

# SZÖVEGBÁNYÁSZAT, ADATBÁNYÁSZAT, ISMERETFELTÁRÁS ÚJ LEHETŐSÉGEK A TUDOMÁNYOS KOMMUNIKÁCIÓBAN

Holl András

informatikai főigazgató-helyettes,  
MTA Könyvtár és Információs Központ  
holl.andras@konyvtar.mta.hu

2014 decemberében a LIBER<sup>1</sup> 25 szakértőt hívott meg a tudományos ismeretfeltárás aktuális kérdéseinek megbeszélésére Hágába, a Holland Nemzeti Könyvtárba. A megbeszélés eredménye a *Hágai Deklaráció* (URL) első fogalmazványa.

A tudományos kommunikáció egyik izgalmas és fontos kérdése az exponenciálisan bővülő tudományos ismeretanyagban való tájékozódás (Holl, 2013). Egyrésztől informatikai támogatás nélkül a kutatók egyre kevésbé képesek hatékonyan tájékozódni az információáradatban, másrésztől az új informatikai lehetőségek kihasználásával új távlatok nyílnak meg előttük. A tájékozódást segítő, az információözünt kezelő eljárások egyik legfontosabb csoportját szövegbányászatnak, adatbányászatnak, ismeretfeltárásnak nevezük. Ebben a cikkben a szövegbányászatra (tudományos publikációk nagy méretű hal-

mazaiban való adatbányászatra) összpontosítunk. Mivel tudományos publikációkban a szövegen túl adatok, képek is találhatóak, a tudományos szakirodalom informatikai eszközök segítségével történő feltárásában a szövegbányászat gyakorta adatbányászattal is társul – az angol nyelvű szakirodalom ezért a *text and data mining (TDM)*, vagy a *knowledge discovery* kifejezéseket alkalmazza.

A szövegbányászat a nagyméretű, szöveges adatbázisokban való, teljes szövegű kereséssel kezdődik, de további lehetőségeket is magában foglal, olyan eljárásokat, amelyeknek a szöveg, a szövegösszefüggések elemzése is része. A keresőszolgáltatásokat nyújtó vállalkozások hatalmas adatbázisokat építenek az interneten elérhető szövegekből, melyekben megkereshetjük például egy idézet forrását, szöveggörnyezetét. A keresőmotorok arról is gondoskodnak, hogy egy-két karakter elgépelése, a némileg pontatlanul felidézett szövegtöredék használata is jó eséllyel eredményt adjon. A kommersz keresők a kutatóknak is gyakorta jó szolgálatot tesznek. Ebben a cikkben olyan alkalmazásokkal foglalkozunk,

---

<sup>1</sup> LIBER – Ligue des bibliothèques européennes de recherche / Association of European Research Libraries. Az MTA Könyvtár és Információs Központ is tagja a szervezetnek.

amelyek nagy publikációs adatbázisok használatával, informatikai eljárások segítségével felhasználást találhatnak a tudomány területén.

A tudományban azon túl, hogy az egyszerű teljes szöveges keresésnél gyakran bonyolultabb eljárásokra van szükség, különbözik a szöveges adatbázis is. A szövegbányászat jól körülhatárolt adatbázison: egy szélesebb vagy szűkebb szakterülethez vagy tudományterülethez sorolható publikációk nagyméretű halmazán operál. A publikációk digitális szövegei származhatnak a kiadóktól vagy repozitóriumokból. A következőkben az ismeretfeltárás néhány példáját ismertetjük.

\*

Az MTA Nyelvtudományi Intézetének egyik projektje a Magyar Nemzeti Szövegtár. A nagyméretű szövegtár (1,5 milliárd szövegszó) különböző forrásokból, eltérő stílusrétegekből épült fel: például a sajtóból, a szépirodalomból, a szociális médiából. A tudományos (inkább népszerűsítő) szövegek csak kis részét teszik ki, forrásuk a Magyar Elektronikus Könyvtár (Oravecz et al., 2014). Ebben az esetben a szövegbányászat célja a szavak összegyűjtése – a korpusz és a ráépülő szolgáltatások nyersanyagul szolgálnak további kutatásokra, nagyszótár létrehozására.

