Kokas Károly az SZTE Klebelsberg Könyvtárban.
Fotó: Nagy Dóra

A netes információkeresés egyik problémája, hogy a web túlságosan mély. Egyes adatok azért maradnak rejtve, mert a keresés nem volt eléggé összetett. Hogyan lesz jobban megtalálható a könyvtári adat ebben a mélységben?

Kokas Károly: – A szemantikus web döntő pillanata, hogy a könyvtári rekordok indexelődnek is a Google-ban. Ha most beírjuk a Google-ba az "Egri csillagok képregény" kifejezést, akkor webáruházakat, gyűjtői oldalakat kapunk. A szemantikus webtérben, nagy eséllyel az első találatok könyvtári címleírások lesznek. Mivel a Google figyeli, hogy hol vagyunk, így akkor a hozzánk legközelebb eső könyvtár katalógusát fogja mutatni.

Nagy Gyula: – Ami indexelhetővé, jobban láthatóvá teszi ezeket a könyveket, az a szemantikus webnek egy belső szabványos késztetése, mégpedig, hogy minden egységnek, például személynek, fogalomnak stb., egyedi URI-val, vagyis állandó hivatkozással kell rendelkeznie. A mélywebes tartalmak viszont sokszor dinamikusak, nincs állandó címük, hanem egy adatbázisból generálódnak ki. Ez az állandó URI teszi lehetővé, hogy a keresők körbe tudják járni és feltérképezni a tartalmat. Másfelől, ha személyre, helyre keresünk a Google már most wikipediás tartalmat tesz az első helyre, sőt, oldalra a saját dobozát is elhelyezi belőle. Ezt a Wikidata teszi lehetővé, amely strukturált, gép által olvasható. Így például egy személy születési dátuma fel van címkézve, s nem

a Google-nak kell kitalálnia, hogy az a négyjegyű szám évszám-e.

A könyvtár már több, mint a könyvtár volt

A könyvtári adatvagyon hány százalékban jelenik meg a magyar nyelvű weben? Milyen arányban van digitalizálva a könyvtárakban elérhető tartalom?

Nagy Gyula: – Ez utóbbi könyvtáranként nagyon változó. A Klebelsberg Könyvtárban mintegy 3 millió oldal digitalizált tartalomnál tartunk, ami nagy munka, de a teljes állomány elenyésző százaléka. Ha a weben elérhető digitális tartalmat nézzük, a magyar kiadású napilapok, folyóiratok 80–90 százaléka ott van valamelyik adatbázisban. Az ADT és a Hungaricana oldalain 30–35 millió oldalnyi tartalom érhető el. Az Akadémiai Könyvtár repozitóriuma néhány millió oldalas. Ha összeadjuk a többi néhány millió oldalas tartalmat, kijönne még 35 millió. Ugyanakkor az összes magyar digitális sajtótermék, könyv együtt több száz millió oldal. Vannak országok, mint Ausztrália vagy Új-Zéland ahol elmondhatják, hogy a periodikák tartalma 100 százalékban elérhető a weben.

Kokas Károly: – Ha nyernénk 2 milliárd forintot, egy légkondis épületben, felszerelve modern eszközökkel 60 tanítvánnyal 3 műszakban digitalizálva 5 éven belül az egész magyar könyvtári kultúra még nem digitalizált tartalmát mindenestül talán be lehetne digitalizálni. Csakhogy az, hogy egy könyv elolvasható-e a neten fizetősen vagy ingyenesen, alapvetően nem könyvtári, hanem jogi kérdés. Amíg a globális szerzői jogi szabályozás nem változik, addig ezt nem lehet könyvtár-informatikával megoldani.

Mit gondolnak, a könyvtárak látogatottsága növekedni fog-e attól, ha a Google-ban az első találat vezet a könyvtári katalógusokra?

