Magyar Tudomány 179(2018)7, 1001–1016
DOI: 10.1556/2065.179.2018.7.8

ADATKEZELÉS – A (KUTATÁSI) ADATOK KEZELÉSE A MAGYAR TUDOMÁNYOS ÉS MEMÓRIAINTÉZMÉNYEKBEN

DATA HANDLING – RESEARCH DATA MANAGEMENT IN HUNGARIAN SCIENTIFIC AND MEMORY INSTITUTIONS

Kovács László
osztályvezető, MTA SZTAKI Elosztott Rendszerek Osztály
laszlo.kovacs@sztaki.mta.hu

ÖSSZEFOGLALÁS

Az MTA Könyvtár és Információs Központ, az MTA Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központja, valamint az MTA Számítástechnikai és Automatizálási Kutatóintézet *Kutatási adatok kezelése* címmel műhelykonferenciát rendezett 2015. május 14-én. A műhelykonferencia főbb, csak kéziratban rendelkezésre álló megállapításai mentén mutatjuk be a kutatási adatok kezelésének néhány kérdését, a kutatási szféra elmúlt húsz éve adatkezelési gyakorlatának jellemző vonásait. A tudomány és a kulturális örökség szféráiban lévő (memória) intézmények hasonló adatkezelési problémái és gyakorlata miatt megállapításaink többsége a memóriaintézményekre is érvényes lehet.

ABSTRACT

Some institutions of MTA – the Hungarian Academy of Sciences – as the MTA KIK, the Library and Information Centre, the MTA TK, the Centre for Social Sciences and the MTA SZTAKI, the Institute for Computer Science and Control organized a workshop on 14 May 2015 under the title of *Handling Research Data*. Actual issues of research data handling and the major characteristics of data management practice of the last 20 years of the research sphere are presented here via the yet unpublished notes of the workshop. Similarity between problems and practices of data handling within the research and the cultural heritage spheres leads us to conclusions that may be valid for the memory institutions as well.

Kulcsszavak: kutatási adatok digitális kezelése, DMP, adatkezelési terv, repozitórium, adatsiló, digitalizálás, FAIR-követelmények, adatinfrastruktúra, tartalominfrastruktúra, kapcsolt adatok, hosszú távú digitális megőrzés

Keywords: digital research data handling and management, DMP, Data Management Plan, repository, data silo, digitalisation, FAIR requirements, data infrastructure, content infrastructure, linked-data, long term digital preservation

BEVEZETÉS

Az adatvezérelt tudományos kutatási tevékenységek széles körű elterjedése, a tudományok adatigényének drasztikus növekedése oda vezetett, hogy az adatkezelés és feldolgozás a különféle diszciplínákban szignifikánssá, több diszciplína esetében pedig a kutatási tevékenység alapvető meghatározójává vált.

Az MTA Könyvtár és Információs Központ, az MTA Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központja, valamint az MTA Számítástechnikai és Automatizálási Kutatóintézet *Kutatási adatok kezelése* című műhelykonferenciájának (Kovács et al., 2015) részt vevői közös problémákat és tendenciákat véltek felfedezni a kutatási adatok magyar kezelésével kapcsolatosan.

TUDOMÁNYOS ADAT KELETKEZÉSE

A kutatási és memóriaintézményekben jelentős mennyiségű (tudományos) adat keletkezik. A kezelendő adatok között a nyers adatoktól kezdve a különféle feldolgozottságú adatokon keresztül egészen a publikálásra kerülő adatállományokig mindenfajta és -féle adat megtalálható. Az adatállományok mérete, milyensége, formátuma, többek között, függ az egyes tudományágakban járatos adatkezelési szokásrendszerektől, a konkrét kutatási tevékenységektől és/vagy az alkalmazott mérő és regisztráló rendszerektől.

Az elmúlt húsz évben jelentős elmozdulás volt tapasztalható az analóg adatoktól a már eleve digitálisan keletkező (born digital) adatok felé, a digitális adatkezelési igény manapság már szinte egyeduralkodóvá vált világszerte.

DIGITALIZÁLÁS

Az igény a korábban analóg formában rögzített adatok digitalizálására ugyanakkor nem csökkent, a digitalizálás folyamata eleddig, bár különféle intenzitással, de folyamatos volt, és e folyamat mind a mai napig nem fejeződött be hazánkban. Ennek elsődleges oka a digitalizálás mint tevékenység támogatásának hazai forráshiánya, de mint később látni fogjuk, a digitalizáláshoz szorosan kapcsolódó egyéb tevékenységek, a nem elégséges munkaráfordítás és a szakmai hozzáértés hiányosságai is visszatartják a digitalizálás folyamatát.

A digitalizálás motivációja tekintetében a szakmai közvélekedés három szempontot szokott említeni: az értékmegőrzést (A), a digitális adatkezelés egyszerűségét és olcsóbb voltát (B), valamint a digitális adatterjesztés univerzális lehetőségét (C).

A) Az analóg formátumú adatok, pontosabban az analóg adathordozók (például filmszalagok) kezelését az idő előrehaladtával egyre nehezebb fenntartani, a fizikai romlás látható jeleinek prognosztizálható hatására az adatkezelők előszeretettel választják a digitalizálást mint adatmentési, adatmegőrzési mechanizmust. Teszik mindezt abban a (hamis) tudatban, hogy a digitális térben az adatok megőrzése egyszerűbb, olcsóbb, hatékonyabb.

Mint később látni fogjuk, ez távolról sincs így, e (hamis) tudatot elsősorban a mindennapi életünket lehetővé tévő, mindenütt jelen lévő digitális számítástechnikai és kommunikációs eszközök milliárdjainak megléte és mindennapos használata hozza létre, ugyanakkor megalapozott informatikai/gazdasági megfontolások kevésbé játszanak itt szerepet.

