Z. Karvalics László

Data Hugging – a magunkhoz szorított adatokon innen és túl

Az adatok világában földrengés-szerű változások mennek végbe, és erre a technológia és az üzlet szereplői után és mellett érzékenyen, alaposan és időben reagált a tudomány is[1] Pedig 2011 médiaszenzációja, a közbeszédet és az üzleti kommunikációt beterítő, már-már parttalan „Big Data diskurzus" elkezdett kifáradni: egyre kevesebb tanulmány, cikk és könyv boncolgatta annak hatáskövetkezményeit, hogy milyen technológiai kihívásokat okoz a közösségi oldalak, a nagy ügyfélforgalmú vállalatok, az egészségügyi intézmények vagy a tudományos óriásgépek elképesztő méretűvé duzzadt adattermelése.[2] Ehelyett az elemzés, a feldolgozás, a megosztás és a vizuális megjelenítés kérdései felé fordul a figyelem. Mégis, csak kapkodja a fejét az, aki 2009 előtt szerezte az informatikai vagy üzleti diplomáját. Számtalan új kifejezéssel kell ugyanis rövid idő alatt megismerkednie: olyanokkal, mint a „data hugging" (vagy a DbHd, database hugging disorder) meg a „linked data", nem is szólva az adatvilág nevezéktanának a fentiekkel összefüggésben újraértelmezendő, átalakuló többi darabjáról – amelyekről az alábbi, rövid áttekintésben adunk egyfajta gyorsfényképet.

Sokan a „statisztika világának Mick Jaggereként" és adatcelebként (*data star*) emlegetett *Hans Rosling*hoz kötik a „*data hugging*" kifejezés megalkotását, pedig serény marketingvarázslók már évekkel a kiváló svéd tudósnak a tárgyban elhíresült előadásai[3] előtt evvel a kifejezéssel igyekeztek rossz érzést kelteni azokban, akik nem voltak elég bátrak saját szervereikről a számítási felhőbe áthelyezni az adatbázisaikat. Mint egy véget nem érő, ragacsos ölelés: szorítják és nem eresztik.

Rosling értelmezése aztán felülírta ezt a jelentést. Ő abból indult ki, hogy a különböző adatgazdák által őrzött adatsokaságok nagy szabadságfokkal használhatók arra, hogy a bennük tükröződő tényekből, illetve azok párosításaiból a világ új és új összefüggéseire csodálkozzunk rá. Ha kellő mennyiségben, variációban és minőségbiztosítottan állnak rendelkezésre az adatok, mindig tehetünk fel olyan kérdéseket, amelyekre választ kapunk tőlük, ha tudjuk a módját, hogyan csalogassuk elő vagy szemléltessük vizuálisan az eredményt. Egy kicsit megváltozhat az a mód, ahogyan arról gondolkodunk, hogy hogyan szüntethetünk meg ismerethiányos helyzeteket.

Ha azonban az adatgazdák túlzottan ragaszkodnak a saját adataikhoz, inkább dajkálgatják, mintsem hogy nagy, közös felhasználási terekbe engednék őket, könnyű hozzáférést biztosítva, akkor a nyílt adattér még nem elég vonzó ahhoz, hogy sok szereplő itt keresse a megoldást saját problémájára (mint bármilyen új terméknél vagy szolgáltatásnál, ahol a keresletnek meg a kínálatnak egyszerre kell átlépnie a kritikus küszöböket).

Nagyon szeretik magukhoz ölelni az adataikat az államok (*data-hugging states*). Az amerikaiak (*www.data.gov*) és a britek (*www.data.gov.uk*) szerencsére példát mutatnak a világnak avval, hogy minél több, a közszféra segítségével létrejött adatot publikálnak automatikusan: annak tudatában, hogy valahol, valakiknek éppen az alkoholmérgezésben elhunytak számára van szüksége havi bontásban, más pedig a forgalomból kivont járművekkel kapcsolatban lesz kíváncsi valamire. De még kényelmesek és óvatosak a közintézmények, lassúak és gyanakvóak a cégek. S noha ezt ma már szokás rendellenességnek (*disorder*) hívni, azért nagyon erősen szorítanak a karok. Az adatöleléssel kapcsolatos konferenciákon a legnépszerűbb előadások azokat a tipikus érveket és kifogásokat gyűjtik össze és pellengérezik ki,[4] amelyeket a „miért jobb, ha nem osztjuk meg az adatainkat" kérdésre szokás adni válaszul.