Ugyancsak az MTA Nyelvtudományi Intézetének projektje a MATRICA (Magyar Társadalomtudományi Citációs Adatbázis; Váradi et al., 2014). A projekt keretében mintegy 190 hazai szakfolyóirat több éves teljes anyagának feldolgozásával nyelvtechnológiai eszközöket fejlesztettek irodalmi hivatkozások kinyerésére. Az eljárás a lényege a hivatkozás szövegben való felismerése, majd a hivatkozott mű bibliográfiai adatainak azonosítása. A cél bibliometriai adatok szolgáltatása olyan tudományterületeken, ahol nemzetköz-

zi citációs adatbázisokra nem támaszkodhatunk. (A bölcsészet és társadalomtudományok idézettségi adataival bővebben foglalkozik Dudas Anikó [2013]).

A szövegbányászat speciális esetének tekinthetjük a plágiumkeresést. A digitális könyvtári tartalmak növekedése – talán meglepő módon – nem kedvez a plagizálásnak: a nagymértékben hasonló dokumentumok könnyen azonosíthatóak. A plágiumkereső rendszerek terén említhetjük az MTA SZTA-KI KOPI-t, vagy a tudományos kiadók CrossRef rendszerének CrossCheck szolgáltatását. Ez utóbbi az iParadigms cég iThenticate szoftverével működik. A szerző saját gyakorlatából is beszámolhat a plágiumkeresés gyakorlati hasznáról. A Library and Information Systems in Astronomy VII. konferencia kiadványának szerkesztésekor is sikerült a szerző által önplagiarizált (korábbi konferenciakiadványban már megjelent, szóról-szóra egyező) kéziratot kiszűrni. A KOPI már egy mobiltelefonnal lefotózott oldal alapján képes az eredeti szerzőt megtalálni (Pataki et al., 2014).

A bibliometriából a szcientometria területe átlépvé megemlítjük a kutatási nagyberendezések használatának és hasznosulásának vizsgálatát a szakfolyóiratokban megjelenő adatok elemzésével. A csillagászat területén számos obszervatóriumi rendszerben működő – a távcsőidőt (műszeridőt) pályázati rendszerben elosztó – nagyberendezés működik, mint a Hubble űrtávcső, vagy az Európai Déli Obszervatórium (ESO) nagytávcsövei a chilei Andokban (köztük a négy 8,2 méteres óriástávcsőből álló Very Large Telescope). A kutatók pályázatokon nyernek műszeridőt, s a pályázatok esetében követelmény az elért eredményekről való beszámolás és a pályázati azonosítók szerepeltetése a megfigyelések alapján publikált cikkekben. Elvárható, hogy

a kutatók a létrejött elsődleges publikációikat a beszámolóokban felsorolják – de a beszámolás többnyire határidőhöz kötött, és a később létrejött, az adatokat felhasználó további vagy másodlagos cikkekről a nagyberendezések tulajdonosa már nem feltétlenül értesül. A nagyberendezésekkel készült megfigyelések adatai gyakran egy idő után szabadon elérhetőek lesznek. A tudományos etika és a publikálás alapkövetelményei szerint az adatok forrását a felhasználásukkal készült cikkek meg kell említsék – de a berendezésekre való irodalmi hivatkozások összegyűjtése már pusztán emberi erővel nem végezhető el. Chris Erdmann és Uta Grothkopf (2010) az *ESO telbib* teleszkóp-bibliográfiai adatbázisát, és a készítéséhez használt *ESO Full-Text Search Toolt (FUSE)* ismerteti. Jill Lagerstrom (2015) a szakmában kialakult „jó gyakorlatról” számol be. Ezekben az esetekben a szövegbányászat pusztán a publikációk leíró adataiban és a teljes szövegben való kifejezések keresésével történő nyers bibliográfiai lista létrehozására szorítkozik, a tényleges műszerhasználat azonosítása könyvtárosok munkájával történik. A korábban ismertetett *MATRICA* gépi szövegfeldolgozást alkalmaz, de a rendszer jelenlegi állapotában ott is szükség van még emberi felügyeletre. Az alkalmazott eszközök fejlődésével a jövőben lehetőség lesz mind kutatók, mind pályázati projektek, nagyberendezések vagy intézmények eredményeinek szcientometriai követésére és elemzésére.

Korábban beszámoltunk már (Holl, 2013) az *Astrophysics Data System* szakirodalmi tájékozdást segítő funkcióiról: egyszerű lehetőség egy relevánsnak bizonyult cikkhez hasonló további cikkek keresésére a tartalmi kivonatban, vagy a teljes szövegben előforduló szavak súlyozott gyakoriságának összehasonlítása (Kurtz – Henneken, 2012).