B) A digitális adatkezelés finanszírozhatóságát lényegesen befolyásolja az informatikai piaci verseny árcsökkentő hatása, ugyanakkor a közbeszerzések pontosan az ellentétes mozgásokat indukálják. A szférában kezelendő adatmennyiség exponenciális növekedése (eltekintve a *Big Data*-tárolás és -feldolgozás speciális problémakörétől), az adathordozók, tárolók áralakulása révén, időben meglepően állandó finanszírozási igényt mutat.

C) A digitális adatok és információk átvitele, szállítása nagyságrendekkel olcsóbbá vált, köszönhetően az új kommunikációs lehetőségeknek és kiépült kapacitásoknak. A web általánossá válása a világméretű hálózaton keresztüli adatterjesztést forradalmasította. A digitalizálás és a digitális entitások interneten történő közzététele valóban jelentős motivációs erő, mely előmozdítja a digitalizálás folyamatát.

ADATKÖZLÉS WEBEN KERESZTÜL

Az állagmegőrzés, illetve az univerzális digitális terjesztés motiválta magyar digitalizálási gyakorlatok során azonban legtöbbször nem vették figyelembe azt, hogy mekkora a digitális információk felhasználásának tervezett kiterjedése, a potenciális felhasználói bázis prognosztizálható mérete, és az adat- és információfelhasználásnak milyen tovagyűrűző hatása keletkezik a tudományban magában vagy a gazdaság, az oktatás stb. szféráiban.

Hatáselemzések hiányában számos kielégítetlen adatigény mellett jelentős számosságú adatközlő mű kapott finanszírozást (web-honlap, -portál formájában) a 2000-es években. E webszolgálatok létrehozása egyrészről valóban megteremtette az univerzális és globális adathozzáférést hálózaton keresztül, ugyanakkor katasztrofális következményekkel járt, ha ezen webszolgálatok fenntarthatóságát tekintjük.

A digitális információk webes (például honlapon keresztüli) közlése az informatikailag sokszor alulképzett döntéshozók számára a digitális adattárolás és

adatközlés együttes, egy csapásra történő megoldásának potenciális lehetőségét jelentette. A modern honlap-/portáltechnológiák (a korai honlapoktól eltérően) ténylegesen tartalmaznak adattárolásra szolgáló adatbázis-kezelőt, hozzáférést adatbázis-kezelőkhöz és ezzel párhuzamosan, a webes kiszolgálásra szolgáló adatközlő modulokat is. A beépített adatbázis-kezelők azonban egyértelműen a webszolgáltatás sajátos, informatikamotiválta igényeit szolgálták/-ják ki, és csak igen kevéssé a tárolt adatok kezelésének, tisztításának, megőrzésének stb. céljait. Egyszóval az adatközlést, nem pedig az adatkezelést támogatják.

A portáltechnológiák alkalmazásának ilyetén félreértése nemzetgazdasági szinten vezetett veszteségekhez, ugyanis a fent nevezett honlapalapú projektek elkészülte után pár évvel azzal a problémával szembesültek a fenntartók, hogy a webvilág technológiai változási sebessége az egyik leggyorsabb a világon.

A portálok esetében három-öt évenként biztosan cserélni-módosítani-javítani szükséges a meghajtó portálmotorokat, különben a webszolgálat használhatósága szignifikánsan romolhat, vagy szélsőséges esetben az adatok hálózati elérhetősége akár meg is szűnhet. A webböngészők folyamatos változása, a felhasználói felületek újabb kiszolgáló technológiái és az új interakciós lehetőségek fejlődése (és néha persze a technológiai divat) állandó fejlesztési készenlétet és/vagy beavatkozást igényel. Mindez folyamatos és szignifikáns finanszírozási feladatot jelent.

Összefoglalóan az történt, hogy a digitális adatkezelés/tárolás/közvetítés webes kontextusba helyezésével egyidejűleg egy folyamatos technológiakövetési feladatot is magukra vettek a döntéshozók anélkül, hogy ennek műszaki feltételrendszerét, szervezeti és munkaigényét, finanszírozhatóságát előre látták és biztosíthatták volna. A webszolgálatok fenntarthatatlansága pedig nem csupán az adatelérhetőséget, de magát az adatok létének fennmaradását is veszélyeztette, vagy ténylegesen meg is szüntette, jelentős nemzetgazdasági károkat okozva ezzel.

A tudományos adatok és információk (mint értelmezett adatok) webes közlése tehát nemhogy megoldotta volna az analóg adattárolás során felmerülő problémákat, de egy újabb problémával tetézte azokat. Ebből a nézőpontból az elkésett magyar fejlődésnek, a digitalizálásban észlelhető, a fejlett országokhoz képesti lemaradásunknak (mely kb. tíz-tizenöt évre becsülhető) kicsiny pozitív hozadéka lehet (ha az analóg adathordozók fizikai állapota, illetve tervezett élettartama megengedi) e csapdahelyzet (részleges) kikerülése és a 2010-es évek elején-közepén kezdődő, szakmailag igenelhető adatkezelési gyakorlat művelése.

REPOZITÓRIUMOK MAGYARORSZÁGON

A digitális repozitóriumok magyarországi elterjedésének kezdete a 2010-es évek elejére datálható, ami szintén legalább egy évtizednyi lemaradást jelent a fejlettebb országok gyakorlatához képest. A repozitóriumok felállítására vonatkozó

döntések a közvetlen felhasználási igényeken túlmenően többek között azon alapulnak, hogy az adatkezelés, az adatkuráció műszakilag, technológiailag és tevékenység szempontjából így célszerűen elválasztható az adatközlési feladatoktól. Adattárolók, repozitóriumok, digitális könyvtárak és archívumok létesítésével egyidejűleg az adatkezelési feladatok (tárolás, tisztítás, válogatás, hosszú távú megőrzés stb.) funkcionális szétválasztása és az eltérő technológia támogatása így megvalósulhat. A repozitóriumokban a tárolt adatokhoz társított metaadatok sémája egyértelműsíthető, jól meghatározható, ami által a későbbi visszakereshetőség és a (szemantikusan) helyes felhasználhatóság alapja teremtődik meg.