Pedig Rosling mellett nem kisebb név, mint a World Wide Web megalkotója, *Tim Berners-Lee*

kezdte el néhány éve népszerűsíteni a gondolatot, hogy az adatvilág egyes darabjai a hozzáférhetőségen túl strukturált módon kölcsönösen összekapcsolhatóvá és kereshetővé is tehetők. Ez az összekapcsolt adat, a *linked data* világa − amely nem azonos a technológiai alaprendszerek szintjén használt adatkötéssel (*data binding*).

A korábbi időszakok hatalmasra duzzadt intézményi (kormányzati, vállalati, kutatóintézeti stb.) kezelésű adatvilágai mellé (valójában: fölé) egy új rendszerszint jön létre, megállíthatatlanul. Egy globális adattér (*global data space*),[5] egy új adatökoszisztéma (*data ecosystem*), amely legalább akkora előrelépést hozhat, mint a hypertextelv megvalósulása, a WWW világa alig húsz éve a szövegek birodalmában. Tegyük hozzá: az adatbázis-hálózatok, a kutatói citációs indexek és a könyvtári katalógusrendszerek már megkezdték a beköltözést ebbe a térbe, mert „házon belül" elvégeztek nagyszámú, fontos összekapcsolást. Most eljött az idő, hogy „egymás között" is relációk létesüljenek, hogy a publikációk mellett az adatbázisok közzététele is a tudományos gépezetek működésének mindennapi rutinjává váljon, s hogy az így megnyitott térbe egyre több adatgazda kapcsolja be a maga állományát. Ez egyik oldalról az adatokat strukturálatlanul, ömlesztve kezelő irtózatos silók (*data silos*) megszüntetését, másrészt az adatközi tér folyamatos újrarendezését eredményezi, ami kihat a „privát" kezelésben maradó adatbázisokra is. Adatcsomópontok (*data hubs*) emelkednek ki, amelyekhez különösen sok link vezet. Adatgyűjtemények (*data compendiums*) jönnek létre, amikor meghatározott célból, projektszerűen kapcsolnak össze adatcsoportokat. Az adatvagyon (*data asset*) immár az apró tudásműveletek millióival összeabroncsolt adattér teljes állományát jelenti, részben az emberiség egyfajta közjószágaként (*common good*), részben az adattulajdonosok mind értékesebb „tőkeformájaként". Adattőke (*data equity*)[6] ugyanis akkor keletkezik, amikor a rendelkezésre álló adatvagyon nagy teljesítményű analitikai megoldásokkal gyors és sikeres döntéseken keresztül aktualizálódó üzleti előnyszerzésbe fordul. Nem véletlen, hogy magas szinten igazgatni, alakítani, irányítani, kormányozni kell majd a folyamatokat, a vállalaton belül kialakított szervezeti kultúrát és a nemzetközi szervezetek adatgazdálkodási gyakorlatát átfogóbb szintre emelve (*global data governance*).[7]

A vállalati−intézményi szinten a továbbiakban adatkészletek (*data sets*) kezelése folyik majd, a jól bevált adattárházas−adatbányászós módon. Az új adattérben új értelmet nyer az adat-deduplikáció (*data deduplication*) is, az ésszerű leegyszerűsítés és „kompresszálás", hiszen az őrzési helyek, másolatok dzsungelében minden átvitelnek és minden tárolási műveletnek jelentősége van – a technológiának követnie kell tudni a megnövekedett igényeket támogató tranzakciótípusokat. Ezért válik az elérhetőség (*accessibility*) mellett egyre inkább meghatározóvá a mozgathatóság, áthelyezhetőség (*data portability*) is. A megnövekedett adattömeg kezelésére szakosított adatközpontok (*data centers*) legfrissebb generációjával új típusú „megatechnológia" jön létre, akár az elfoglalt alapterület, az üzemi méretek, az egy négyzetméterre eső gépi komponensek száma, az automatizált folyamatok száma vagy az energetikai megoldások (áramellátás, hűtés) volumene és innovativitása szempontjából tesszük mérlegre ezeket a gigászi komplexumokat[8].

