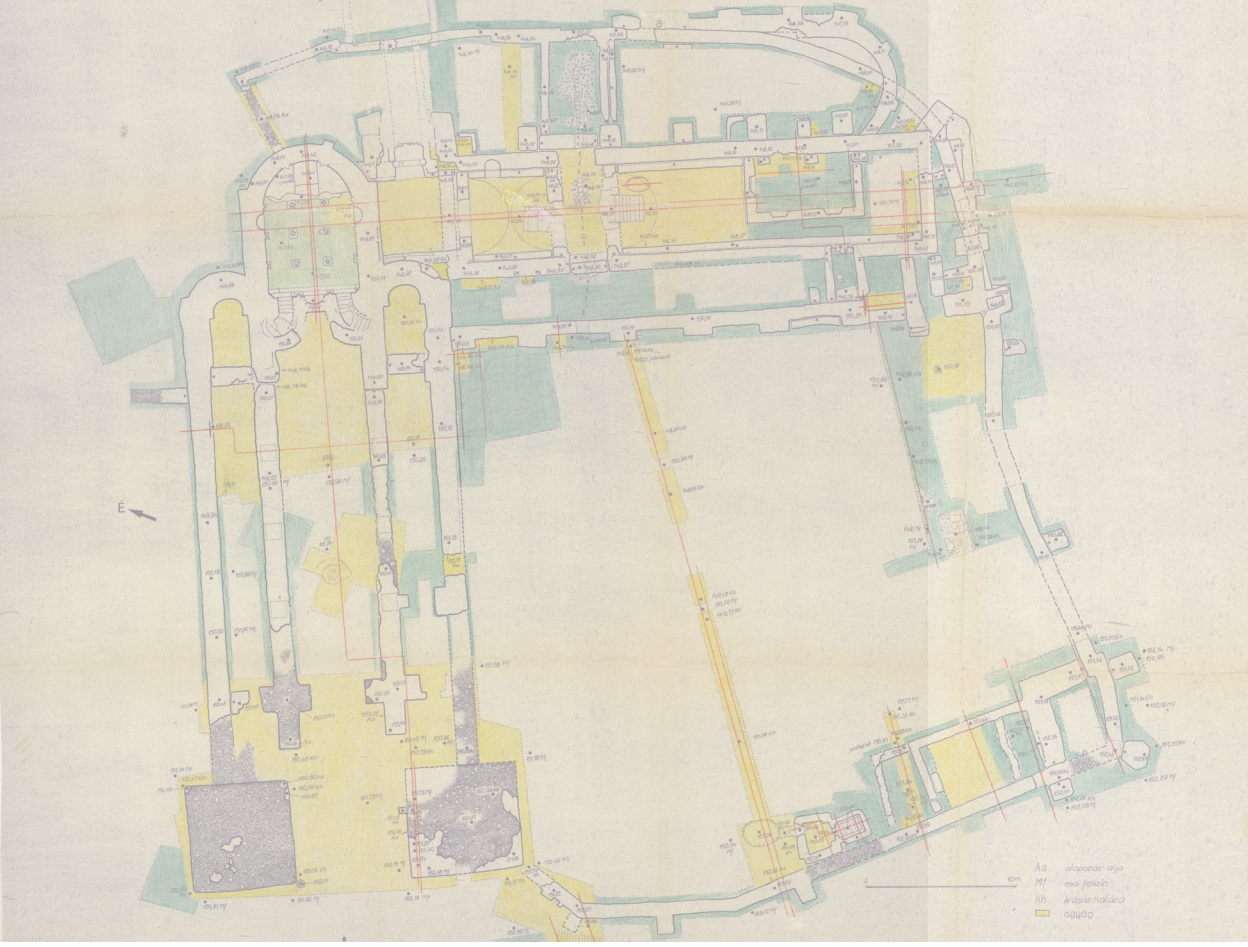


A KUTATÁSIADAT-KEZELÉS GYAKORLATA

BEVETT ELJÁRÁSOK ÉS KÍSÉRLETI PROJEKTEK

47 (122)
ÚJ SOROZAT

A MAGYAR TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRÁNAK KÖZLEMÉNYEI

MTA K 1826

A kutatásiadat-kezelés gyakorlata:
bevett eljárások és kísérleti projektek

A Magyar Tudományos Akadémia Könyvtárának Közleményei
Publicationes Bibliothecae Academiae Scientiarum Hungaricae