Magyarországon a digitális repozitórium fogalmán egy, elsősorban a digitális dokumentumok (például publikációk) avagy multimédia entitások (hang- és videóanyag stb.) tárolására szolgáló rendszert értenek, a repozitórium sokkal kevésbé jelent (ma még) adattároló (data repository) rendszert. Míg a magyar tudományosságban ténylegesen kiépült a publikációk nyílt hozzáférését előíró MTA-rendelet indukálta országos repozitóriumhálózat (az MTA REAL és a kutatóintézetekhez, egyetemekhez kötött intézményi repozitóriumok hálózati rendszere) (URL1), addig szándékosan és célzottan a csupán tudományos adattárolás célú intézményi repozitóriumok létrehozására még csak az első lépések történtek meg.

Létrejöttek az első, célzottan kutatási adatokat tartalmazó hazai adatrepozitóriumok: MTA KRTK Adatbank (URL2), MTA TK KDK (Micsik–Gárdos, 2014), (URL3). A kutatók és adatállományok azonosítására alkalmas azonosítók elérhetők Magyarországról is: ORCID (URL4), DOI (URL5). A REAL alkalmas DOI-val ellátható adatok elhelyezésére is, bár erre még kevés a konkrét példa. Az MTMT (URL6) adathivatkozást is képes kezelni. Létrejött a hazai repozitóriumokban tárolt entitások közös országos keresője (URL7).

Az adatpublikálás problémáiról és az azzal párhuzamos Open Data mozgalomról itt és most nem ejtünk több szót, megtették azt már mások korábban (Micsik–Gárdos, 2014; Holl, 2015, 2016), csak a repozitórium rendszerek létrehozását és fenntarthatóságát elemezzük.

A repozitóriumokat – Magyarországon is – sokkal inkább intézmények, mintsem kutató közösségek hozzák létre és tartják fenn; elsősorban a saját munkatársaik kiszolgálását célozva. A tárolók létrehozásához szükséges források megszerzése, a tárolók fenntartása intézményi keretek között tervezhető, és a hosszú távú fennmaradás esélye is nagyobb.

A magyar intézményi repozitóriumok és azok országos hálózatának stabilitását segíti elő a repozitóriumok minősítési rendszerének bevezetése (URL8), mely (nagyon helyesen) leginkább a repozitóriumok hosszabb távú fenntartásának feltételrendszerét kéri számon.

Határozott előrelépés lehetne az országban egy, a REAL-hoz hasonló, de célzottan adattárolásra alkalmas, országosan megszervezett adatrepozitórium- és/

vagy adatsilóhálózat felállítása, valamint a projektek befejeztével a kutatási adatállományok megkövetelt elhelyezése ezekben a silókban, legalább a közpénzen finanszírozott kutatások esetében. E koncepció azonban pontosításra szorul a legújabb technológiai fejlemények hatására.

ADATSILÓK LÉTREHOZÁSA

Az individuális, egy adott intézményhez kötött, egy-egy adatállományt biztosító adattárolók, repozitóriumok mellett az adatsilók megjelenése okozott nagyobb változást. Az adatsiló definíció szerint leginkább az intézmények közötti, ha tetszik intézményfüggetlen adattárolás céljait szolgáló rendszer, amelynek fenntartása akkor is biztosított, ha az intézmények megszűnnek, átalakulnak, avagy belső infrastrukturális változtatás miatt a korábban üzemeltetett intézményi repozitórium működése esetleg veszélybe kerül. Ezzel ellentétben az adatsilók fenntartása legtöbbször nagyvállalatok, egyetemek, kutatóintézetek vagy azok kisebb részlegei hatáskörébe tartoznak, és ezért nem szükségszerűen biztosítják az intézményfüggetlenségi elvárást.

MULTIDISZCIPLINÁRIS ADATSILÓK

Az adatsilók tartalmuk szerint lehetnek egy tudományágat támogatók vagy multidiszciplinárisak. A tudományban az adatsilók létrehozásának gyakorlata eltér az egyes diszciplínák között. Diszciplínákhoz kötött, tematikusan homogén adatsilók mellett a multidiszciplináris adatsilók megjelenése új szintre emeli az adatfelhasználhatóságot, ugyanis az az inhomogén, eltérő sémákkal rendelkező, tematikusan, diszciplinárisan eltérő adatok adattársításának lehetőségét segíti elő.

A multidiszciplináris adatsiló biztosítani tudja a mostanában kiemelkedő adattudomány (data science) analitikai eszközeinek hatékony felhasználhatóságát, közvetlen és egyszerű hozzáférést engedve e társítható, esetleg eltérő sémákkal rendelkező adatállományokhoz. Mint később látni fogjuk, az adattársítás direkt és közvetlen módszerei ma még beláthatatlan kihatással kecsegtetnek, és a tudomány globális fejlődésének újabb forradalmát eredményezhetik.

ADATKEZELÉSI GYAKORLAT

Világszerte jelentős eltérés tapasztalható az egyes kutatási intézmények, sőt azon belül az egyes kutatási projektek adatkezelésével, annak céljaival, minőségével, szervezettségével kapcsolatosan. A különbségek leginkább a kis és közepes adat-

állományok kezelése tekintetében jelentősek, míg a nagytömegű adatkezelés (Big Data), a jól kialakult adatkezelési szokások és az adatkezelés, -felhasználás és adatanalízis nagy, infrastrukturális, időben stabil komplex rendszereinek nyomatéka miatt (például elemi részek fizikája, CERN adatkezelése) jobban szervezett. A különbségek másik, el nem hanyagolható része az eltérő diszciplínákban járatos eltérő adatkezelési eljárásokból és technológiákból származik.

