TMT 2015/5

Előadás, Budapesti Corvinus Egyetem Központi Könyvtár
Publikációmenedzsment műhelysorozat
2015 március 25.

Holl András

Kutatási adatok kezelésének nemzetközi trendjei

A kutatási adatok – angolul *research data* – a tudományos kutatás nyersanyagai, új tudományos eredmények megalapozói. „Létrejöhetnek megfigyelések, kísérletek, szimulációk eredményeképpen, vagy korábban gyűjtött adatok összegyűjtésével, válogatásával, feldolgozásával."[1] „Kutatási adatokat – más információtípusoktól különbözően – eredeti tudományos eredmények létrehozására irányuló elemzés céljából gyűjtenek, figyelnek meg vagy hoznak létre."[2] „A kutatási adatok rögzített tényjellegű anyagok, melyeket a tudományos közösség elfogad és megőriz a kutatási eredmények igazolásához."[3]

Sok esetben a kutatási adatokat konkrét vizsgálatok számára állítják elő egyéni kutatók vagy kutatócsoportok, kutatási projektek keretében. Ekkor hozzáférhetővé tételük az adott kutatás ellenőrizhetőségét, reprodukálhatóságát célozza. Más esetekben a kutatási adatok gyűjtése és felhasználása egymástól elválik: az adatokat egy felmérési (*survey*) jellegű program keretében hozzák létre, nem konkretizált, de előre láthatóan fontos jövőbeli kutatások nyersanyagául, és gyakorta szabadon hozzáférhetővé teszik (pl. *Human Genome Project*[4]). Előfordul az is, hogy az egyedi, konkrét kutatási projektek céljára létrehozott adatokat adatbázisba szervezik, és további kutatások céljára hozzáférhetővé teszik (pl. *Hubble Space Telescope* adatai a *MAST*-ban[5]). A nagy adatbázisok célja a kutatási adatok újrahasznosítása. A HST esetében az adatok másodlagos felhasználásából keletkező tudományos publikációk mennyisége mára meghaladja az eredeti megfigyelők cikkeinek számát.[6]

Tudományos adatarchiválás a digitális korszak előtt is létezett: adattárak, rajztárak, fotótárak, dokumentációtárak, lelettárak voltak, vannak számos intézményben. Ezeket a kvalifikált kutatók látogathatták, a tartalomhoz hozzáférhettek. Az adatokat számos esetben publikálták. A tudományos szakcikkek megszokott elemei a számadatok, táblázatok, grafikonok és fotók. Azonban a nagy mennyiségű adat publikálása költséges, nem mindig lehetséges. *Konkoly Thege Miklós*t megrótták a nagy mennyiségű adat publikálása miatt (Vargha, 2001). Számos folyóirat indított külön kiegészítő sorozatot (*Erganzungschrifte*, *Supplement Series*) nagy mennyiségű adat közlésére. Mára az adatok túlnyomó része digitális formában keletkezik, és a hozzáférhetőség, kezelhetőség, feldolgozhatóság, sok esetben a megőrzés érdekében a régi adatokat digitalizálják.

Az interneten való hozzáférhetőség jelentősen megkönnyíti az újrafelhasználást. Az adatok felhasználását korábban gyakorta engedélyhez kötötték, mára terjed a nyílt hozzáférés alkalmazása.

Az Európai Bizottság ajánlása[7] szerint:

"*Open access policies aim to provide readers with access to peer-reviewed scientific publications and research data free of charge as early as possible in the dissemination process, and enable the use and reuse of scientific research results.*"

"*Open access to scientific research data enhances data quality, reduces the need for duplication of research, speeds up scientific progress and helps to combat scientific fraud.*"

Az MTA TK Kutatási Dokumentációs Központja weblapján[8] a következő érveket olvashatjuk az adatok megosztásának hasznáról:

- *mert elősegíti a tudományos vitát,*
- *mert a cikkeinkhez be lehet linkelni az adatokat,*
- *mert adatfelhasználók és adatlétrehozók közötti új együttműködést tesz lehetővé,*
- *mert elősegíti az átláthatóságot és számonkérhetőséget,*
- *mert hozzájárul a tudományos módszerek fejlődéséhez,*
- *mert csökkenti a tudományos kutatás költségeit,*
- *mert hatásosabbá és láthatóbbá válnak a tudományos eredményeink,*
- *mert növeli a tudós elismertségét,*
- *mert kiváló oktatási anyag lehet belőle.*