47 (122)
Új sorozat

Sorozatszerkesztő
Gaálné Kalydy Dóra

A kutatásiadat-kezelés gyakorlata:

bevett eljárások és kísérleti projektek

Magyar Tudományos Akadémia
Könyvtár és Információs Központ
Budapest 2022

Szerkesztette
Holl András és Maróthy Szilvia

Technikai szerkesztő
Vas Viktória

Korrektor
Egyed-Gergely Júlia

Nyomta és kötötte az Alföldi Nyomda Zrt., Debrecen
Felelős vezető: György Géza vezérigazgató

ISBN 978-963-7451-85-0
ISSN 0133-8862
DOI: 10.36820/MTAKIK.KOZL.2022.KutAdat

A címlapon szereplő illusztráció:
A dömösi prépostság szintezési felmérése a templom és a palota alaprajzával.
(© MTA – Bölcsészettudományi Kutatóközpont Régészeti Intézet Adattára
RK_14655a, készítette: Léh Ernő 1979).
https://hdl.handle.net/21.15109/CONCORDA/6OAIJS, CONCORDA, V1

Tartalomjegyzék

Bevezető. Kutatási adatok kezelése: könyvtári feladat?

A címben szereplő kérdést az amerikai tudományos szakkönyvtárak már több mint egy évtizede megválaszolták. Az Association of Research Libraries felmérése szerint[1] a tagkönyvtárak stratégiákat készítettek a kutatási adatok kezelésére – még ha intézményenként eltérő módon is. Az MTA Könyvtárának Közleményeiben (45.) egy évvel ezelőtt foglalkoztunk már kutatási adatokkal a nyílt tudomány kérdései között. A sorozat 47. kötetében most visszatérünk a témára.

2021-ben az Eötvös Loránd Kutatási Hálózat titkárságának megbízásából, a Research Data Alliance magyar csoportja szakmai támogatásával az MTA Könyvtár és Információs Központ kísérleti programot bonyolított le kutatási adatok kezelésének témájában. Kötetünk első részében a projekt összefoglalóját, és négy támogatott projekt beszámolóját olvashatjuk.

Az adatok elhelyezése történhet repozitóriumokban, szakterületi adatközpontokban, de a tudományos szakfolyóiratok is gyakorta lehetőséget adnak a cikkek mellett adatok elhelyezésére a digitális felületen. Az Akadémiai Kiadó adatmelléklet-kezelési gyakorlatát ismerteti a második rész első cikke. Ezt két szakterület: a bölcsészettudomány és a csillagászat adatkezelési gyakorlatának bemutatása követi.

Mind az adatkezelési kísérleti program lebonyolításával, mind e kötet kiadásával a szakkönyvtárak ez irányú elkötelezettségét igyekszünk alátámasztani.

Holl András

1 E-Science and Data Support Services: A Study of ARL Member Institutions. Soehner, Catherine; Steeves, Catherine; Ward, Jennifer. ARL, 2010. https://eric.ed.gov/?id=ED528643

Tanulságok az ELKH-HRDA adatrepozitórium pilot projektek végrehajtása alapján

Holl András

MTA Könyvtár és Információs Központ
ORCID: 0000-0002-6873-3425

Az Eötvös Loránd Kutatási Hálózat Titkársága a kutatási adatok kezelésére való felkészülés támogatására egyéves programot indított. A projekt szakmai támogatását a Research Data Alliance magyar tagozatának segítségével biztosította, a projekt technikai lebonyolítását és a költségvetés kezelését az MTA Könyvtár és Információs Központ végezte 2021-ben.

E program súlyponti részeként kutatásiadat-kezelési pilot projektek támogatására szolgáló pályázati kiírás jelent meg, melyre nyolc pályázatot adtak be az arra jogosult intézmények – az ELKH kutatóközpontjai, intézetei és kutatócsoportjai. Egy kivétellel minden pályázat támogatására mód nyílt, ugyan egy esetben csak csökkentett költségvetéssel.

A projektek egy részének eredményeiről készült beszámolók megtalálhatóak e kötetben.