Az adatkezelés funkcionális tartalma a tudományban legtöbbször a következőket jelenti: az adatok forrásainak és az adatelérés módozatainak meghatározása, interfészek, kommunikációs formák és kommunikációs szolgálatok, technológiák és rendszerek, hozzáférési kommunikációs protokollok (például OAI-PMH (URL15) adataratási logikák) meghatározása. Az adatokon végzett transzformációs műveletek összessége, így például az adatok begyűjtése, felvétele, rögzítése, rendszerezése, szűrése, válogatása, tárolása, megváltoztatása, formátum transzformációi, az adatok felhasználása, importálása, exportálása, szállítása, továbbítása, nyilvánosságra hozatala, publikálása. Ide tartozik még az adatok összekapcsolása, zárolása, törlése és megsemmisítése, az adatok további felhasználásának biztonságos megakadályozása. Az adatvesztés kiküszöbölése és az adatokhoz való hozzáférés szabályozása, az adatvédelem ugyancsak integráns része e funkcionális fogalmi kiterjedésnek.

Mint látjuk a tudományos adatok kezelése rendkívül összetett és szerteágazó feladatot jelent, és mint ilyen, digitális adatkezelési szakértelmet követel meg. E szakértelem elméleti alapjait az informatika- és a könyvtártudomány, gyakorlati megvalósulását pedig az informatika gyakorlata teremti meg. E szakértelem tehát vagy magukban a kutatókban, vagy a kutatók és informatikusok/digitális könyvtárosok kooperációjában testesülhet meg.

A műhelykonferencia résztvevőinek szinte egyöntetű véleménye alapján e téren jelenleg nagyfokú szakértelemhiány mutatkozik a magyar tudományosság szinte minden szférájában és diszciplínájában. A kutatók informatikusokkal való együttműködési igényének kielégületlenségét jelen sorok szerzője saját, az MTA kutatóintézeti hálózatából származó közvetlen tapasztalataival tudja megerősíteni.

A modern digitális adatkezelési informatikai technológiák, módszerek, gyakorlatok ismeretének hiánya megdöbbentően elmaradott adatkezelési, adattárolási rendszerek meglétét és aktuális használatát jelenti Magyarországon. Tapasztalatunk alapján a magyar bölcsészet- és társadalomtudományok művelői jelentős hátrányban vannak ilyen tekintetben a természet- és műszaki tudományok művelőihez képest. A magyar memóriaintézmények jelenlegi helyzete pedig tragikus, ezen intézményekben mindenhol (minőségi) informatikushiány mutatkozik. Mindez az alkalmazott adatkezelési rendszerek minőségében, technológiai fejletlenségében csapódik le, pontosabban a nem alkalmazott fejlett technológiák és a *state-of-the-art* ismeretének hiányában.

ADATKEZELÉS KÖLTSÉGE

A magyar kutatási projektek előkészítése során rendszerint hiányzik a projektek adatkezelésének megtervezése, így általában nem tervezik ennek költségeit sem. Az adatkezelési feladatokat legtöbbször a kutatók végzik, az (informatikai) eszközök beszerzésének terve mellett az explicit adatkezelési költség- és tevékenységbecslés ritka.

Jelentős probléma, hogy hiányzik a rendszerek és bennük az adatszolgáltatások hosszú távú, a projekt befejezése utáni fenntartási költségeinek tervezése is. A magyar finanszírozó szervezetek legtöbbször nem gondolkoznak a projekt futamidején túlmenően, alig követelik meg a kutatási projektek utáni követési, fenntartási, hasznosítási feladatok keretében az adatállományok túlélésének biztosítását és persze e feladatok reális finanszírozásának megteremtését és/vagy támogatását. Ennek súlyos következménye a projektekben keletkező vagy az ott kezelt (és ezáltal jelentős mennyiségű élőmunkát, értéket hordozó) adatállományok, adatszolgáltatások továbbélésének, hasznosulásának, egyáltalán fennmaradásának veszélyeztetése.

Hiányzik egy, legalább ágazati szintű elvárásrendszer, adatkezelési stratégia, szabályzat, útmutatás arról, hogy hogyan óvjuk meg ezeket az adatállományokat a projektek befejezte után. Nemzetközi pályázatoknál (például Horizon 2020) ugyanakkor elterjedt az adatkezelési terv (DMP – Data Management Plan) megkövetelése. Ilyen projektek esetében az adatkezelési terv a projektek során begyűjtött, feldolgozott és/vagy létrehozott kutatási adatok kezelésének teljes életciklusára ki kell hogy terjedjen, információkat biztosítva a FAIR-követelmények tervezett megvalósulásáról.

FAIR-KÖVETELMÉNYEK

A FAIR-követelmények (Findable, Accessible, Interoperable and Re-usable) a kutatási adatok projektek futása közbeni és utáni megtalálhatóságát, a széles körű hozzáférés biztosítását, az adatok csereszabatosságát, illetve az újrahasznosításhoz szükséges feltételek és metaadatok meglétét követeli meg (Wilkinson–Dumontier, 2016). Mindez azt jelenti, hogy az adatkezelési tervben egyértelműen definiálni kell azt, hogy a projekt során milyen adatokat gyűjtenek be, azokat hogyan, milyen módszerrel dolgozzák fel, valamint milyen új vagy származtatott adatok keletkeznek a projekt során. Az adatformátum szabványoknak való megfelelés, a használt szabványok egyértelmű meghatározásán túlmenően az adatkezelési terv tartalmazza az adatok közzétételének, megosztásának tervezett módozatait, a hozzáférés biztosításának módszereit, jogi, műszaki, szervezeti feltételeit. Információ szükséges arról is, hogy a kutatási

adatok feldolgozási folyamata során milyen adatkurátori (válogatás, szűrés, aggregáció stb.) munkát terveznek végezni, és a projekt befejezte után hogyan fogják az adatokat megőrizni és/vagy újrahasznosítani, különféle időintervallumokat feltételezve, és azt, hogy az újrahasznosítást milyen származási/nyomkövetés jellegű metaadatok (provenance metadata) támogatják. Az adatkezelési terv, elvárás szerint, foglalkozik az adatkezelés tervezett költségeivel, a projektek lezárulása után felmerülő, hosszú távon jelentkező fenntartási költségekkel is.

