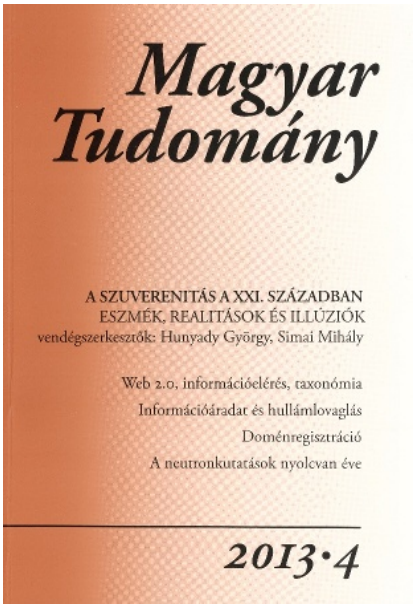

Magyar Tudomány, 2013/4. 473.o.

Információáradat és hullámlovaglás

Holl András
MTA CsFK CsI, MTA Könyvtára
holl.andras@konyvtar.mta.hu

Bevezetés

A tudományos szakirodalom exponenciális növekedését Derek de Solla Price ismerte fel a 20. század közepén, a *Philosophical Transactions of the Royal Society* majd két évszázad során publikált köteteit időrendben feltornyozva (lásd Ekers, 2009). Michael J. Kurtz és Edwin A. Henneken (2012) évi 3,5%-os növekedést említ: az évente publikált cikkek mennyisége húsz év alatt megduplázódik. Philip Young (2009) szerint a növekedés üteme évi 3%. A kutatói elme kapacitása feltehetően nem növekszik, ám a kutatók száma igen. Ennek következménye az egyre erőteljesebb specializálódás. A növekvő ismeretanyag kezelése – ha nem fogadjuk el a növekvő elszigetelődést, pontosabban a csupán kevéssé kapcsolódó szakterületek kialakulását, ha nem adjuk fel a „teljes kép” áttekintésének igényét – egyre nagyobb kihívást jelent mind a kutatóknak, mind a kutatást irányítóknak.

A nehézségek már régóta nyilvánvalóak a tudományos szakkönyvtárak számára. A tudományos folyóiratok előfizetési díjai gyorsabb ütemben növekszenek, mint amekkorát az infláció indokolna (Peine, 2010) – tegyük hozzá, az infláció és a terjedelemnövekedés együttese által indokoltnál is[1]. A tudományos folyóiratok kiadását nem szabályozza hatékonyan a piac: mivel az egyik folyóirat nem helyettesítheti a másikat, minden cikk egyedi (Young, 2009). Nem a költségvetés jelenti a könyvtárak egyetlen gondját: azokat a folyóiratokat, amelyeket meg tudnak rendelni, egyre nehezebb a polcokon elhelyezni.

A tudományos adatok növekedése is exponenciális, de a növekedési ütem a szakirodalom növekedésénél sokkal gyorsabb. Ennek oka, hogy a kutatók számának növekedésénél sokkal gyorsabban emelkedik az adatokat gyűjtő műszerek és számítógépek száma, ráadásul az adatgyűjtő képességük is rohamosan nő[2].

1 A 2007-es árnövekedés 7% Young (2009) szerint, Mike Peine (2010) az 1989 és 2011 közötti árnövekedésre ugyanekkora adatot közöl az USA-ban kiadott folyóiratokra.
2 A mindennapi életben is megfigyelhető jelenség a CCD-csipek növekedése a fényképezőgépekben és a telefonokban, ami együtt jár a képek méretének növekedésével.

Az adatokra is igaz, amit a szakirodalomra nézve leírtunk: a tárolás, feldolgozás megoldható a felhasznált számítógépek számának növelésével – a sok gépre, sok merevlemezre elosztva gyorsabb ütemben növekedhet a tárolt, feldolgozott adatok mennyisége, mint az egyes számítástechnikai alkatrészek kapacitása (bár az önmagában is gyorsan növekszik, mint azt például a Moore-törvény kimondja). De felmerül az igény a roppant nagy adatállományok összevetésére, az összes, adott kérdésben rendelkezésre álló adatokban való kutatásra – és itt már a technológiai fejlődés, a több számítógép önmagában nem segít. Az adatáradatról és kezelésének kihívásairól Szalay Sándor és Jim Gray (2006) cikkében olvashatunk, az adatmennyiség szerintük évente megduplázódik. Az adatmennyiség növekedését szemlélteti a csillagászati égboltfelmérések története: az 1950-es évek végére elkészült National Geographic Society – Palomar Observatory Sky Survey fotólemezeit az 1990-es évek közepén digitalizálták, és százkét CD-ROM-on adták ki (mind maga az égboltfelmérés, mind a digitalizálása több évig tartott). A Large Synoptic Survey Telescope egy évtized múlva több mint egy petabájt adatot fog szolgáltatni évente (ami nagyjából kétmillió CD-ROM-on férne el).