Kokas Károly: – A könyvtári szolgáltatást ma már szélesebben kell érteni. Van olyan hallgató, aki úgy gondolja, hogy neki nincs már dolga a könyvtárral, mert mindent weben intéz. Eközben pedig a napja úgy indul, hogy megnézi a repozitóriumokat, s a könyvtár által előkészített tartalmak között keresgél. Ő talán arra szavazna: nincs szükség könyvtárra. Még az egyetemi térben is sokan gondolhatják, hogy van a könyvtár, ahol kardigános nénik őrzik a melegben az olvasókat, és van az internet, ahol nagyszerű dolgok fejlődnek, a könyvtáraktól függetlenül. Ugyanakkor pedig a Klebelsberg Könyvtárban több, mint félmilliárd dokumentum érhető el a katalóguson keresztül. Van olyan kutató, aki 10 éve nem használt más digitális tartalmat, mint amit a könyvtár megvásárol. Ez azt jelenti, hogy ki tudjuk szolgálni az egyetemi kutatás igényeit.

Nagy Gyula: – A Klebelsberg Könyvtárban azért nem aggódunk ezen a kérdésen, mert az elmúlt pár évben mind a könyvtárba bejövők száma, mind a virtuális térben valamelyik anyagunkat meglátogatók száma növekedett. A könyvtár közösségi tér funkciója pedig kivált erősödni látszik. Idén talán először fordult az elő, hogyha még néhány tucat olvasó bejön délelőtt, már nem mindenkit tudtunk volna leültetni az olvasótermekben, pedig mintegy 1000 ülőhelyünk van összesen.

Kokas Károly: – Ebben a külföldi hallgatók sokasodása is szerepet játszik, mivel nekik az a munkastílusuk, hogy reggel bejönnek, lepakolnak, elmennek az óráikra, majd visszatérnek a könyvtárba délutánra, mert itt találják meg a szakkönyveiket és a közösségeiket. Itt szeretnek dolgozni. A magyar hallgató viszont jellemzően kikölcsönzi a könyveit és hazacipeli. Sokszor akkor is, ha egy 5 oldalas tanulmányt az olvasóteremben gyorsabban elolvasott volna, mint ameddig a kölcsönzés folyamata tartott.

Több száz millió adat keres kapcsolatot

Ha jól értem, a könyvtárak több száz milliós nyomtatott adatkincsének katalóguscéduláit kell most minél több szempont szerint egymással összefüggésbe hozni. Mi ennek a módszere?

Kokas Károly: – Ahhoz, hogy a szemantikus weben hibátlanul tudjanak működni, a Klebelsberg Könyvtár katalógusain adattisztítást végzünk. Király Péter a göttingai egyetemen a könyvtári MARC rekordok, vagyis géppel olvasható katalógusadatok adattisztításából írt disszertációt. Az ő algoritmusaival is végeztünk elemzést és kitűnő kollégánk, Bernátsky László is már évek óta végez adatkonszolidációs műveleteket a katalóguson. Mihelyt megvannak a típushibák, le lehet futtatni a javító algoritmusokat, s csak ezután jön a kézi feldolgozó munka. Az adattisztításban és az összefüggések építésében a közösségi munkavégzés, vagyis a felhasználó bevonása is nagy lehetőség. A Nemzeti Névtér oldalán már felmerült, hogy szívesen hoznának létre hozzáférést azok számára, akik a saját szakterületükön többet tudnak az adott személyiségekről, mint amit a névtér tud. Amikor a Délmagyarországot digitalizáltuk, mi

is álmodoztunk arról, hogy a felhasználókat megkérjük, pontosítsák a cikkek tárgyszavazását. Ha valakit a kézilabda érdekel, arra kaphatna elérést, hogy a Délmagyar adatbázisában ezzel a sportággal foglalkozó cikkeket tárgyszavakkal lásson el, és hozzon egymással összefüggésbe. A megfelelő szabályokat betartva, kézilabda témában a Délmagyarország archívuma szinte 100 százalékosan be tudna kapcsolódni a szemantikus webbe.

Nagy Gyula: – Én a mesterséges intelligencia bevonásában is hiszek: egy 100 évfolyam fölötti napilap 300 ezer oldalnyi információját emberi erővel nehéz befogni. Egyelőre alacsonyabb színvonalon vannak, de a kézi munkánál gyorsabbak a már most létező szövegbányászati algoritmusok. Az automatikus kulcsszavazás szintén egyre jobban terjed. Ezek a módszerek egymással kiegészítve jól működhetnek: a mesterséges intelligencia segítségével előfeldolgozott anyagot a felhasználó bevonásával lehet feldolgozni. Ezután civil szereplők és könyvtárosok ellenőrizni, utófeldolgozni tudnák az anyagot.