A szakirodalmi tájékozdás fejlettebb, nyelvtechnológiai eszközöket alkalmazó lehetőségeiről számolt be Frédérique Lisacek, Christine Chichester, Aaron Kaplan és Sándor Ágnes (2005). A szerzők által a szövegben alkalmazott retorikai elemek is árulkodnak a tudományos tartalom újdonságértékéről. A *PubMed* adatbázisból származó orvosbiológiai cikkek kivonatainak elemzésével listát készítettek olyan fehérjékről, amelyeket a cikkekben neurodegeneratív betegségekkel hoztak kapcsolatba. A listában szereplő fehérjék egy részének neurodegeneratív betegségekkel való kapcsolatát a vizsgálat időpontjában az áttekintő cikkek nem említették – viszont a következő évek során publikáltakban már szerepeltek. Informatikai eszközökkel gyorsan lehet a szakirodalomból fontos információkat kiszűrni. A 2013-as genfi OAI8 (Open Archives Initiative – CERN Workshop on Innovations in Scholarly Communication) konferencián Sándor Ágnes a Xeroxnál fejlesztett szoftvereszközök további alkalmazásairól is beszámolt: lehetséges a cikkeken belül a lényeges részek, kulcsfontosságú állítások automatikus kiemelése (Sándor – Vorndan, 2010).

Egy-egy csillagászati objektum akár több tucatnyi névvel, azonosítóval rendelkezhet: ugyanazt az objektumot katalogizálhatták például infravörös és rádiótartományban végzett felmérések során. A strasbourgi *SIMBAD*-adatbázis összefoglalja az objektumok adatait és azonosítóit. A szakirodalom feldolgozása során egy *DJIN* nevű szoftvert alkalmaznak az objektumnevek kinyerésére (Lesteven et al., 2010). A *DJIN* alkalmazása is szakértői felügyeletet igényel – de jelentősen megkönnyíti az emberi munkát. Egyes csillagászati folyóiratok a cikkekben felcímkézik az objektumneveket (így jár el az *Astronomy & Astrophysics* és az *MTA Csillagászati és*

Földtudományi Kutatóközpont Konkoly Thege Miklós Csillagászati Intézet (CsFK CsI) által kiadott kis folyóirat, az *Information Bulletin on Variable Stars*), ami megkönnyíti a cikkek indexelését. A többi folyóiratnál informatikai eszközök – mint a DJIN – alkalmazására van szükség, a szakirodalom mára már pusztán emberi munkaerővel nem preparálható.

A georeferálás (különböző objektumok földrajzi koordináta-rendszerbe illesztése) mintájára beszélhetünk genoreferálásról is: egy valamiféle kísérlet során meghatározott bázispársorozat (DNS-darabka) elhelyezéséről az adott élőlény teljes genomjában. A text2gene projekt (Haeussler et al., 2011) célja a szakirodalomban publikált bázispár-sorozatok összekapcsolása genetikai adatbázisokban fellelhető génekkel, a genom szakirodalmi annotálása. 2011-ben a *PubMed Central* 150 000 cikkét dolgozták fel. A kutatók dolgoznak további szövegforrások feldolgozásán – de a különböző folyóiratok kiadóinak engedélyét megszerezni a szövegbányászatra hónapokba, évekbe kerül tapasztalatuk szerint. A projekt eredményeként a genom adatbázishoz kapcsolható lesz az egyes génszakaszokkal foglalkozó szakirodalom – mint ahogy a digitális várostérképeken megtekinthető, milyen fotót töltek fel egy adott utcasarokról.

\*

A tudományos folyóiratok cikkeit begyűjteni – még ha a szövegbányászattal próbálkozó kutató intézménye érvényes előfizetéssel rendelkezik is – nem feltétlenül egyszerű. Bár a kutatók az egyes cikkeket egyenként letölthetik, a robotokkal történő tömeges cikletöltés könnyen az elérés blokkolását eredményezheti. Egyes kiadók, folyóiratok támogatják a szövegbányászatot – a saját elképzelése-

ik szerint. Ilyen az Elsevier és a nyílt hozzáférésű Public Library of Science (*PLOS*). A fogadtatás vegyes: a kutatók részben az egyes szabályokat vitatják, részben azt, hogy az egyes kiadók szabályozásai egymástól különböznek, és a hozzáférési jogok megszerzése jelentős terhet jelent. A szövegbányászatban rejlő lehetőségek jobb kihasználásához részben nyílt hozzáférésre, részben egységesítésre lenne szükség.