Hasonló pilot projektek támogatására nem csupán itthon, de külföldön sem ismerünk példát. A nemzetközi – és immár hazai – kutatási pályázatokban mára már követelménnyé vált az adatkezelési tervek készítése, s így a kutatásiadat-kezelés költségei is elszámolhatóak. Kifejezetten kutatásiadat-kezelés megvalósítására azonban nem szoktak pályázatokat kiírni. A hazai kutatói társadalomnak csupán kis része – a legutolsó európai pályázati fordulókban támogatást nyert projektek, vagy a nemzetközi együttműködésben folytatott nagyprojektek résztvevői – találkozhatott a kutatásiadat-kezelés mára elfogadott FAIR

kritériumrendszerével. Az ELKH intézmények által megvalósított pilot projektek egyedülálló lehetőséget biztosítottak az adatkezelési gyakorlat fejlesztésére, a szabványok megismerésére, esetenként a szabványosítási folyamatba való bekapcsolódásra, a korábbi adatkezelési gyakorlatok megújítására.

Projektek

Támogatott projekt	**Intézmény / szervezeti egység**
Sokcsatornás, nagy téri felbontású in vivo elektrofiziológiai adatok archiválása*	Természettudományi Kutatóközpont, Kognitív Idegtudományi és Pszichológiai Intézet, Integratív Idegtudományi Kutatócsoport
A Társadalomtudományi Kutatóközpontban (illetve annak jogelődjében) végzett kutatások veszélyben lévő kutatási adatainak megóvása*	Társadalomtudományi Kutatóközpont, Kutatási Dokumentációs Központ
Régészeti rajzgyűjtemény kutatási adatainak feltárása és FAIR közreadása*	Bölcsészettudományi Kutatóközpont, Régészeti Intézet
Kis dózisoknál megfigyelhető hiperszenzitivitással és indukált sugárrezisztenciával kapcsolatos adatok gyűjtése és közzététele*	Energiatudományi Kutatóközpont, Energia- és Környezetbiztonsági Intézet, Környezetfizikai Laboratórium, Sugárbiofizikai Kutatócsoport
Fúziós kísérleti adatok tárolása és metaadatolása a FAIR elveknek megfelelően	Energiatudományi Kutatóközpont, Fúziós Plazmafizika Laboratórium

* Ezen projektek beszámolói megtalálhatók a kötetben.

Támogatott projekt	Intézmény / szervezeti egység
Az OpenBioMaps biológiai adatbázis keretrendszer publikus adatrepozitórium „láb" fejlesztésére	ELKH–DE Viselkedésökológiai Kutatócsoport
Funkcionális anyagok adatainak archiválása	Wigner Fizikai Kutatóközpont, Részecske- és Magfizikai Intézet, Nukleáris Anyagtudományi Osztály

A pilot projektek látványosan demonstrálták a kutatási adatok és kezelésük diverzitását. Nemhogy tudományterületek és -ágak között, de többnyire ugyanazon témában is alapvetően eltérő adattípusok fordulnak elő, melyek kezelése eltérő megközelítést kíván.

Az Energiatudományi Kutatóközpont Fúziós Plazmafizikai Laboratóriuma által megvalósított projektben a videodiagnosztikai és a nyalábemissziós spektroszkópiai mérésekben keletkező adatok mennyisége és feldolgozása is különböző. E projekt esetében a tárhelyszükséglet nagyságrendekkel haladta meg más projektekét – a tárolás és az adatmozgatás aspektus különbözteti meg a többitől. A nemzetközi szervezetek szabványosítási törekvéseibe való bekapcsolódás hangsúlyos eleme volt a munkának. Ugyancsak e projekt részeként valósult meg publikált cikkek kiegészítése mérési adatokkal.

A Régészeti Intézet projektje archív rajzdokumentáció digitalizálását, leírásának fejlesztését és adatrepozitóriumba helyezését célozta. Ennél a projektnél erőteljesen kidomborodott a történeti aspektus (nem a kutatott korszakokat, hanem a kutatás történetét tekintve): a hagyományos rajztár sok évtizedet felölelő, ugyanakkor folyamatos újrafelhasználási potenciállal bíró anyagainak digitális elérhetőségét és kereshetőségét kellett megalapozni, egyúttal lehetőséget adva a leíró adatok modern szempontok szerint történő gazdagítására, javítására. Ez a pályázat példázta a más hazai adatbázis (az Archeodatabase) szabványos, hierarchikus szószedeteihez való alkalmazkodást.

