A NYÍLT KUTATÁSI ADATOK KEZELÉSÉNEK HÁROM OLDALA

Holl András
MTA Könyvtár és Információs Központ
holl.andras@konyvtar.mta.hu
ORCID: 0000-0002-6873-3425

DOI: 10.31915/NWS.2018.8

Open Research Data is a component of Open Science. It is among widely discussed topics in Europe, and attention should be paid to it in Hungary too. Research data could only be made FAIR (Findable, Accessible, Interoperable and Reusable) with the co-operation of three parties: the researchers, the IT and the Library.
Keywords: research data, data stewards, Open Science

1. Nyílt hozzáférésű kutatási adatok

Egyre gyorsuló ütemben bővül a tudomány: egyre több tudományos közlemény jelenik meg, egyre nagyobb mennyiségben keletkeznek kutatási adatok. Elterjedt a vélemény, miszerint csak a tudomány nyitottságának növelése segíthet abban, hogy ne fulladjunk bele a tengernyi információba, fenn tudjuk tartani – esetleg gyorsítsuk is – a fejlődés ütemét és gazdaságosabbá tudjuk tenni a kutatási folyamatot. A nyílt hozzáférésű tudomány – az Open Science – gyűjtőfogalom. Ide tartozik a nyílt hozzáférés – Open Access – valamint a nyílt kutatási adatok – Open Data – is. Míg az előbbi általánosan ismert, alkalmazása folyamatosan terjed, az utóbbi Magyarországon jobbára ismeretlen, de világszerte is kihívást jelent.

Kutatási adatok alatt mindazokat a megfigyelési, kísérleti, felmérési, modellezési, adatbányászattal vagy archívumban való kereséssel gyűjtött adatokat, dokumentumokat értjük, amelyeket kutatási projektek során készítenek, amelyeket a vizsgálatokhoz, elemzésekhez, a következtetések levonásához felhasználnak, amelyekre alapozva közleményeket publikálnak.

A kutatási eredmények reprodukálásához, ellenőrzéséhez szükség van az adatok hozzáférhetőségének biztosítására (de hasonlóképpen szükség van a kutatási folyamat további összetevőinek nyilvánosságára, mint például az adatok tisztításához, elemzéséhez használt számítógépes programokra). Ugyanakkor a kutatási adatok esetenként újra felhasználhatóak is lehetnek: az adatgyűjtés eredeti céljától eltérő tudományos kérdések megválaszolását is segíthetik jelentős költséggel elvégezhető új mérések, adatgyűjtés nélkül.

Míg a nyílt hozzáférés tárgyai – a tudományos közlemények – a meglévő „műfaji" különbségek ellenére viszonylag hasonló módon kezelhetőek, a kutatási adatok sokkal inkább eltérőek. Mennyiségük, formátumuk, a közreadásukkal járó esetleges etikai kockázatok és megannyi más tulajdonságuk különböző. Lehetnek akadályok a közleményekhez való nyílt hozzáférés előtt is, ám a kutatási adatok

esetében nehezebb a nyílt hozzáférést megvalósítani. Az általános alapelv az, hogy legyenek annyira hozzáférhetőek, amennyire csak lehetséges, ám a nyílt hozzáférés legyen korlátozható, amennyiben erre alapos ok van (as open as possible, as closed as necessary). A tudományos közleményekhez hasonlóan a kutatási adatok esetében is alkalmazható az embargó: megadott ideig csak az adatok létrehozója fér hozzájuk, csak az embargó lejártával válnak nyilvánossá. Esetenként alkalmazandó a személyes adatok védelmére szolgáló anonimizálás, indokolt esetben az adatok korlátlan ideig zárolhatóak.
Kívánatos, hogy a kutatási adatok kezelése a FAIR alapelveknek megfelelően történjen. A FORCE11 szervezet által megfogalmazott kritériumok szerint a kutatási adatok legyenek megtalálhatóak, hozzáférhetőek, szabványosak és újrafelhasználhatóak (Findable, Accessible, Interoperable, Re-usable)[1].