Mindez egyben komoly lehetőségeket is kínál. Egy, az Európai Bizottság számára készült jelentés (*A hullámot meglovagolva* [*Riding the Wave*, HLEG, 2010]) megvizsgálja, hogyan használhatja ki az Európai Unió a tudományos adatok özöne által teremtett esélyeket.

Cikkünkben megvizsgáljuk az információáradat kezelésének két aspektusát: a tudományos információk szabadon hozzáférhetővé tételét, valamint a kereshetőség megteremtését mind a kutatók, mind az érdekükben eljáró elektronikus programok számára.

Az információk szabadon hozzáférhetővé tétele

Mind a tudományos szakirodalom, mind az adatok esetében fontos szempont a jelenlegi finanszírozási rendszer átalakítása. A kutatási eredmények – publikációk és adatok – közreadásának, megőrzésének költségeit a kutatási projektek finanszírozásába kell beépíteni. A publikációk esetében nem szerencsés és nem fenntartható a jelenlegi gyakorlat, miszerint ugyan alkalmasint a projektekben is szerepelnek publikációs költségelemek (például a gyakran fizetendő közlési hozzájárulás (page charge), a színes ábrák díja), de a költségek nagyobb része a könyvtáraknál jelentkezik. A kutatók és a könyvtárak közötti gyenge csatolás (Young, 2009) hozzájárul a költségek ellenőrizhetetlen növekedéséhez. Egy szűk tudományterületen – a nagyenergiás fizikában – meg is indult az átalakulás, elindult a SCOAP[3] projekt[3] (Mele, 2010).

A tudományos adatok esetében az adatok tisztításának, dokumentálásának, metaadatokkal való ellátásának, megfelelő formátumra alakításának költségei követelik meg az archiválásuknak és hozzáférhetővé tételüknek a kutatási projektek keretében való kezelését. Utólagosan összegyűjteni, hosszú távú megőrzésre, publikus felhasználásra alkalmassá tenni az adatokat reménytelen és megfizethetetlen lenne[4]. Ha a projektet támogatók (kutatási alapok) nem ellenőrzik, valószínűleg elmarad az archiválás és közreadás. S ha elmarad, a kutatás utólagos ellenőrizhetőségének lehetősége csökken, és nem nyílik mód az adatok másodfelhasználására. Mind több nagy tudományos projekt választja az adatok nyílt

3 Sponsoring Consortium for Open Access Publishing in Particle Physics (URL1)

4 Az adat-infrastruktúráknak, a hosszú távú megőrzésnek és elérhetővé tételnek mindenképpen lesznek a projektekre nem hárítható költségei, azonban proaktív megközelítés nélkül ezek sokkal magasabbak lesznek.

hozzáférhetőségét (esetleges embargó időszak után), elég a Human Genome Projektet vagy a Hubble Űrtávcsövet említeni.

A tudományos szakirodalomhoz való nyílt hozzáférés (Open Access – OA) igényét 2001 végén Budapesten fogalmazták meg (Budapest Open Access Initiative). Az OA célja – a már említett költség- és tudományirányítási szempontokon túl – a publikációk hatásának (impaktjának) növelése és az adatbányászat lehetőségének megteremtése (erre a következőkben még visszatérünk). Az OA megteremtésére több út is kínálkozik: az „arany út": eleve OA-folyóiratban közölni; és a „zöld út": repozitóriumban elhelyezni. (A repozitóriumok teljes szövegű publikációk szakszerű elhelyezésére szolgáló elektronikus könyvtári rendszerek.) Az Open Access egyes kérdéseiről a *Magyar Tudomány*ban már írtunk (Holl, 2010). Újabb hazai fejlemény az MTA elnökének OA határozata (URL2).

Az OA közzétételi kötelezettség egyre szélesebb körben jelenik meg. Az EU 7-es keretprogramjában már kísérleti jelleggel megjelent (URL3), a *Horizont 2020* már kötelezővé fogja tenni a nyílt hozzáférést, és a publikációkon kívül már a tudományos adatokkal is foglalkozik (URL4). A publikációkra koncentráló DRIVER- és OpenAIRE programok után az OpenAIREplus-ban már a tudományos adatok problémája is megjelenik (URL5). Az Alliance for Permanent Access/PARSE.Insight (URL6) jó tájékozódási lehetőséget kínál az adatok megőrzésével és hozzáférhetővé tételével foglalkozó projektek között.