Bár az adatkezelési terv megkövetelése magyar viszonyok között nem jellemző, ugyanakkor diszciplínaspecifikus adatkezelési mintatervek kidolgozása és elterjesztése jelentősen segíthetné a jelenlegi magyar adatkezelési gyakorlat javítását.

A magyar tudomány jelenleg ugyancsak kevéssé alapozhat a tudományos adat- és tartalomkezelés olyan nagy léptékű modellkísérleteire, amelyek mint „best practice", megfontolandó, esetleg átvehető mintaként szolgálhatnak az egyes kutatóhelyeken.

Hiányoznak vagy hiányosak a kutatási adatok és digitális tartalmak kezelését lehetővé tévő funkcionális, architekturális, technológiai, működési (és egyes esetekben üzleti) modellek és rendszerek, szabványok, szabályzatok, jogok, adat- és információszolgáltatások, regiszterek és repozitóriumok (adatbázisok, adattárak, adattárházak, digitális gyűjtemények), valamint ezek interoperábilis rendszere, tehát mindaz, amely a tudományos adatkezelés nemzeti és intézményi szintű rendszerkontextusát, tágabb értelmű infrastruktúráját adná.

ADAT- ÉS TARTALOMINFRASTRUKTÚRA

Az adatkezelés infrastrukturális megközelítése, mint új fejlemény, azon a felismerésen alapul, hogy az adatszolgáltatások egyedileg, önmagukban nem életképesek, hanem más adatszolgáltatások egymáshoz harmonikusan, informatikailag és szemantikusan is illeszkedő rendszerében, egy tervezett hálózatban tudnak csak rendesen létezni, működni. Egy könnyen átlátható példa a magyarországi névterek problematikája.

A memóriaintézmények regisztereiben az intézményközi névtérkezelés azonban már legalább húsz éve megoldatlan az országban, annak ellenére, hogy ezalatt szinte folyamatos (volt) a nevekkel, névterekkel való foglalkozás, névtérállomány-építés és -kezelés, egyéni és intézményi szinteken is.

A névtérkezeléshez szükséges, országos szinten jelentkező globális szervező, megvalósító, finanszírozó tevékenységeket eddig sem az állam, sem pedig valamilyen intézményi önszerveződés nem tudta fenntartható módon megvalósítani. Ennek következménye a memóriaintézményekben jelentkező, feleslegesen párhu-

zamos névkezelési munkák miatti erőforrás-pazarlás, egyben a névkezelés-minőség optimumának elérhetetlensége, mely gátolja az oktatási, kutatási, kulturális, sőt még a kormányzati szféra különféle tevékenységeit is.

Az Európai Unióban országokon keresztülnyúló hálózatok, hálózati infrastruktúrák támogatják a digitális tudományos adatkezelés legújabb, adatfelhő alapú megközelítéseit (például nemzetközi DARIAH-infrastruktúra a digitális bölcsészettudományok területén [URL9]), sőt azon túlmenően a tudomány elektronikus művelése, az eScience-funkciók teljes vertikumát. Magyarországon ugyanebben az időben a legfelső szinten kérdőjelezik meg a digitális bölcsészet diszciplináris létét, az azt támogató felhőalapú digitális infrastruktúrák létesítésének szükségességét, az ilyen célú projektek támogathatóságát, lásd például (MTA BTK–MTA SZTAKI–DE–ME, 2016) GINOP-pályázat, a nemzetközi főáramtól való leszakadást indukálva e tudományok területén.

A magyar tudományban jellemző, hogy széles körben hiányzik az adatkezelés infrastrukturális megközelítése, az, hogy a digitális kutatási adatok létrehozását, feldolgozását, megtalálását, tárolását, szállítását, felhasználását, megőrzését stb. – egyszóval a digitális adat- és tartalomkezelést lehetővé tévő rendszereket – egységes digitális adattartalom-infrastruktúráknak tekintsük, és mint ilyeneket (országos vagy ágazati szinten) központilag tervezzük, létrehozzuk és fenntartsuk. A nemzetközi, hasonló célú kezdeményezésekbe, infrastruktúrákba való belépésünk, csatlakozásunk, az adatkapcsolati szintű kapcsolatunk (adat import/export) csak akkor lehetséges, ha rendelkezünk ilyen célú, stabilan működő hazai adatinfrastruktúrákkal. Míg a kommunikációs és internethálózatokat a közfelfogás infrastruktúrának tekinti, az ezekre szervesen épülő adat- és tartalominfrastruktúrák fogalma ma még nem elterjedt a közbeszédben.

Ezekben az adatinfrastruktúrákban, az adatszolgáltatások között különleges szerepet kapnak az interoperabilitást elősegítő szolgáltatások (protokollregiszterek, egymással interoperábilis névtér- és ontológiakezelők, a metaadatsémák, felhasználási profilok (application profile) regiszterei, a szótárak, szószedetek és egyéb nyelvi szolgáltatások stb.), egyszóval azok a központi adatszolgálatok, melyek egységes rendszerré fűzik fel az egyedi, kutatás célú adatszolgálatokat. Az adatszolgálatok és azok megbízható interoperabilitásának létrehozása szabványos informatikai megoldásokat, egységes mérnöki tervezést, megvalósítást és persze szabályozást igényel.