Casey Bergman, a Manchesteri Egyetem bioinformatikusa 2012-es blogbejegyzésében (URL2) fel is teszi a kérdést, miért nem használják ki jobban a *PubMed Central* által nyújtott adatbányászati lehetőségeket? Mindazonáltal felsorolja azokat a cikkeket, amelyek a repozitóriumban található teljes Open Access anyag felhasználásával készültek.

A Hágai Deklaráció – e cikk írásának idején nem végleges formában, szabadon kommentálhatóan – az ismeretfeltárás jogi kérdéseire összpontosít. Azonban a kiadók fenntartásai között technikai jellegűek is vannak: a nagyméretű PDF-fájlok tömeges letöltése túlzott terhelést jelenthet a kiszolgáló számítógépeken.

A cikkek robotok által való, tömeges szüretelésének engedélyezése véleményünk szerint nem elegendő – technikai támogatásra és megegyezésekre is szükség van. A folyóiratok honlapjain található cikkek emberi „fogyasztásra” készültek. A szöveg kibontása PDF- vagy HTML-állományokból nem feltétlenül könnyű. Véleményünk szerint a következő lépésekre lenne szükség:

- a cikkek szövegének gépi formában való feldolgozásra alkalmas formában (XML, TXT) is elérhetőnek kellene lennie;
- külön tartalmi kivonatokat kellene készíteni automatikus feldolgozásra, a cikk lényeges állításainak szemantikus web

szabványok szerinti kódolásával, nano-publikációs (URL<sub>3</sub>) formában;

- automatikusan feldolgozható változatban mellékelni kellene a táblázatokat;
- az adatokból rajzolt ábrákhoz mellékelni kellene az adatokat, a képekhez metaadatokat kellene társítani;
- a cikkeket arató robotok számára érthetővé kellene tenni a cikkhez tartozó állományok viszonyát, azt, hogy gépi vagy humán felhasználásra valók, továbbá a szövegbányászati jogosítványokat.

A tartalmak egyszerű szöveges vagy XML-formátumban való elérhetővé tétele választ adna a terheléssel kapcsolatos aggodalmakra is. Az öt felsorolt javaslat közül az első és az utolsó egyszerűen megvalósítható lenne. A hazai *Information Bulletin on Variable Stars* számai szabadon letölthetőek LaTeX-formátumban. Az XML sokkal jobb lenne, de a LaTeX is megfelel a szövegbányászati céljaira. Mindössze arra lenne szükség, hogy a szüretelő robotoknak jelezhessük, melyiket töltsék le a rendelkezésre álló formátumok közül, és melyik a letöltött állomány emberi szemnek szánt változata. Jelenleg a robotok válogatás nélkül mindent letöltenek.

A többi három javaslat – és az első javaslat XML opciója – nehezebben megvalósítható. Mind a kiadóknak, mind a kutatóknak viszonylag nagyobb mértékben változtatni kellene a jelenleg követett gyakorlaton. Ahhoz, hogy a javaslatok kivitelezhetőek legye-

nek, szabványokra volna szükség, és arra, hogy a műszergyártók és a szoftvergyártók ezeket termékeikbe beépítsék. Az új formátumokra, szoftverekre, szabványokra való törekvés már egy idő óta jelen van a tudományban – a bölcsészettudományokat is beleértve (Kecske-méti, 2014). A javaslatok részben a szövegbányászati megkönnyítést célozzák – részben a feje tetejéről a talpára állítják az információfel-tárás kérdését. A megfelelően preparált információban sokkal könnyebb keresni – a Google nyers ereje a metaadatok alkalmazásával szemben. Ahogy Barend Mons megfogalmazta: „Minek az információt eltemetni, ha úgyis ki akarjuk bányászni?”