Az is jól látható, hogy az adatok fizikai tárolása és védelme mellett a megbízhatóság, hitelesség és standardizáltság válik különösen fontossá a makroszinten, így a rendszerben felértékelődik az adatforrás (*data source*), az összeilleszthetőség biztosítása és az ezt akadályozó mozzanatok (*data inconsistency*) megszüntetése. Ezért sokat tehet egy éppen formálódó, új terület, az adatújságírás (*data journalism*) is.[9]

A szervezett összekapcsolás révén újra életre kelhetnek olyan adattömegek is, amelyekről korábban azt hihették, hogy már nem lesznek jók semmire: az eddig csak „házon belül" értelmezett és alkalmazott másodlagos, ismételt vagy többszörös felhasználás (*data re-use*) révén számos állomány kapcsolódhat az adatkörforgásba. Amikor pedig marad idő és energia, hogy a látszatra már semmilyen hasznot hajtani nem tudó adatszeméthez (*data trash*) is kapcsolatok épüljenek, a szakirodalomban eddig óvatosan pedzegetett adatújrahasznosítás (*data recycling*) is értelmet nyerhet. Ez persze nem lesz ok arra, hogy akik eddig az adatszmoggal (*data smog*) riogattak minket,[10] most igazolva lássák félelmeiket. Az új adattér ugyanis akkor működik jól, ha a megőrzésre és összekapcsolásra érdemtelen adatszemét valóban adatpocsolyába, adat-meddőhányóba (*data puddle*) kerül. A nagyvállalatoknál már korábban is használták az adattemető (*data cemetery*) kifejezést a döntéshozatalt nem támogató, amiatt felesleges és eldobandó adatokra. (Egy német vállalati kontroller 2004-ben adattemető-gondnoknak (*Verwalter des*

Datenfriedhofes)[11] nevezte azt, aki a szűrő szerepet ellátja.) Az új adattérben mindez egy új szakma képviselőjére, az adathulladék-kezelőre (*waste data handler*)[12] vár, akire talán már korábban is szükség lesz, mint 2030, amikorra alkalmazásba állását jósolják. Személyiségünk és magánéletünk megvédése érdekében a virtuális térben elszórt nyomaink biztonságos eltüntetése[13] ugyanis sokkal fontosabbá válik, mint az értéktelen vállalati, állami vagy tudományos adatszemét kiválogatása és „elhamvasztása". Persze neki is vigyáznia kell, nehogy valamit úgy töröljön, hogy majd egy adatrégésznek (*data archeologist*) kelljen helyreállítani az adatokat.

Valaki elmosolyodott, hogy ez talán már mégis túlzás? Nos, nem az. Adatrégészek 1993 óta léteznek, és éppen az a feladatuk, hogy a különböző hordozókról újra hozzáférhetővé tegyenek adatokat.