A kutatási adatok megosztása a kutatók számára nem mindig könnyű. Itt is megfigyelhető az a kettősség, ami a publikációkhoz való nyílt hozzáférésnél: vannak, akik lelkesen gyakorolják már hosszú idő óta, mások tartanak tőle, ellenállnak. A legfontosabb ellenérv talán az, hogy a kutatási adatok értékesek: pénz- és munkabefektetés árán jönnek létre, és gyakorta további publikációk készítéséhez való információkat rejtenek. Ám tudjuk, hogy a kutatók sokszor „ülnek” az adatokon, másnak nem adják, maguk csak komótosan foglalkoznak vele, vagy egyáltalán nem. Ez esetben a kutatást finanszírozó, a munkáltató avatkozhat be – a kutatásba fektetett összeg jobb hasznosulása érdekében elvárhatja a nyilvánosságra hozatalt, többnyire türelmi idő elteltével. A már említett HST esetében a megfigyelő csoport dolgozhat az adatokkal egy évig – utána nyilvánossá válnak. Ez egyrészt nyomást gyakorol a kutatókra, hogy publikáljanak hamar, másrészt növeli a másodlagos publikációk számát. A módszer bevált.

Nehezebben kezelhető, etikai problémák jelentkeznek az orvostudományban vagy szociológiában – ahol az adatok a kutatások alanyainak személyiségi jogait, érdekeit érinthetik. Ezekre a kutatásokra jelenleg általában szigorú szabályozások vonatkoznak – az összegyűjtött adatokat az elsődleges felhasználás után gyakorta meg kell semmisíteni, további kutatásokra már nem használhatók fel. Jelentős tudományos haszonnal járhatna, ha például a különböző gyógyszerekre vonatkozó klinikai vizsgálatok adatait egy nagy

adatbázisba lehetne tölteni. Egy lehetséges megoldás az anonimizálás – de ez sem mindig tökéletes.

A kutatási adatok mások által való felhasználhatóságát sok munkabefektetéssel lehet megteremteni – megfelelő dokumentációt, metaadatokat kell biztosítani. Ha ezt a munkát a kutatók értékelésénél nem veszik figyelembe, kevés adat lesz nyilvános. A nyilvánossá tett adatok – és a rájuk történő hivatkozások számbavétele megkezdődött, a *Thomson Reuters* például létrehozta a Data Citation Indexet. Nehéz a kutatási eredményeket ellenőrizni, reprodukálni, ha az adatok nem hozzáférhetők, nyilvánosak. Nem szabadna előfordulnia, hogy a kutatók az adataik manipulálásával befolyásolják az eredményeiket. A tudományos csalás eseteinél gyakrabban fordul elő az adatok kozmetikázása. Itt leginkább a tudományos folyóiratok szerkesztőségeinek van lehetősége beavatkozni. Egyre több folyóirat követeli meg a cikkekhez használt kutatási adatok nyilvánosságra hozatalát. Érdemes a *Public Library of Science* gyakorlatát említeni.[9]

Hogyan lehet a kutatási adatokat hozzáférhetővé tenni és megőrizni? Ma is léteznek adatközlésre (vagy legalábbis közzétett adatok leírásainak publikálására) szakosodott folyóiratok – ilyen a *Scientific Data* (*Nature Publishing Group*) vagy a *Journal of Astronomical Data*. A *Nature* általánosságban követeli meg az adatok elérhetővé tételét.[10] A hazai *Information Bulletin on Variable Stars* a cikkek mellett közli az adatokat is az interneten. Többnyire azonban a kiadók nem kívánnak adatokkal foglalkozni – az adatokat repozitóriumban kell elhelyezni és DOI azonosítóval ellátni. Ezekre az azonosítókra lehet a cikkekben hivatkozni. Léteznek általános, adatok megosztására használható repozitóriumok, mint a *figshare*, és vannak egyes tudományterületeken használtak, mint a *Dryad*[11]. Adatok kerülhetnek intézményi repozitóriumokba is – az MTA KIK REAL-jában is vannak DOI azonosítóval ellátott, egy, a PLoS ONE-ban megjelent cikkhez kapcsolódó adatállományok.