Az Energiatudományi Kutatóközpont másik, dozimetriai projektje kis sugárdózisoknál megfigyelhető hiperszenzitivitás és indukált sugárrezisztencia modellezéséhez szükséges, a szakirodalomból gyűjtött adatok feldolgozását célozta. Ebben az esetben tehát rögtön megvalósult a korábbi, más kutatócsoportok által mért adatok újrafelhasználhatóvá tétele és újrafelhasználása: az összegyűjtött és közreadott adatok a modell javításán és ellenőrzésén túl további kutatások számára is hozzáférhetővé váltak.

Ismét másik oldalát mutatta meg a kutatásiadat-kezelésnek a TK KDK projektje. Ez esetben egy már régóta működő kutatási adatrepozitórium volt a pályázó, a megvalósított feladat pedig hanganyagok archiválása volt. Kiemelendő a hanganyagok kezelésének szoftveres megoldása, és a társadalomtudományok terén fontos adatvédelem, anonimizálás.

A Természettudományi Kutatóközpont projektje esetében is fontos tényező volt a nemzetközi adatleírási szabványokhoz való alkalmazkodás. Ennél a projektnél merült fel a publikációhoz társuló adatnyilvánossági követelmény is – a megvalósítás idején szembesültek a szerzők egy benyújtott közleményük bírálója kérésével, miszerint a felhasznált adatokat és az elemzésben alkalmazott kódot is tegyék elérhetővé.

A Wigner Fizikai Kutatóközpont kutatási programja keretében három (megjelent vagy elbírálás alatt lévő) közleményhez is elhelyeztek adatokat a Concorda-ban. A beszámolóban megjegyezték, hogy az adatrepozitóriumok a projektekben résztvevő, esetenként különböző intézményekből érkező kutatók közötti kommunikációban is fontos eszközök lehetnek: az adatok már a kísérletek során repozitóriumba kerülhetnek, és az arra jogosultaknak hozzáférhetőek lehetnek.

Újabb facettáját csillantotta meg a kutatásiadat-kezelésnek az MTA–DE Viselkedésökológiai Kutatócsoport projektje. Az OpenBioMaps egy kutatási célú adatbázis-infrastruktúra, melyhez adatrepozitálást elősegítő szoftveres megoldásokat fejlesztettek. Igen fontos a kutatáshoz használt eszközök (beleértve a szoftvereket és adatbázisokat) FAIR archiválást támogató funkciókkal való bővítése. Megfelelő

infrastruktúra nélkül a kutatók nem lesznek képesek a FAIR adatkezelés követelményei miatt megnövelt költség- és munkaigényeknek megfelelni.

Érdemes a program során előtérbe került sokféle követelményt felsorolni (még az ismétlés ódiumát is vállalva):

- nagy adatmennyiségek;
- kis kutatási projektek („Little Science");
- archív anyagok;
- publikációkhoz kapcsolódó adatok;
- hazai szabványos nevezéktanok/szótárak használata;
- kereszthivatkozások hazai adatbázisokra;
- bekapcsolódás a nemzetközi szabványosítási folyamatba;
- média digitalizálási technológiák alkalmazása;
- adatbázisrendszerek kapcsolódásának kialakítása;
- korábbi adatok javítása;
- egyedi azonosítók használata.

Tanulságok

A HRDA tagjai/vezetősége köréből kikerült bíráló bizottság igen jó véleményt alakított ki a pilot eredményeiről. Lényeges eredmény volt, hogy olyan kutatókat és kutatócsoportokat is érzékenyíteni lehetett az adatarchiválás és a FAIR szempontrendszer követelményeivel, akik ezzel korábban nem találkoztak. Az ELKH épülő adatrepozitóriuma számára is lényeges volt a valós kutatói igényekkel való szembesülés, a korai kapcsolatépítés.