A megnövekedett adattömeg az egyes szervezeteken belül is erős szakosodást indított már el, hiszen újfajta adattermékek (*data products*) egész sora jelent meg.[14] Kezd kettéválni az adatgazdász (*data steward*) és az adatgondozó/kezelő szerep (*data custodian*). Előbbi a tartalmi, használati és üzleti szempontokra érzékeny, utóbbi adja a technológiai hátországot. Az adatspecialisták új generációja pedig egyenesen a legpiacképesebb és legkeresettebb szakemberek közé tartozik a munkaerőpiacon. A sajátos igényeknek megfelelően ma már nem egyszerűen adatkonzultánsokat (*data consultants*), adatkönyvtárosokat (*data librarians*), adatmenedzsereket (*data officers),* adatkurátorokat *(data curator)*[15] adatannotátorokat *(data annotators)*[16] sőt adattudósokat (*data scientists*)[17] keresnek álláshirdetés útján, hanem ezek a jól körülírható ismeretek azonnal gyűjtőkategóriává lettek. Mindegyikből többféle van, attól függően, hogy milyen tudomány (genetika, geográfia, csillagászat) adattömegével birkóznak, illetve. milyen üzleti (közösségi hálózatok, autóipar, légi közlekedés) vagy kormányzati (egészségügyi, térképészeti, nemzetbiztonsági) környezetben van rájuk szükség. E cikk kézirata lezárásának pillanatában csak adatkönyvtárosból tizennégy(!) különböző felhasználói világba keresnek online oldalakon szakértőt – nem véletlen, hogy a felsőoktatás leginnovatívabb amerikai szereplői egymással versengve kezdték meghirdetni az új adattérhez igazodó kurzusaikat. (Európában a skóciai Dundee egyeteme az úttörő.)

Szükség is van a gyors reagálásra, hiszen – ahogy néhány éve bekerült az internetes szállóigék közé – az adatok sosem alszanak (*data never sleeps*). És előttünk áll még az ipari internet (*Industrial Internet*) forradalma (amely a gyártás-előkészítéstől a megvásárolt termék használata vagy szervizelése során képződő adatokig valamennyi termelési komponens, eszköz, ember, folyamat, illetve a változatos szenzorhadsereggel felszerelt összes kibocsátott termék új nagyságrendű adat-ökoszisztémáját hozza létre), valamint a „minden dolgok internetje" (*Internet of Everything*), amely az emberek, tárgyak, folyamatok és adatok új nagyságrendű kapcsolatvilágát teremtheti meg.[18] Nem véletlen, hogy a humán tudományok is új szemmel kezdenek az adatok és az adatbázis problémakörével foglalkozni – szinte bizonyos, hogy az ezredforduló után elindult izgalmas diskurzus a következő években robbanásszerűen fejlődik majd.[19]

Jegyzetek

1 Elsősorban a statisztikusok jeleskedtek, de több áttekintés vizsgálja a pedagógiai hatáskövetkezményeket is (pl. Holcomb, 2011) vagy a közgazdaságtudományi kihívást (Utóbbira magyarul ld. Bőgel, 2011).

2 Ennek az irodalomnak talán legizgalmasabb darabja (Smolan és Erwitt, 2012) egyben a legfrissebb is (2012 novemberének végén jelent meg).

3 Ld. pl. ezt a prezentációt *http://unstats.un.org/unsd/statcom/statcom_2010/Seminars/Communication/default.html* vagy az alábbi interjút (*http://news.bbc.co.uk/today/hi/today/newsid_8076000/8076488.stm*).

4 *http://learningtheworld.eu/2011/excuses-for-data-hugging/*

5 Heath és Bizer (2011)

6 A kérdéskör legjobb elemzését ld. Mohamed and Ismail, 2012. A fogalom maga 2011-ben született, a The Economist szakírója, Adrian Wooldridge használta először egy 2011 májusi cikkében.

7 Ladley, 2012.

8 Érdeklődők vessenek néhány pillantást a délbelgiumi St. Ghislain mellett felépített Google Data Center-re: *http://googlepolicyeurope.blogspot.hu/2013/02/a-flower-of-computer-history-blooms-in.html*, vagy nézze meg a Wikibon blogon a világ tíz legnagyobb adatközpontját *http://wikibon.org/blog/inside-ten-of-the-worlds-largest-data-centers/*. A kínai Langfang-

ban 2016 során átadásra kerülő adatközpont mérete meg fogja haladni a Pentagonét.

[9] Gray és tsai, 2012 Orr, 2011. A Digital Journalism című újság 2013 júliusáig várja 2015-ben (!) megjelentetni tervezett tematikus számához a tanulmányokat, amelyet teljes egészében az adat-újságírásnak szentelnek majd (*Journalism in an Era of Big Data*). *http://sethlewis.org/call-for-papers-journalism-in-an-era-of-big-data-special-issue/*

[10] Az adatszmog 2004-ben került az Oxford English Dictionary-be. A kifejezést David Shenk (1997) könyve tette ismertté (noha nem ő használta először). Az adatszemét egy szerzőpáros leleménye (Kroker-Weinstein, 1994).