Az adatok hozzáférhetővé tételének fontos eleme a rájuk való hivatkozás lehetőségének megteremtése. A kutatások támogatói érdekeltek a ráfordításaik minél nagyobb mértékű megtérülésében: az egyszer megszerzett adatokból minden csepp tudományos haszon kipréselésében, a másodlagos hasznosításban. A kiadók feladata őrködni a minőség felett – a *Nature* például megköveteli a cikkekhez használt adatok elérhetővé tételét[5]. A kutatók érdekeltségét viszont az adatok idézhetőségének megteremtésével lehet biztosítani. Az idézhetőséghez állandó és feloldható azonosítókra van szükség, a publikációkhoz hasonlóan. A DataCite (URL8) DOI-azonosítókat biztosít adatállományokhoz. Ez már a következő témánkhoz vezet el bennünket: az információk kereshetővé tételéhez.

A publikációk és az adatok kereshetővé tétele

A szabad elérhetőség követelménye nem csupán annyit jelent, hogy a weben elérhetővé kell tenni a cikkeket vagy az adatokat. Gondoskodni kell arról, hogy az információk hosszú távon elérhetőek és felhasználhatóak maradjanak. Ezért van szükség repozitóriumokra, az információkat gondozó könyvtárosokra és adatkönyvtárosokra, valamint egyedi, feloldható azonosítókra[6]. Míg az adatoknál a már említett DataCite kínál azonosításra megoldást, a publikációknál a CrossRef szervezet biztosítja a dokumentumok DOI-val való címkézését.

Az elérhetőség, kereshetőség lehetőségét a tudományos szakirodalom esetében üzleti alapon működő adatbázisok teremtik meg jelenleg: a *Web of Science* (WoS) és a *Scopus*. A

5 "...authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications." [a szerzők kötelesek az anyagokat, adatokat és hozzájuk tartozó eljárásokat késedelem, és indokolatlan megkötések nélkül az olvasók számára elérhetővé tenni] (URL7)

6 Az egyedi azonosító létrejöttétől kíséri az adatállományt vagy dokumentumot, ilyen például az ISBN a könyveknél, és ilyen a Digital Object Identifier (DOI). A feloldhatóság azt jelenti, hogy van egy központi adatbázis, amelyből az azonosító alapján meg lehet tudni a leíró adatokat, és el lehet jutni magához a dokumentumhoz (adatállományhoz).

publikációk adatainak adatbázisba gyűjtése vagy költséges apparátus kiépítését követeli meg, vagy nagyobb együttműködést igényel a kiadók részéről. Az utóbbira példa egy tudományterületi bibliográfiai adatbázis, a csillagászatot lefedő *Harvard-Smithsonian Astrophysics Data System*, amely ingyen kínálja az információkat. Amennyiben nagyobb mértékben támaszkodnának a kiadók a már létező, szabványos bibliográfiai metaadatokat közreadó protokollra, az OAI-PMH-ra, sokkal kisebb költséggel lehetne publikációs adatbázisokat üzemeltetni.

Magyarországon a *Magyar Tudományos Művek Tára* (MTMT) gyűjti össze a tudományos publikációkat, és amellett, hogy statisztikai adatokat szolgáltat, portált is biztosít majd a hazai tudományos eredményekhez. Az egységes keresés lehetőségeit az MTMT teremti meg, és utat nyit a szabadon elérhető teljes szövegek felé – már amennyiben ilyen rendelkezésre áll a kiadóknál vagy a repozitóriumokban. A WoS vagy éppen a *Scopus* nem helyettesítheti az MTMT-t – már csupán azért sem, mert a humán- és társadalomtudományokat, a magyar nyelven publikált cikkeket ezek nem reprezentálják. A természet-, élet- és műszaki tudományokban a mérce nemzetközi[7], ám a nemzeti nyelvet, a hazai kiadású, esetleg (még) impaktfaktor nélküli folyóiratokat semmibe venni mégsem szerencsés. A szerb gyakorlat (a doiSerbia által közvetített azonosítók és a SCIndeks által mért impakt) 2005 utáni hatása a tudomány eredményességére szembeszökő (Šipka, 2012, az előadás 9. diája). Nemzeti tudományos bibliográfiai adatbázisokra példa az SCIndeks-en túl a holland NARCIS vagy a malajziai MyCite (URL9).