A projektbeszámolók alapján kiderül, hogy a pilot eredeti célkitűzésein túl is elért eredményeket:

„értékes információkhoz jutottunk mind az eredmények reprodukálhatóságát illetően, [...] mind pedig a tanulmány eredményeinek megbízhatóságát tekintve"[1]

„egy másik [a pályázatban nem résztvevő] kutatócsoport [...] is megismerkedhetett a magyar adatrepozitóriummal [Concorda]"

„Hadd jegyezzük meg, hogy a kutatási adatok repozitóriumban való elhelyezése a kutatási projektek végrehajtása során, még az eredmények közlése előtt is egy nagyon hasznos eszköz lehet a kutatók kezében. Lehetőségeket nyújt, hogy az egyes adatcsomagokhoz – privát URL-en keresztül – a közreműködő kollégák hozzáférjenek, ami nagymértékben megkönnyítheti a kutatók munkáját, főleg nagyobb adatmennyiségek esetén."

Megállapíthatjuk, hogy

i. a kutatásiadat-kezelés megfelelő méretű és biztonságú tárolóhelyek biztosítását igényli;
ii. nemzetközi publikálás esetén egyre gyakrabban kötelező az archiválás és a hozzáférhetővé tétel;
iii. a megfelelő kutatásiadat-kezelés munkaigényes;
iv. az eredményes adatkezelés feltétele a megfelelő eszközök (szoftverek, protokollok, szolgáltatások) megteremtése;
v. szaktudásra, támogatásra – adatgazdászok alkalmazására – van szükség.

Projektzáró, folytatás

A projektek 2021 decemberében lezárultak, a zárókonferencia 2022. január 18-án volt.[2] A pilot sikerét leginkább az tanúsítja, hogy az ELKH Titkársága folytatásként másfél éves futamidejű Adatrepozitórium Platform (ARP) projektet indított.

1 Az idézetek a projektbeszámolókból származnak.
2 https://openaccess.mtak.hu/event/kutatasiadat-archivalasi-pilot-projektek-az-eotvos-lorand-kutatasi-halozathoz-tartozo-kutatokozpontokban-intezetekben-es-csoportokban/

Idegtudományi kutatási adatok archiválása a CONCORDA magyar adatrepozitóriumba

Fiáth Richárd

ELKH TTK Kognitív Idegtudományi és Pszichológiai Intézet
ORCID: 0000-0001-8732-2691

1. Idegtudományi Kutatások

Szinte triviális kijelentés, hogy az idegtudományi kutatások során keletkező adatok és eredmények rendkívüli értékkel bírnak, ennélfogva ezek hosszú távú, biztonságos megőrzése, illetve nyílt hozzáférésű megosztása a tudományos közösség egyik alapvető fontosságú feladata kell hogy legyen. A központi idegrendszerünk összetett felépítése és működése, illetve az agykutatási adatok létrehozásához használt sokféle különböző kutatási módszer és technológia ezeket az adatokat lényegében egyedivé, pótolhatatlanná teszi.

Mivel módszertani lehetőségek széles tárháza áll rendelkezésünkre, az agykutatáshoz kapcsolódó tudományterületeken számos különféle adattípussal találkozhatunk. A főbb típusok közé tartoznak például a különböző mikroszkópokkal készített anatómiai képek és videók, vagy az agyi elektromos tevékenység vizsgálatára kifejlesztett elektrofiziológiai módszerekkel rögzített adatok, mint például az elektroenkefalogram (EEG). A leggyakoribb kutatási adatfajták közé sorolhatók a különféle képalkotó eljárásokkal – például mágneses rezonancia képalkotással (MRI és funkcionális MRI) vagy pozitronemissziós tomográfiával (PET) – készített többdimenziós képek, a kísérletek során keletkezett hang- és videóanyagok (pl. rágcsálókon végzett viselkedéses vizsgálatokban), a kognitív tesztek és kérdőívek, vagy az összetettebb idegrendszeri folyamatok szimulálására és modellezésére létrehozott számítási modellek is. De ide tartoznak a nyers adatok feldolgozott és elemzett változatai, vagyis a származtatott adatok, továbbá az adatfeldolgozáshoz és elemzéshez készített és használt szoftverek és forráskódok, illetve az adatokat leíró metaadatok is.