Mit is biztosít a nyílt kutatási adatok elvének alkalmazása, ha hozzáférést nem (nem feltétlenül)? Átláthatóságot, jó adatkezelési gyakorlatot. A kutatási pályázatok kiírói Nyugat-Európában gyakorta megkövetelik az adatkezelési tervek – Data Management Plan – benyújtását. Ha az adatok esetleg nem is lesznek – rögtön vagy soha – nyilvánosak, legalább pontosan lehet tudni, milyen adatok keletkeznek a kutatási program során, hogy kezelik, archiválják azokat, milyen hozzáférési szabályokat alkalmaznak. Az adatkezelési tervnek az ajánlások szerint ki kell térnie az adatok leírására, a dokumentációra és a minőség-ellenőrzésre, a tárolás és mentés gyakorlatára, a felmerülő jogi és etikai kérdésekre, valamint a megosztásra és a hosszú távú megőrzésre vonatkozó elképzelésekre. Előnyös, ha az adatkezelési terv a pályázat bírálói számára is látható. A Science Europe újabban adatkezelési tervek helyett javasolja a tudományterületi adatkezelési protokollok – Data Domain Protocol – alkalmazását mindazokban az esetekben, amikor a kutatási projektben az adott területen megszokott, szabványos eljárásokat alkalmazzák[2]. Ezekben az esetekben nem lenne szükség egyedi adatkezelési terv benyújtására, elegendő lenne az elfogadott szakterületi protokollra hivatkozni.

Mindaddig nehéz lesz a kutatókkal elfogadtatni, hogy elengedhetetlen az adatok megfelelő kezelése és közzététele, amíg ez pusztán kényszer, de számukra előnyt nem jelent. A tudományos közlemények publikálása és a közleményekre kapott hivatkozások jelentik a tudományos előmenetel alapját. Csak úgy lehet a publikus adatokat újra felhasználni, ha megfelelően hivatkoznak rájuk – a hivatkozások és azok nyilvántartásának technikai alapjait a DOI[3] azonosítók használata jelenti. Amennyiben a nyilvánosságra hozott adatokra kapott hivatkozások is segíteni fogják a tudományos karriert, az megfelelő ösztönzést jelent majd a kutatók számára. Immár az adathivatkozások nyilvántartásának eszközei is rendelkezésre

1 FAIR data principles https://www.force11.org/group/fairgroup/fairprinciples
2 Presenting a Framework for Discipline-specific Research Data Management http://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf
3 Digital Object Identifier http://doi.org

állnak: ilyen a Clarivate Analytics Data Citation Index-e, de a Magyar Tudományos Művek Tárában is lehet publikált kutatási adatokat rögzíteni és az ezekre kapott hivatkozásokat is képes a rendszer nyilvántartani. (Az MTA Könyvtár és Információs Központ pedig díjmentesen tud a DataCite szervezeten keresztül DOI-azonosítókat biztosítani a hazai kutatóintézetek számára.)

Mivel a kutatási adatok kezelése erős technológiai hátteret igényel, a szükséges szabványosítás előmozdítására, az érdekeltek tevékenységének összehangolására létrejött a Research Data Alliance[4]. Az Európai Unió a tagállamok kutatásiadat-infrastruktúráit a European Open Science Cloud kezdeményezéssel kívánja összefogni[5].

2. Kutatási adatok kezelése – a három oldal

Igen összetett, kihívást jelentő feladat a kutatási adatok kezelésének megszervezése, csak háromoldalú, a kutatókat, informatikusokat és a könyvtárosokat bevonó együttműködéssel valósítható meg. A kutatók számára a feladat többletterhelést jelent, mindamellett szükségük lehet informatikai, személyes adatvédelmi, etikai vagy éppen archiválási szakértelemre is.

2.1 Kutatók

A kutatási adatok még egy szűk tudományterületen belül is nagyon sokfélék lehetnek – leírásuk, tárolásuk, kezelésük, felhasználási lehetőségeik különbözhetnek. Csak a kutatási folyamatban résztvevőknek lehet esélyük egyes adatvédelmi, etikai kockázatok felismerésére – még ha kezelésükhöz szakértői segítségre is szükségük lehet. Leginkább a szakterületet ismerő kutatók képesek az adatok újrafelhasználásának esélyeit felmérni. Csak akkor van értelme az adatokat archiválni és elérhetővé tenni, amennyiben a publikációkhoz hasonlóan szakmai bíráláton mennek keresztül. Valószínűleg az a legszerencsésebb, ha a bírálatot a közleményt bíráló kutatók végzik az adatok esetében is. A kutatók legtöbbször az adatok tulajdonosainak érzik magukat – még abban az esetben is, ha nyilvános adatbázisból merítettek vagy a megfigyelésekre rendelkezésükre bocsájtott nagyberendezést üzemeltető szervezet a mérési adatok valódi tulajdonosa. Természetes, hogy a kutató részt vegyen a „saját" adatai utóéletének megtervezésében. Lehetetlenség a kutatást végzők nélkül végezni az archiválást és közzétételt: mivel az archiválási és közzétételi szempontoknak már az adatgyűjtés és adatfeldolgozás során meg kell felelni.