A cím, szerző vagy megjelenési adatok alapján való kereshetőség nélkülözhetetlen, de az exponenciálisan bővülő szakirodalom áttekintésére egyre kevésbé alkalmas. Szerencsére az informatika és a bibliográfiai adatbázisok új tájékozódási lehetőségeket is teremtenek. A kutatók többsége valószínűleg találkozott már a weben vásárolva, termékinformációkat keresve olyan funkciókkal, amelyek a korábbi látogatók vásárlási és böngészési szokásai alapján kínáltak segítséget: mit vettek (néztek) még azok, akik ugyanezeket a termékeket tették a kosarukba (vagy nézték meg)? A tudományos publikációk esetében nemcsak a letöltési adatok, de az idézési háló is információforrást jelent. Az Astrophysics Data System-ben már megjelennek az idézési és idézettségi mintázatokat felhasználó cikkajánló funkciók (Kurtz – Henneken, 2012; Kurtz et al., 2002). Segítségükkel pillanatok alatt meg lehet keresni egy, a felhasználó számára új tudományterületen a legfontosabb referenciamunkákat és a legújabb áttekintő cikkeket.

A bibliometriai adatok és az idézettségi hálózat elemzése segíthet a kutatóhelyek értékelésében is. Nem csupán rangsorok kialakításához járulhat hozzá, de a kutatási struktúra, az erősségek és hiányosságok, a kutatóhelyek közötti kapcsolatok is vizsgálhatók. A Thomson Reuters InCites, illetve az Elsevier SciVal Experts és SciVal Spotlight a WoS, illetve a Scopus adataira épülő elemzést kínál. Az MTMT adatai – a folyamatban lévő TÁMOP[8] projekt keretében folyó adatbázis-feltöltés befejezése és az adatminőség javítása után – felhasználható lesz ilyesféle tudományelemzési célokra is[9].

7 A humán- és társadalomtudományokban egy magyar nyelvű, hazai vonatkozású cikk idézettsége nehezen mérhető össze egy, mondjuk műszaki tárgyú, angol nyelvű, világszerte kutatott témában megjelent cikkével.

8 TÁMOP-4.2.5.A-11/1-2012-0001 A Magyar Tudományos Művek Tára (MTMT) publikációs adatbázis szolgáltatások országos kiterjesztése.

9 Ezen már dolgozik a TÁMOP-projekt keretében az MTA Könyvtárának Tudománypolitikai és Tudományelemzési Osztálya.

Kecsegtető a teljes szövegekben való keresés lehetősége – ehhez viszont alighanem nélkülözhetetlen a publikációkhoz való nyílt hozzáférés. Nem elég azonban az a lehetőség, hogy karaktersorozatokra kereshessünk, személyekre (szerzőkre), objektumokra (vegyületekre, fajokra, égitestekre) is kell tudni keresni. A szerzők azonosítására nyújt lehetőséget az ORCID (URL10). Egy hazai folyóiratban, az MTA CsFK CsI-ben (Csillagászati és Földtudományi Kutatóközpont Konkoly Thege Miklós Csillagászati Intézete) kiadott *Information Bulletin on Variable Stars*ban való szemantikus keresést ismertet Holl András (2012). A szemantikus web irányába mutató kísérletet jelentenek a nanopublikációk (Mons et al., 2011). Ez a technológia többet kínál a szövegbányászatnál, „hiszen eleve minek eltemetni [az adatokat], ha végül úgyis ki akarjuk bányászni?" (Mons, 2005). A lényegállítások, mint például: „a malária terjesztője az Anopheles szúnyog", „a HLA-B27 gén kapcsolatban áll a Bechterew-kórral" vagy éppen a „GSC 2936-297 csillag delta Scuti típusú változó" leírhatók számítógépek által értelmezhető módon, alanyból, állítmányból és tárgyból álló tripletek segítségével. Barend Mons szerint a nanopublikációk felváltják majd a publikációkat. Véleményünk szerint talán inkább a cikkek legfontosabb állításait kódolják majd a géppel olvasható tartalmi kivonatokban.

Az adatok kereshetővé és újrafelhasználhatóvá tételében jelentős sikereket ért már el a csillagászat Virtuális Obszervatórium projektje. Ellenőrzött szótárakat, az adatokat gazdagon dokumentáló, szabványos formátumokat, adatforrásokat tartalmazó jegyzékeket, keresési protokollokat, megjelenítőprogramokat hoztak létre az adatözön kezelésére (Berriman et al., 2012).

A különböző tudományágakban nehéz lenne ugyanazokat a módszereket és eszközöket alkalmazni a publikus adatok kezelésére. Az a követelmény közös, hogy a kutatási projektekben a keletkező adatok közzétételének és ennek finanszírozásának meg kell jelennie, és közös annak a szükségszerűsége, hogy az információözönt szabályozzuk, sőt meg tudjuk lovagolni a hullámait.