Az adatállományok azonosítására mára egyértelműen a DOI használata terjed. Kifejezetten adatok azonosítására szerveződött a *DataCite*[12] ügynökség, melynek az MTA KIK is tagja. A DOI azonosítóhoz leíró, az állomány megtalálásának céljára való (discovery level) metaadatokat kell megadni. A cél itt elsősorban az idézhetőség megteremtése. (Az adatállományok leírására bonyolultabb, hierarchikus sémák szolgálhatnak, ezek a jogi, technikai, származási jellegű metaadatok mellett részletes szakmai metaadatokat is kell, hogy tartalmazzanak.) A szakmai leírás követelménye miatt sok esetben tudományterületi, szakosodott repozitóriumokban való elhelyezést kívánhat. A megfelelő leírás követelménye erős érv az adatok és publikációk (cikkek, monográfiák) szoros kapcsolata mellett. Már foglalkoztunk azzal, hogy a cikk szempontjából miért szükségesek az adatok. De az adatok szempontjából is szükséges a kapcsolat megteremtése olyan cikkekkel, amelyekben az adatgyűjtés motivációja, módszere, az adatkezelés le van írva, sőt, ahol esetleg egy tudományos felhasználás is szerepel. Ha a kutatási adatok nem a folyóiratoknál, nem a cikkekhez kapcsolva találhatóak meg, hanem egy repozitóriumban, adatbankban, fontos hogy a metaadatok között az adatokat leíró, az adatokat feldolgozó tudományos közlemények bibliográfiai azonosítói is bekerüljenek.

A kutatási alapok, tudományfinanszírozók egyre gyakrabban követelik meg az adatok elérhetővé tételét. Ez a feltétel hosszú ideje szerepel már az OTKA szerződésekben is. Egyre több kutatási alap követeli meg a pályázatok beadásánál, szerződéskötésnél az adatok kezelésének tervezését (Data Management Plan).

Ugyanakkor sokszor be lehet tervezni az adatok feldolgozásának, archiválásának költségeit is.

Megjelenik az EU Horizont 2020 programjában is a tudományos adatok kezelésének kérdése. Az adatkezelésre való felkészülés szerepel az OpenAIRE projekt jelenlegi és ezelőtti fázisában is. Az Európai Bizottság idézett állásfoglalása a tagállamokat is szabályozás kimunkálására kötelezi. Hasonló irányelvek, célkitűzések megjelennek az OECD és az UNESCO különböző dokumentumaiban is. Kathleen Shearer (COAR) készített egy összeállítást a kutatási adatokra vonatkozó politikákról[13].

A kutatás minőségének emeléséhez lényegesen hozzájárul az előzetes tervezés, a megfelelő adatkezelés. A legnagyobb nyertes talán nem is a tudományos közösség, hanem az adott projekt. Ennek érdekében azonban nem csak a kutatók tehetnek, hanem a műszer- és szoftvergyártók is. A megfelelő procedúrák alkalmazása, a szabványos formátumokban való rögzítés, valamint az elegendő mennyiségű és minőségű metaadat alkalmazása, amennyiben a kutatási folyamatba, és az alkalmazott eszközökbe beépülnek, nem jelentenek túlzott terhet a kutatók számára.

Kutatási adat sokkal több fajta lehet, mint publikáció. Az adattípusok száma egy-egy szakterületen belül is nehezen számbavehető. Szöveg, táblázat, hang, kép, videó és rengeteg más, komplexebb adatstruktúra. A tudományos adatmenedzsment nem húzható egy kaptafára – legfeljebb nagyon magas szinten. A megaprojektek adatainak megvannak a maguk adatbázisai – mint például az Large Hadron Collider vagy a Human Genome Project esetében. Még egy-egy tudományban is nehéz szabványosítani, de nem lehetetlen. Példa a Virtuális Obszervatórium[14] a csillagászat területén.

Vajon hozzáférhetőek, olvashatóak, felhasználhatóak lesznek-e a mai adatok évtizedek múlva? Persze erre nem mindig lesz szükség – akkorra már lehet, hogy pontosabb, jobb adatok egy korábbi vizsgálat adatait elavulttá teszik. De sokszor éppen a hosszú távú adatgyűjtés teremti meg egy tudományos kérdés vizsgálatának lehetőségét. Sokszor hosszú idő múltán merül fel a kérdés: egy korábbi cikk megállapításai vajon megalapozottak voltak-e? A hozzáférhetőség biztosítására szolgáló megoldás a DOI-k alkalmazása: az adatállomány URL-jének megváltozását a DOI linkek alkalmazása követni tudja. Az adatformátumok megfelelő megválasztása, szabványos formátumok alkalmazása elősegítheti az olvashatóság fenntartását, az adatok integritását, romlatlanságát, az adatbiztonságot informatikai megoldások garantálhatják. A jó dokumentáció a felhasználhatóság alapfeltétele. Minderről a kutatási projekt esetén, az adatot gyűjtő, mérő kutatóknak kell gondoskodniuk, amíg a projektben erre van pénz. A gondosan előkészített adatok hosszú távú tárolása már nem kerül sokba, és a költségeket, mindaddig amíg az adatok mennyiségének és a technológia fejlődésének exponenciális növekedése tart, a kutatási költségvetés biztosítani tudja (mint ahogy egy exponenciálisan növekedő populációban a dolgozók befizetései a nyugdíjakat fedezni tudják).