4 RDA https://www.rd-alliance.org/

5 Mons, Barend, Neylon, Cameron, Velterop, Jane, Dumontier, Michel, da Silva Santos, Luiz, Olavo Bonino, Wilkinson, Mark D. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, 37. 1. sz. (2017), 49-56 https://doi.org/10.3233/ISU-170824

2.2 Informatikusok

Adatintenzív projektek esetében jelentős informatikai kihívással szembesülhetnek a kutatók. Esetenként igen nagy mennyiségű adat tárolását és továbbítását kell megoldani, amit a kutatók saját projektjük keretében nem tudnak biztosítani (a keretek alatt nem csak a költségeket, de az időt is értve – a kutató nehezen tud a projekt lezárta utánra tervezni). A technikai igények szükségessé tehetik az informatikusok bevonását a projekt tervezési fázisában. Esetenként az informatikusok tudnak tájékoztatást adni az adatok tárolására leginkább alkalmas szabványos formátumokról. A formátumok pontos ismerete nélkül nem oldható meg az archivált adatok validálása és szükség szerinti migrálása. A formátumok avulása miatt szükségessé váló adatmigrációt is az informatikusok tudják észlelni és végrehajtani. Bonyolultabb adatok kezelése – akár megtekintése is – speciális szoftvereket igényelhet. Mindezeken túl az adatkezelés nem választható el az adatok létrehozásához, feldolgozásához használt szoftverek hosszú távú biztosításától sem.

Sok esetben szükség van az adatkezelés során különböző informatikai infrastruktúrák – szuperszámítógépek, grid, felhő, valamint azonosítási és jogosultságkezelési rendszerek – használatára. Igen nagy mennyiségű adat használata során felmerülhet az a kérdés is, vajon az adatokat célszerű-e az elemzés helyére eljuttatni vagy az elemzéshez használt kódot az adatokhoz?

2.3 Könyvtárosok

Mint memória-intézmény, a könyvtár a megfelelő hely a hosszú távú megőrzés biztosítására (az adatok fizikai elhelyezése történhet az intézményi adatközpontban vagy éppen valamilyen felhőszolgáltatásban). A könyvtárosoknak van gyakorlata a metaadatok használatában (még ha a szakterület-specifikus leírási követelményeket a kutatóknak is kell megadniuk). Fontos szempont a publikációs kapcsolatok gondozása. Az adatok legjobb leírását a felhasználásukkal készült szakcikkek adhatják. Ezeknek a cikkeknek az azonosítóit az adatállományok leíró adatai között is el kell helyezni. Az adatok újrafelhasználása esetén a másodlagos felhasználás azonosítóival is bővíteni kell a metaadatokat. Éppilyen fontos, hogy a publikációban is megfelelő hivatkozások legyenek a felhasznált adatokra. A publikációk, adatállományok és szerzők egyedi azonosítóinak (DOI, ORCID) ismerete – az előbbieknél adminisztrálása is – könyvtárosi kompetencia, ugyanúgy, mint a kutatási adatok tudománymetriai nyilvántartása.

Nem utolsó sorban a kutatási adatok kezelésében való részvétel újabb lehetőséget teremt a könyvtárak és a kutatócsoportok kapcsolatának megerősítésére, a kutatási folyamatot támogató, a kutatást végző szervezeti egységekbe beágyazott könyvtárosi pozíciók létrehozására.

2.4 Három az egyben: az adatgazdász

Gyakori, hogy a kutatási projektekben résztvevő kutatók nem rendelkeznek az adatkezeléshez szükséges ismeretekkel, de még ha rendelkeznének is, a gondos, korábban meg nem követelt dokumentáció és archiválás munkaigényes, újabb résztvevők bevonását teszi szükségessé.

Már a projekt előkészítése folyamán – az adatkezelési terv kialakításakor –, de a projekt lezárását követően is új teendők keletkeznek. Az adatkezeléssel foglalkozó szakemberek, a kutatási adatok kezelésére specializálódott könyvtárosok, informatikusok vagy „kiugrott" kutatók lehetnek: a data steward-ok (vagy Bereczky Áron magyarításával adatgazdászok)[6]. A becslések szerint a European Open Science Cloud működtetéséhez hosszú távon félmillió adatgazdászra is szükség lehet[7]. Mint már említettük, a könyvtár megőrző szerepe indokolja, hogy az adatgazdászok a könyvtárak kötelékében működjenek.

Még ha nem is fogadjuk el a fenti becslést, még ha a hazai részesedést pesszimistán (vagy reálisan) is ítéljük meg, akkor is számos hazai szakember képzéséről és alkalmazásáról kell gondoskodni, amihez képzési programokat is kell szervezni és akkreditáltatni.

6 Barend Mons, Data Stewardship for Open Science: Implementing FAIR principles. New York: Chapman and Hall/CRC, 2018. https://doi.org/10.1201/9781315380711

7 Barend Mons becslése http://e-irg.eu/news-blog/-/blogs/we-need-500-000-respected-data-stewards-to-operate-the-european-open-science-cloud