Az ilyen nagy nemzeti infrastruktúrák ugyan létre tudnak jönni projektfinanszírozási logikával, de hosszú távú (akár több évtizedre szóló) fenntartásuk csak az erre a célra létrehozott speciális intézmények keretében, jól kidolgozott üzleti-finanszírozási és/vagy működési modellek alapján lehet reális. Ezeket az intézményeket meg kell alapítani, létre kell hozni, működtetni és persze finanszírozni kell.

ADATKEZELÉS MINT KUTATÓI ERŐFESZÍTÉS

Az adatkezelést lehetővé tévő hardver- és szoftvereszközök használata esetén felmerül az a kérdés, hogy milyen szakértelemre van szükség az adatfeldolgozási folyamatban, annak teljes életciklusában, a különféle munkafázisokban? Megvan-e a szükséges szakértelem az adatkezelést legtöbbször *de facto* végző kutatókban, illetve a kutató-informatikus, a kutató-könyvtáros együttműködési viszonylatokban? Az eltérő diszciplínák eltérő adatkezelési szokásrendszerei hogyan viszonyulnak a state-of-the-art adatkezelési lehetőségekhez? E felmerülő kérdésekre más vizsgálatok adhatnak pontos válaszokat, itt és most csak egyetlen kapcsolódó aspektusra kívánjuk felhívni a figyelmet, nevezetesen a kutató szerepére az adatfeldolgozási folyamatban.

A hivatkozott *workshop* résztvevői annak a gyakorlati tapasztalatuknak adtak hangot, hogy a képviselt diszciplínák (leginkább társadalom- és bölcsészettudományok) legtöbb kutatása esetében a kutató élő, közvetlen és napi kapcsolata a tudományos adatfeldolgozás különféle fázisaihoz elengedhetetlen. A feldolgozási folyamatban az adatkurátori munkákat ugyanis csak diszciplináris tudással bíró kutató tudja szakszerűen, a diszciplína általános és az adott kutatási projekt specifikus elvárt követelményei és céljai ismeretében elvégezni. Minőségi tudományos adatok előállítása és kezelése tehát a kutató közvetlen hatása és munkája nélkül elképzelhetetlen.

Amennyiben a kutató napi tevékenységének szignifikáns, netalántán túlnyomó részét az adatkurátori és persze az ehhez társuló klasszikus adatfeldolgozási (gyakorlati) munka alkotja (ez ma már nem csupán a memóriaintézményekben, de a kutatóhelyeken is megjelenő jelenség), akkor a kutatói lét megalapozását és általános teljesítménymérését jelenleg kvázi egyetlen paraméterben mérő publikációs tevékenység csorbát szenvedhet. Leginkább azért, mert a (magyar) tudományosságban nem alakult ki a tudományos adatközlés és adatpublikálás, tudományon belüli, jutalmazási és elismerési rendszere. A minőségi tudományos adatok publikálása mint olyan nem érvényesíthető tudományos teljesítményként, sem egyénileg, sem pedig intézményi szinten. Az adatvezérelt tudományok jelenlegi felívelő korszakában ez nyilvánvalóan felülvizsgálandó kérdés, melyet a végletekig feszít majd az újabb keletű publikációs formák (nanopublikáció [URL10], kapcsolt adatok [URL12] publikálása) elterjedése.

A korrekt adatpublikálás az adat-újrafelhasználás előfeltétele. Az adatpublikálás helyes végrehajtásához a publikálásra kerülő adatállományokhoz részletes eredetinformációk társítása (provenance metaadatok) szükséges. E metaadatok előállítása de facto multi- és interdiszciplináris szaktudást követel meg. Az adateredet és -feldolgozás történetiségének metaadatolása ugyanis, a forrástól a végfelhasználásig nem csupán a mérő és regisztráló eszközök működésének és beállított paramétereinek ismeretét (műszaki ismeretek) és pontos regisztrálását,

az adatfeldolgozási lépések és (informatikai) módszerek feltárását és rögzítését, de a korrekt (könyvtárosi, archiválási) metaadatolást, osztályozást stb. is igényli, vagyis egy interdiszciplináris szaktudást, melyben a diszciplináris tudás és a fenti szakterületeken történő jártasság együttes jelenléte elengedhetetlen.

Az adatfeldolgozási folyamat legtöbbször mint önálló rész nem választható el mechanikusan a kutatási *workflow*-tól. Vannak olyan diszciplínák (például digitális bölcsészet) ahol az adatfeldolgozási workflow informatikai és szaktudományi ismeretrendszere annyira összeolvad, hogy szétválasztásuk nem is lehetséges, vagyis a kutatónak egy személyben kell ismernie az informatikai és a szak- (példánk esetében bölcsészet) tudományi ismereteket lényegében teljes mélységben. A kutatási és az adatfeldolgozási workflow egyetlen integráns egységet képez.

Mindkét esetben (adatpublikációt előállító kutató és az integrált kutatási-adatfeldolgozás workflow alapján dolgozó kutató esetében) a publikálás jelenlegi nehézségei elrettenthetik a kutatót e területektől, mely visszavetheti a tudomány fejlődését ezen a ma még marginálisnak látszó, de a fősodor felé navigáló területen.

KAPCSOLT ADATOK

A géppel értelmezhető és automatikusan feldolgozható adatok világméretű hálózatának, a szemantikus web gondolatának mint víziónak korai felvetése (Berners-Lee, 1998), a kialakított szemantikus web-architektúra és technológiai szabványrendszer rendkívüli komplexitása nem tette vonzóvá az idea gyakorlati megvalósulását, és mindeddig nem vált a fősodor részévé. Az utóbbi évek leglátványosabb fejlődése azonban a kapcsolt adatok (linked data) kvázi mint 'a szemantikus web' megjelenése és rapid elterjedése (Bizer et al., 2009).