[11] *http://www.controllingportal.hu/M&C_levelek/10_M&C_level_2004._szeptember_8.Azt_mondd_amit_tenyleg_mondani_akarsz*

[12] Gyarmati (2010)

[13] *http://fastfuture.com/wp-content/uploads/2010/01/future_jobs_sheet.pdf*

[14] A friss típusokat rendszerezi: Loukides (2011)

[15] Ld. pl. *http://www.gabormelli.com/RKB/Data_Curator*

[16] Távmunkában és rész-munkaidőben végezhető összefoglaló-adatoló munka, jellemzően a mozgóképi, képi vagy audio tartalmak kivonatoló, szöveges leírását jelenti.

[17] Az adattudomány mibenlétére ld. az alábbi friss összefoglalót (Loukides, 2012).

[18] Részletesebben ld. Z. Karvalics (2013).

[19] Kezdeteire ld. Manovich (2009/2001), hazai recepciójára ld. Dragon Zoltán és Sághy Miklós 2011-es, sok fordulós párbeszédét az Apertúra oldalain *http://magazin.apertura.hu/tag/lev-manovich*.

Irodalom

BŐGEL György: Az adatrobbanás mint közgazdasági jelenség. = Közgazdasági Szemle, 58. köt. 2011. október, p. 877–889.

GRAY, Jonathan − CHAMBERS, Lucy – BOUNEGRU, Liliana: The Data Journalism Handbook. O'Reilly Media, 2012.

GYARMATI Andrea: A testrész-készítőtől a kapcsolatháló-építő munkásig, avagy foglalkozások 2030-ban. = Információs társadalom, 2010/2, p. 84−92. *http://epa.oszk.hu/01900/01963/00033/pdf/infotars_2010_2_084-092.pdf*

HEATH, Tom – BIZER, Christian: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.

HOLCOMB, Edie L.: Data Dynamics: Aligning Teacher Team, School, and District Efforts Solution Tree, 2011.

KROKER, Arthur – WEINSTEIN, Michael A.: Data Trash. The Theory of the Virtual Class. St. Martin's Press, New York, 1994.

LADLEY, John: Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program Morgan Kaufmann, 2012.

LOUKIDES, Mike: The Evolution of Data Products. O'Reilly Media, 2011. (Kindle Edition)

LOUKIDES, Mike: What Is Data Science? O'Reilly Media, 2012. (Kindle Edition)

MANOVICH, Lev: Az adatbázis, mint szimbolikus forma Apertúra 2009 ősz *http://apertura.hu/2009/osz/manovich* A Kiss Julianna által fordított szöveg az alábbi könyv egy részlete: Lev Manovich: *The Language of New Media*. Cambridge: MIT Press, 2001. 194–122.

ORR, James C.: Data Governance For The Executive Senna Publishing, L.L.C., 2011.

SCHENK, David: Data Smog: Surviving the information glut. HarperCollins, 1997.

SHEHAN, Mohamed – OSMAN, Ismail (eds.): Data equity. Unlocking the value of big data
(Cebr) Centre for Economics and Business Research Ltd, London, 2012.
http://www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf

SMOLAN, Rick − ERWITT, Jennifer: The Human Face of Big Data. Against All Odds Productions, 2012.

Z. KARVALICS László: Minden dolgok Internetje (Internet of Everything)
It-Business Online Széljegyzet 2013. április 5.
http://www.itbusiness.hu/Fooldal/publicisztika/Z_Karvalics_Laszlo/minden_dolgok_internetje.html

Beérkezett: 2013. III. 27-én.

Z. Karvalics László
a Szegedi Tudományegyetem
BTK Könyvtár- és Humán
Információtudományi Tanszék
egyetemi docense.
E-mail: *zklmaildelivery@gmail.com*