Kokas Károly: – „Az olvasó a gép”, hogy a korán elhunyt szegedi irodalomtörténész, Labádi Gergely címével éljek, misztikus az átlagember számára. De valójában a számítógép alkalmassá tehető arra, hogy személyes olvasás és gondolkodás útján felállított tudományos hipotéziseket igazoljon vagy cáfoljon. Még az irodalmi stílus kérdései is felvethetők számítógépes módon. Rá lehet-e venni a gépet, hogy különbséget tegyen Móricz és Mikszáth irodalmi stílusa között? Matematikailag meghatározható ez a különbség? Még ebben az évtizedben kiderült, hogy igen, a gép nemcsak keresni tud a szövegben, de már stilisztikai jellemzőket is felismer. Ez óriási továbblépés lesz ahhoz képest, hogy egy adott név előfordul-e abban a szövegben.

„Ha még néhány tucat olvasó bejön délelőtt, már nem mindenkit tudtunk volna leültetni” Fotó: Kokas Károly

Van olyan dokumentum, amit senkit sem keresett az idők során?

Kokas Károly: – Most még igen, de amint ez a tartalom fel van indexelve és ontológiákkal kiegészítve, akkor a benne végzett keresés egyben kutatást is jelent majd; *search as research*, ahogyan angolul mondják. A jó keresés önmagában tudományos eredményeket képes produkálni. Vannak adatok, amelyeket nem is lehetett volna felfedezni, mivel azelőtt nem voltak összekapcsolhatók.

Nagy Gyula: – Biztosan vannak nem keresett dokumentumok: a könyvtári adatkincsbe a kínai szakpublikációk is beletartoznak. Éppen az óriási adatmennyiség miatt van szükség az új módszerekre, mint a névtér és a szemantikus web, hogy ne csak a felszínét érjük el az adatmennyiségnek. A Google kereséseknél is csak az első egy-két oldalt nézzük meg: ha a legjobb találat nincs ott, akkor az nem létező tudás marad. A Klebelsberg Könyvtárból elérhető kb. 500 milliós dokumentumnál ez még inkább így van, mivel nincs a keresés mögött egy Google-szintű súlyozó algoritmus. A Google is éppen e probléma miatt kezdett szemantikus webes fejlesztésekbe.

Mit gondolnak, a felhasználók okosabbak, tudatosabbak lesznek az információ könnyebb elérésétől?

Nagy Gyula: – Ezt egyelőre nem állítanám. A Facebook ajánló algoritmusaiból látszik, hogy ez a tudás a buborék-hatás jelenségét hordozza magában, vagyis az ember csak azzal találkozik, amivel egyetért. Ez nem segíti a tudás terjedését.

Kokas Károly: – Okosabbak magunktól nem leszünk. A bennünk lévő tudás fog megtöbbszöröződni. Nekünk kell akarni keresni, s tudni, hogy mit várunk, várhatunk el. Mindez inkább óriási lehetőségeket nyit meg a felhasználók számára. Olyan ez kicsit, mint amilyen talán az önvezető autó lesz 5-10 év múlva. Ha ez a rengeteg tudás-előkészítés és elő-rendszerezés bekapcsolódik a mai világhálózatba, a kereséseink és a találataink biztonságosabbak lesznek. Mi vezetünk, mi mondjuk meg hova megyünk, de az intelligens gépek, a rendszerezettebb adatok sokat segítenek, hogy oda is érjünk, ahová tényleg menni akarunk. Vagyis, hogy a valóban fontos és hitelt érdemlő dolgokat találjuk meg. Gyorsan és biztonságosan.

Forrás: https://www.delmagyar.hu/kultura/helyi-kultura/a-kodexektol-a-magara-ismero-szovegig-4745004/?fbclid=IwAR2ICpMhe7qHZABJl06f9ujZtJ5dTJueIzXSqfZlFr6D09TDbL7ow142MfE

Válogatta: Fonyó Istvánné