Az agykutatásban használt modern, nagy térbeli és időbeli felbontással rendelkező módszereket alkalmazva manapság már nem ritka, hogy egy kísérlet során több száz gigabájt nyers adat keletkezik. Sőt, egy adott kutatási kérdés vizsgálatára elvégzett hosszabb kísérletsorozat, vagy egy több kutatóintézet együttműködésében megvalósított nagyobb volumenű kutatási projekt esetén akár a petabájtos nagyságrendet is elérheti a projekt életútja során létrejött nyers és származtatott kutatási adat összes mennyisége. Bár az utóbbi példák még viszonylag szélsőséges eseteknek számítanak, és egy átlagos, egy kutatócsoport által megvalósított kutatási projekt esetén a kutatási adatok tárolásához szükséges tárterület általában még nem igényel többet néhány terabájtnál, korunk gyors technológiai fejlődésének eredményeként a létrehozott kutatási adatok mérete is rohamosan növekszik. Ezt a nagy adattömeget egy kutatócsoport – sőt néha egy kutatóintézet is – csak ritkán tudja hosszú távon megőrizni és publikussá tenni. A kutatási pályázatokat kiíró és azokat finanszírozó hazai és nemzetközi támogató szervezetek azonban már gyakran megkövetelik, hogy egy nyertes kutatási projekt során keletkező adatok szabadon hozzáférhetőek legyenek mindenki számára. A kutatási adatok egyszerű és hatékony archiválására és megosztására nyújthatnak ideális megoldást a tudományos adatrepozitóriumok.

Jelen tanulmány célja vázlatos képet adni az idegtudományi kutatási adatok archiválásának folyamatáról és a szerző ezen folyamathoz kapcsolódó tapasztalatairól. Ennek érdekében először a kutatócsoportunkban (Integratív Idegtudományi Kutatócsoport, Kognitív Idegtudományi és Pszichológiai Intézet, Eötvös Loránd Kutatási Hálózat Természettudományi Kutatóközpont)[1] alkalmazott invazív elektrofiziológiai módszerekkel kinyert adatokat (agyi elektromos tevékenység) és ezek főbb tulajdonságait szeretném dióhéjban bemutatni. Ezt követően néhány korábbi kutatási eredményünk alapjait adó elektrofiziológiai mérést, illetve ezen adatok archiválásának és megosztásának folyamatát fogom ismertetni, röviden kitérve a tudományterületen elterjedt főbb szabványos adatformátumokra és az adatok tárolására használt jelentősebb nemzetközi adatrepozitóriumokra.

1 http://www.ulbertlab.com/

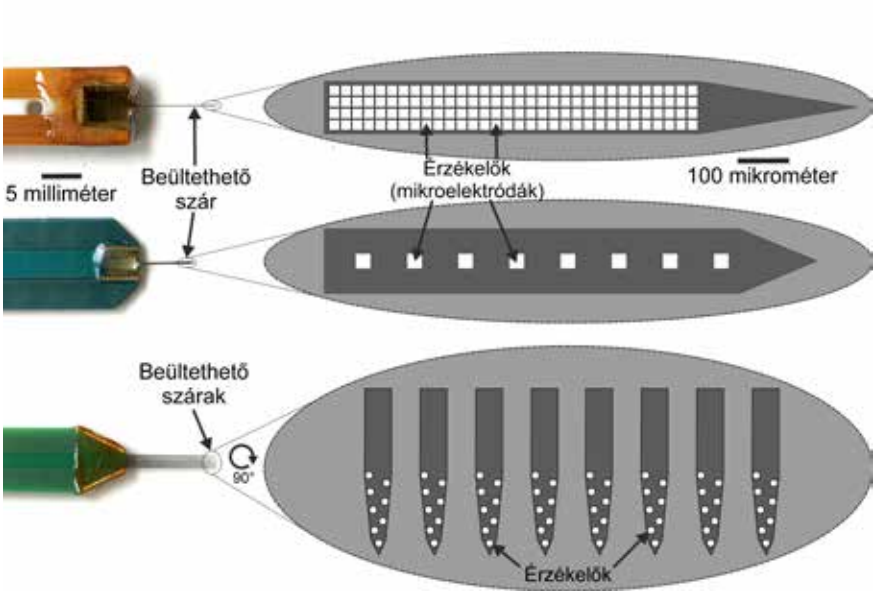
A bemutatott kutatási adatok elhelyezése a közelmúltban kifejlesztett CONCORDA (Concentrated Cooperation on Research Data)[2] magyar tudományos adatrepozitóriumba történt, egy pilot projekt keretein belül, melyet a Hungarian Research Data Alliance, a Magyar Tudományos Akadémia Könyvtár és Információs Központ, valamint az Eötvös Loránd Kutatási Hálózat (ELKH) Titkársága támogatott.