A szövegbányászati nagy mennyiségű, digitális formában elérhető publikáció feldolgozása alapszik. A publikációk begyűjtése történhet a kiadóktól, de repozitóriumokból is. A repozitóriumok száma öröndetesen nő, tartalmuk gyarapszik itthon is. Az MTA KIK (Könyvtár és Információs Központ) repozitórium – a REAL – gyarapításánál is szempont a majdani szövegbányászati felhasználás lehetősége. Egyszerű funkciókat – mint a teljes szövegű keresés – már használni lehet. A hazai repozitóriumok aggregálásának első lépése pedig az MTA Számítástechnikai és Automatizálási Kutatóintézete (SZTAKI) által fejlesztett közös kereső lehet.

**Kulcsszavak:** *könyvtártudomány, információtudomány, szöveg- és adatbányászat, repozitóriumok*

## IRODALOM

- Dudás Anikó (2013): *Hivatkozásokra vezérlő kalauz – bölcsészet és társadalomtudományok*. NETWORKSHOP 2013. <http://nws.niif.hu/ncd2013/docs/ehu/045.pdf>
- Erdmann, Chris – Grothkopf, Uta (2010): Next Generation Bibliometrics and the evolution of the ESO Telescope Bibliography. LISA VI Proceedings,

ASP Conf. Ser. 433, 81. • <http://adsabs.harvard.edu/abs/2010ASPC...433...81E>

- Haeussler, Maximilian – Gerner, M. – Bergman, C. M. (2011): Annotating Genes and Genomes with DNA Sequences Extracted from Biomedical Articles. *Bioinformatics*. 27, 980. • <http://dx.doi.org/10.1093/bioinformatics/btr043>

- Holl András (2013): Információáradat és hullámlövágás. *Magyar Tudomány*. 4, 473–478. • <http://www.matud.iif.hu/2013/04/13.htm>
- Kecskeméti Gábor (2014): Electronic Textual Criticism. In: *Dávidházi Péter (ed.): New Publication Cultures in the Humanities*. Amsterdam Univ. Press • <http://www.open.org/search?identifier=515678>
- Kurtz, Michael J. – Henneken, Edwin A. (2014): Finding and Recommending Scholarly Articles. In: Cronin, Blaise – Sugimoto, Cassidy R. (eds.): *Beyond Bibliometrics*. MIT Press. • <http://arxiv.org/abs/1209.1318>
- Lagerstrom, Jill (2015): *Best Practices for Creating and Observatory or Telescope Bibliography from the IAU Commission 5 Working Group on Libraries*. LISA VII Proceedings, ASP. Conf. Ser. 492, 99
- Lesteven, Soizick et al. (2010): *DJIN: Detection in Journals of Identifiers and Names*. LISA VI Proceedings, ASP. Conf. Ser. 433, 317 • <http://www.iucaa.ernet.in:8080/jspui/bitstream/11007/1722/1/Lesteven.pdf>
- Lisacek, Frédérique – Chichester, C. – Kaplan, A. – Sándor Á. (2005): *Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases*. First International Symposium on Semantic Mining in Biomedicine, Cambridge, UK • <http://www.aaronkaplan.info/publications/2005-smb.pdf>
- Oravecz Csaba – Váradi T. – Sass B. (2014): *The Hungarian Gigaword Corpus*. *Proceedings of LREC 2014*. • [http://www.lrec-conf.org/proceedings/lrec2014/pdf/681\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf)
- Pataki Máté – Micsik A. – Kovács L. – Szabó M. (2014): *KOPI-Fotó: Plágiumkeresés egy lefotózott oldal alapján*. Informatika a felsőoktatásban konferencia, Debrecen, 2014. augusztus 27–29. • <http://eprints.sztaki.hu/8019/>
- Sándor Ágnes – Vorndan, Angela (2010): *The Detection of Salient Messages from Social Science Research Papers and Its Application in Document Search*. Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires, Argentina.
- Váradi Tamás – Mittelholcz I. – Blága Sz. – Harmati S. (2014): *Magyar társadalomtudományi citációs adatbázis: A MATRICA projekt eredményei*. MSZNY 2014. Magyar Számítógépes Nyelvészeti Konferencia, Szeged. JATEPress, Szeged, 269–279.
- URL1: The Hague Declaration, <http://thehague.declaration.com/>
- URL2: Casey Bergman: *Why Are There So Few Efforts to Text Mine the Open Access Subset of PubMed Central?* <https://caseybergman.wordpress.com/2012/03/02/why-are-there-so-few-efforts-to-text-mine-the-open-access-subset-of-pubmed-central/>
- URL3: <http://nanopub.org/wordpress/>