A meteorológiai előrejelzés úgy született meg, hogy az 1854-es balaklavai csatában egy szélvihar súlyos veszteségeket okozott az angol–francia hajóhadnak. A francia kormány megbízta Urbain Jean-Joseph Le Verrier csillagászt, vizsgálja meg, hogy az akkor létező meteorológiai állomások adatait használva előre jelezhető lett volna-e a vihar? Vajon hány tudományos felismerést szalasztunk el azért, mert az amúgy létező adatok nem érhetőek el, és nem illeszthetőek össze, hogy teljes képet alkossanak?

Kulcsszavak: Open Access – nyílt hozzáférés, bibliográfiai adatbázisok, repozitóriumok, tudományos adatok

IRODALOM

Berriman, G. Bruce et al. (2012): *The Role in the Virtual Astronomical Observatory in the Era of Massive Data Sets*. SPIE Conference 8448: Observatory Operations: Strategies, Processes, and Systems IV, arXiv:1206.4076 [astro-ph.IM]

Ekers, Ronald D. (2009): *Big and Small. Accelerating the Rate of Astronomical Discovery.* IAU GA, Rio de Janeiro, Brazil, arXiv:1004.4279 [astro-ph.IM]

HLEG (2010): *Riding the Wave. Final report of the High Level Expert Group on Scientific Data*. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Holl András (2010): Nyílt hozzáférés a tudományos szakirodalomhoz – hazai fejlemények. *Magyar Tudomány*. 1, 58–61. http://www.matud.iif.hu/2010/01/11.htm

Holl András (2012): Information Bulletin on Variable Stars—Rich Content and Novel Services for an Enhanced Publication. *D-Lib Magazine*. 18, 5/6, doi:10.1045/may2012-holl http://www.dlib.org/dlib/may12/holl/05holl.html

Kurtz, Michael J. – Eichhorn, Günther et al. (2002): Second Order Bibliometric Operators in the Astrophysics Data System. Proc. SPIE 4847, Astronomical Data Analysis II, 238, doi:10.1117/12.460438

Kurtz, Michael J. – Henneken, Edwin A. (2012): *Finding and Recommending Scholarly Articles*. arXiv:1209.1318 [cs.IR]

Mele, Salvatore (2010): *Open Access Publishing in High-Energy Physics: the SCOAP^3 Initiative*. (*ASP Conference Series* 433) 156–166. http://articles.adsabs.harvard.edu/full/2010ASPC..433..156M

Mons, Barend (2005): Which Gene Did You Mean? *BMC Bioinformatics*. 6, 142, doi:10.1186/1471-2105-6-142 http://www.biomedcentral.com/1471-2105/6/142

Mons, Barend et al. (2011): The Value of Data. *Nature Genetics*. 43, 281, doi:10.1038/ng0411-281 http://www.nature.com/ng/journal/v43/n4/full/ng0411-281.html

Peine, Mike (2010): 2010 Study of Subscription Prices for Scholarly Society Journals. http://allenpress.com/system/files/pdfs/library/ap_journal_pricing_study_2010.pdf

Šipka, Pero (2012): Bibliometric Quality of Serbian Journals 2002–2011: More than Just a Dress for Success. Fifth Belgrade International Open Access Conference, http://boac.ceon.rs/public/site/Sipka.pdf

Szalay, Alexander – Gray, Jim (2006): 2020 Computing: Science in an Exponential World. *Nature*. 440, 413–414. doi:10.1038/440413a

Young, Philip (2009): *The Serials Crisis and Open Access*. http://scholar.lib.vt.edu/faculty_archives/YoungP/OAwhitepaper.pdf

URL1: SCOAP^3 Project http://scoap3.org/index.html

URL2: MTA elnöki OA-határozat http://mta.hu/data/cikk/11/97/91/cikk_119791/27_2012_elnoki_hat_Open_Access.pdf

URL3: EU-7 OA közzétételi kötelezettség http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1300

URL4: *Horizont 2020* http://europa.eu/rapid/press-release_IP-12-790_hu.htm

URL5: OpenAIRE: http://www.openaire.eu/hu/component/content/article/326-openaireplus-press-release

URL6: Alliance for Permanent Access – APA/ PARSE.Insight: http://www.parse-insight.eu/

URL7: Nature http://www.nature.com/authors/policies/availability.html

URL8: DataCite: http://datacite.org/

URL9: SCIndeks: http://scindeks.ceon.rs/Default.aspx?lang=en

URL10: ORCID – Open Researcher and Contributor ID http://about.orcid.org/