Kapcsolt adatok esetén explicit ábrázolásra kerül (URL11) két adatentitás közötti reláció, mely valamilyen tudásközösségben közmegegyezéssel létrehozott, ontológiákkal pontosan meghatározott viszony. A tudásábrázolás ilyen atomi szintű megközelítése teszi lehetővé azt, hogy a puszta adatok kezelése és/vagy publikálása az adott tudásközösségen, diszciplínán túlmenő adatfelhasználás esetén is szemantikusan helyesen történhessen meg, egyben a technológia egyszerűsége a gyors és széles körű elterjedés alapfeltétele. Ennek a folyamatnak vagyunk manapság tanúi.

Míg korábban egy adatállomány publikálásánál az adatállományhoz társított származási (provenance) adatok hordozták leginkább az adatok értelmezési keretét és kontextusát, mely legtöbbször csak az adott diszciplína, tudásközösség kutatói/tagjai számára volt pontosan értelmezhető, addig a kapcsolt adatok publikálása során minden egyes adatrészecske szemantikája pontosan megadható nyilvánosan elérhetővé tett ontológiák segítségével. Tehát egy alacsonyabb szintű granuláció és explicit tudásábrázolás váltja fel a korábbi nagy léptékű tudás-

granulációt és implicit, hallgatólagos tudáskövetelményt. A közösségi, közmegegyezett tudás (ontológia) explicit ábrázolása és hálózati publikálása, valamint a közösség kutatási adatainak ugyanilyen módú, nyílt (például Linked Open Data) nyilvánosságra hozatala, valamint e kettő szerves kapcsolatának hosszú távú, mechanikus fenntartása biztosíték arra, hogy a közösség tudása egyrészt beépülhet a tudomány egészébe, másrészt úgy épülhet be (szemantikusan interpretálva), ahogy azt a közösségi tudáslétrehozás/felhalmozás során a hozzáértők feltárták.

A kutatási adatok interoperábilis, multi- és interdiszciplináris felhasználásának az alapfeltétele, a helyes értelmezés lehetőségének univerzalitása valósul meg ezáltal, mely új utakat nyit meg az adatok idegen diszciplínákban történő helyes felhasználására, egyben egy tágabb kontextusban a korábban inkább elveszni látszó univerzális globális tudásközösség újrafelépíthetőségét alapozhatja meg.

Manapság a tudástárolás/tudásmegosztás felhőalapú technológiái terjednek. A korábbi adatrepozitórium, adatsiló megközelítés intézményi szinten ugyan megmarad, de egy olyan informatikai adatmegosztó rendszer-réteg mögé kerül elrejtésre, mely a silókban tárolt adatállományokból közvetlenül konvertál kapcsolt adatokat, és teszi azt elérhetővé az interneten például SPARQL (URL13) nyelvű keresőfelület segítségével. A SPARQL-kereső működésének folyamatos fenntartása révén a külső adatfelhasználók számára mindez úgy jelentkezik, hogy a siló adatállománya állandóan rendelkezésére áll egy alacsony granulációs szinten, mintegy virtuális adatfelhőt létrehozva az interneten. Az adatfelhő adatain, akár következtető (szoftver) gépek segítségével, bonyolult (logikai) adatfeldolgozások hajthatók végre, új felismeréseket, új adatfelhasználási eseteket hozva létre.

Az adatfelhő az elektronikus tudományművelés egyik fontos infrastrukturális alapeleme, melyhez társítva az adatok feldolgozását, valamint a kutatók mindennapos tevékenységét, kommunikációját, kollaborációját stb. támogató szoftvereszközöket és hálózati szolgálatokat, feltehetően a tudomány elektronikus művelésének 21. századi új dinamikáját hozza el.

HOSSZÚ TÁVÚ DIGITÁLIS MEGŐRZÉS

A digitális adatok hosszú távú megőrzése korunk egyik égető kérdése. A hosszú távú megőrzés egyrészről az adatállományok (és persze adathordozók) fizikai megőrzését, másrészről a digitális objektumok eredeti szemantikájának, az adatok korrekt értelmezhetőségének a hosszú távú (100+ évre vonatkozó) megőrzését jelenti. A probléma forrása itt is az informatika gyors fejlődése, ahogy azt már korábban láttuk, ez a web világában is együtt járt a technológiai fejlődés követésének problémájával.

Ez esetben a digitális állományok formátumának gyors avulásával kell megküzdeni. A korábbi szoftververziók által létrehozott fájlformátumok nem használhatóak hosszú távon. Egy idő után a korábbi fájlformátumok fenntartása nem válik lehetségessé vagy kívánatossá. A (szoftver-) rendszerek fejlődése ugyanis magával hozza azt is, hogy a régi adatformátumok már nem adnak elegendő lehetőséget az új, összetettebb értelmezési keretek ábrázolására, ezért új, gazdagabb adatformátumokat definiálnak, és kezdenek el használni szélesebb körben. A mérő- és regisztráló eszközök technológiai fejlődése, szofisztikáltságának fokozódása is ez irányba mutat. A régi fájlformátumok avulása az adatállományok elvesztésének rémével fenyeget. Ha az adatállomány tulajdonosai és/vagy felhasználói nem lépnek időben, akkor egy idő után, még ha az adatállomány fizikailag rendelkezésre áll is, megfelelő szoftver hiányában az nem vagy csak korlátozottan lesz értelmezhető/felhasználható.