Nemzetközi fejlemények. Az OpenAIRE2020-ban Research Data Pilot program indul.[15] A Horizon 2020-ban hét kulcsterületen kötelező lesz az adatok repozitóriumba helyezése. Árva repozitóriumként a CERN-ben fejlesztett *Zenodo*[16] szolgál. A Frontiers kiadó új folyóirata a Frontiers Data Reports.[17] Az ERC 2014-ben rendezett műhelymunkát kutatási adatok kezeléséről és megosztásáról.[18] 2012-ben a Royal Society kiadta a „Science as an Open Enterprise”[19], 2013-ban a LERU a Roadmap for Research Data[20] című dokumentumot.

Hazai fejleményekről is beszámolhatunk. Már említettük az MTA KIK DataCite tagságát. Ennek következményeként térítésmentesen tudunk DOI azonosítókat biztosítani adatállományoknak és szürke irodalomnak.[21] Az MTMT 6.2 verziójában megjelent a kutatási adat típus, lehetővé vált a publikus kutatási adatok és a rájuk kapott idézetek nyilvántartása. Megjelent az első, jelentős mértékben adatok és dokumentációk archiválására és hozzáférhetővé tételére szolgáló repozitórium az MTA TK KDK-ban.[22]

Irodalom

HOLL András: Szövegbányászat, adatbányászat, ismeretfeltárás. Magyar Tudomány. 2015. 6. p. 680-685. http://www.matud.iif.hu/2015/06/05.htm

HOLL András: Információáradat és hullámlovaglás. Magyar Tudomány. 2013. 4. p. 473−478. *http://www.matud.iif.hu/2013/04/13.htm*

MICSIK András − GÁRDOS Judit (2014): Tudományos repozitóriumok az MTA-ban: a KDK és a SZTAKI tanulságai. = Informatika a felsőoktatásban, 2014.08.27−2014.08.29, Debrecen, Hungary. *http://eprints.sztaki.hu/8017/*

VARGHA Domokosné: Konkoly Thege Miklós magyar nyelvű írásai.
Magyar Tudomány, 2001. július, p. 867. *http://www.matud.iif.hu/01jul/vargha.html*

Hivatkozások, megjegyzések

1. Boston University Libraries: Research Data Management *http://www.bu.edu/datamanagement/background/whatisdata/*
2. University of Edinburgh. Idézi: University of Leicester: Research Data *http://www2.le.ac.uk/services/research-data/rdm/what-is-rdm/research-data*
3. Engineering and Physical Sciences Research Council. Idézi: University of Leicester: Research Data
4. HGP: *http://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml*
5. HST adatok a MAST-ban: *https://archive.stsci.edu/hst/*
6. *http://imgsrc.hubblesite.org/hu/db/images/hs-2011-40-a-print.jpg*
7. 2012/417/EU *http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:194:0039:0043:EN:PDF*
8. *http://kdk.tk.mta.hu/adatmenedzsment*
9. PLoS data policy: *http://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy/*
10. *http://www.nature.com/authors/policies/availability.html*
11. Dryad Digital Repository: *http://datadryad.org/*
12. DataCite: *https://www.datacite.org/*
13. Scan of International Funder Policies for Research Data Management/Sharing. Egyelőre nem publikált, csak a COAR levelezőlistán megosztva
14. Lásd a „Virtual Observatory” bejegyzést a Wikipedia-ban
15. *http://europa.eu/rapid/press-release_IP-13-1257_en.htm*
16. Zenodo: https://zenodo.org/
17. *http://www.frontiersin.org/news/Data_Reports_a_new_type_of_peer-reviewed_article_in_Frontiers_journals/1051?utm_source=FRN&utm_medium=MRKT&utm_campaign=TOC_FRN_1502_DATA*

[18] ERC Workshop on Research Data Management and Sharing
http://erc.europa.eu/media-and-events/events/erc-workshop-research-data-management-and-sharing

[19] *https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf*

[20] *http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf*

[21] MTA KIK Szakinformatikai Osztály, DOI Iroda. (elérhető: *doi-info@konyvtar.mta.hu)*

[22] MTA TK KDK repozitórium: *http://openarchive.tk.mta.hu/*

Beérkezett: 2015. IV. 13-án.

Holl András
az MTA KIK informatikai főigazgató-helyettese
E-mail: *holl.andras@konyvtar.mta.hu*