A hosszú távú digitális megőrzés egy, a gyakorlatban használható megoldását, a megőrzés rendszer- és tevékenységmodelljének kidolgozását és szabványosítását tűzte ki célul az OAIS, később ISO szabvány (URL14). A megőrzési folyamat lényeges részei a következők. Figyelni szükséges (obszervatórium-modell segítségével) az adott közösségben alkalmazott fájlformátumok aktuális használatát, és amint egy-egy adott fájlformátum használata kezd leáldozni a közösségben (vagy akár globálisan) akkor az archívumokban, az abban a formátumban tárolt fájlokat vagy transzformálni (migrálni) kell az újabb keletű fájlformátumokba, vagy pedig a formátumokat helyesen értelmező szoftverrendszerek túlélését kell valamilyen, például emulációs technikával biztosítani. Bármelyik módozatot is választjuk, az aktív, legtöbbször élőmunkával társuló beavatkozást igényel, és mint ilyen, jelentősen erőforrás-igényes. A fájlformátum-transzformációk elvégzése vagy az emulátorok programozása egy-egy nagyobb adatrepozitórium, adatsiló esetében ráadásul jelentős időt is vehet igénybe, és mint ilyet, előzetesen tervezni és finanszírozni szükséges.

A hosszú távú digitális megőrzés, beleértve a fizikai megőrzés folyamatát is, erőforrás-igényessége miatt, szokás szerint, a felhalmozott digitális állományok szűrésével, válogatásával, selejtezésével csökkenti a megőrzésre kerülő állományok számát, méretét. Az adatselejtezés felelősségteljes tevékenysége ugyancsak nem lehetséges (inter)diszciplináris tudás nélkül.

A digitálisan keletkezett tudományos adatok hosszú távú megőrzésének problémafelvetése hazánkban még csak most kezdődött el, csak néhány korai kezdeményezésről, projektről van tudomásunk (lásd az Országos Levéltár, az MTA SZTAKI működő, hosszú távú tárolói, az Országos Széchenyi Könyvtár rekonstrukciós projektjének céljai). Így a jelen hazai helyzet a meglévő digitális kutatási adatok jelentős mennyisége elvesztésének rémével fenyeget, akár már középtávon (tízéves távlatban) is.

ZÁRSZÓ

Európai kontextusban ezenfelül olyan kérdésekkel kellene foglalkoznunk, mint az Open Science és/vagy a Science 2.0 adatkezelési trendjei, az RDA (Research Data Allience) és tevékenysége egy globális adatinfrastruktúra felé, a nyílt adatinfrastruktúrák létrehozásának, az adatmegosztáson túlmenően a kutatási workflow megosztásának kérdései, a kutatási adatok metaadatolásának részletei, az adathivatkozások módozatai és szabványos megoldási javaslatai, a kutatási adatok becsomagolásának módszertana, a kutatási objektumok létrehozásának, kutatási kontextus felismerési/tárolási képességének, a kutatási objektumok felhasználásának módozatai, az adatkarbantartás automatikus lehetőségei és újabb technológiái, vagy akár a kutatási adatok nyílt hozzáférésének hatása a tudomány művelésének egészére.

E rövid cikk azonban csak e korábbi, hiánypótló hazai műhelykonferencia hiányzó beszámolójának egyfajta utólagos pótlására vállalkozhatott.

IRODALOM

Berners-Lee, T. (1998): *Semantic Web Road Map.* September, https://www.w3.org/DesignIssues/Semantic.html

Bizer, Ch. – Heath, T. – Berners-Lee, T. (2009): Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5, 3, 1–22. DOI – 10.4018/jswis.2009081901, http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Holl A. (2015): Kutatási adatok kezelésének nemzetközi trendjei. *Tudományos és Műszaki Tájékoztatás*, 62, 5, 177–180. http://real.mtak.hu/24531/1/201505Holl_cikk_TMT.pdf

Holl A. (2016): Tudományos kommunikáció a XXI. században – Open Science. *Magyar Tudomány*, 177, 3, 307–316. http://www.matud.iif.hu/2016/03/08.htm

Kovács L. – Gárdos J. – Holl A. (2015): *Kutatási adatok kezelése az MTA intézményeiben.* Memorandum. Verzió: 0.76, 2015. június 9. Kézirat

Micsik A. – Gárdos J. (2014): Tudományos repozitóriumok az MTA-ban: a KDK és a SZTAKI tanulságai. In: *Informatika a felsőoktatásban 2014.* Debreceni Egyetem Informatikai Kar, http://real.mtak.hu/25200/1/if2014micsikgardoskdk.pdf

MTA BTK – MTA SZTAKI – DE – ME (2016): *Nemzeti digitális bölcsészeti kiválósági központ,* GINOP 2.3.3-15-2 pályázat. Kézirat

Wilkinson, M. D. – Dumontier, M. et al. (2016): The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, Article number: 160018, DOI:10.1038/sdata.2016.18, https://www.nature.com/articles/sdata201618

URL1: Repository of the Acedemy's Library http://real.mtak.hu
URL2: MTA KRTK Adatbank http://adatbank.krtk.mta.hu/nyito
URL3: MTA TK KDK Repository http://openarchive.tk.mta.hu
URL4: ORCID https://orcid.org
URL5: DOI https://www.doi.org
URL6: Magyar Tudományos Művek Tára https://www.mtmt.hu

URL7: Repozitóriumi Közös Kereső http://oaikereso.sztaki.hu/kereso/index.php?type=0
URL8: Repozitóriumminősítő Szakbizottság https://www.mtmt.hu/repozitoriumminosito-szakbizottsag
URL9: DARIAH – Digital Research Infrastructure for the Arts and Humanities http://www.dariah.eu
URL10: Nanopub.org http://nanopub.org/wordpress/
URL11: RDF Resource Description Framework https://www.w3.org/RDF/
URL12: Linked Data https://www.w3.org/standards/semanticweb/data
URL13: SPARQL Query Language for RDF https://www.w3.org/TR/rdf-sparql-query/
URL14: CCSDS – OAIS model (2012) https://public.ccsds.org/pubs/650x0m2.pdf
URL15: OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting https://www.openarchives.org/pmh/