2. Agyszövetbe ültethető implantátumokkal kinyert kutatási adatok és főbb jellemzőik

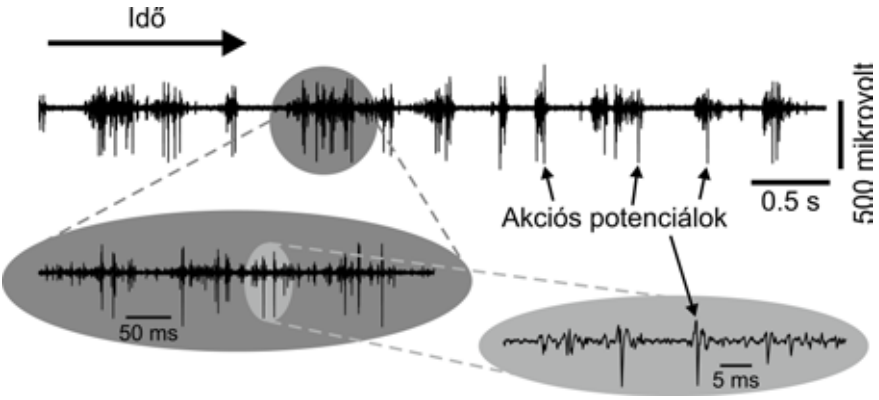
A fő kutatási területem az agyszövetbe ültethető implantátumok, illetve az ilyen típusú eszközökkel rögzíthető agyi elektromos tevékenység vizsgálata, így a továbbiakban ezekre a típusú kutatási adatokra fogok fókuszálni, de néhány esetben kitérek másfajta, idegtudományhoz köthető kutatási adatfajtákra is. A beültethető agyi eszközök alkalmazásának egyik fő célja az eszközön található mikroelektródák (az egyszerűség kedvéért ezeket a továbbiakban érzékelőknek fogom hívni) közelében elhelyezkedő idegsejtek (neuronok) elektromos impulzusainak (szakszóval akciós potenciáljainak) észlelése és rögzítése. Egy ilyen, extracellulárisan (vagyis az idegsejtek közötti térből) mért elektromos impulzus nagyon rövid ideig, nagyjából a másodperc ezredrészéig tart, és ereje (az impulzus alatt mért feszültségváltozás) hozzávetőleg tízezred része egy 1.5 voltos ceruzaelemének. A főként rágcsáló modellben alkalmazott kis méretű agyi implantátumoknak az agyszövetbe ültetett része lényegében egy hajszálvékony (kb. 50 mikrométer vastag), néhány milliméter hosszú tűből (szárból) áll, valamint a tűn található miniatűr (általában kör vagy négyzet alakú) érzékelőkből (1. ábra). Ezekkel az apró érzékelőkkel folyamatosan mérni tudjuk a feszültség változását az agyszövetben, mely feszültséget főként az érzékelők közelében elhelyezkedő idegsejtek tevékenysége befolyásol. Egy adott idegsejt elektromos impulzusa lényegében egy kis tüskeként jelenik meg az extracellulárisan rögzített felvételeken (az angol nyelvű szakirodalomban ezért főként a „spike" kifejezést használják az akciós potenciál szinonimájaként; 2. ábra). Mivel egy-egy érzékelő közelében egyszerre

2 https://science-data.hu/

általában sok neuron aktív (úgy is mondhatjuk, hogy „tüzel"), az elektrofiziológiás felvételeken számos (sokszor egymással átfedő) tüske látható (2. ábra). Egy implantátum általában több érzékelőt is tartalmaz (különféle geometriai elrendezésben és különböző távolságokra egymástól; lásd az 1. ábrát), a száron elhelyezkedő érzékelők tipikus száma 16 és 32 közé tehető, de a tudományterületen az elmúlt években történt technológiai fejlődésnek köszönhetően ez a szám mára már az ezret is átlépte (Raducanu és mtsai, 2017). Sőt, az egyik legmodernebb, nagy elektróda sűrűségű implantátum négy beültethető szárán összességében már több mint ötezer ilyen miniatűr érzékelő található (Steinmetz és mtsa, 2021).



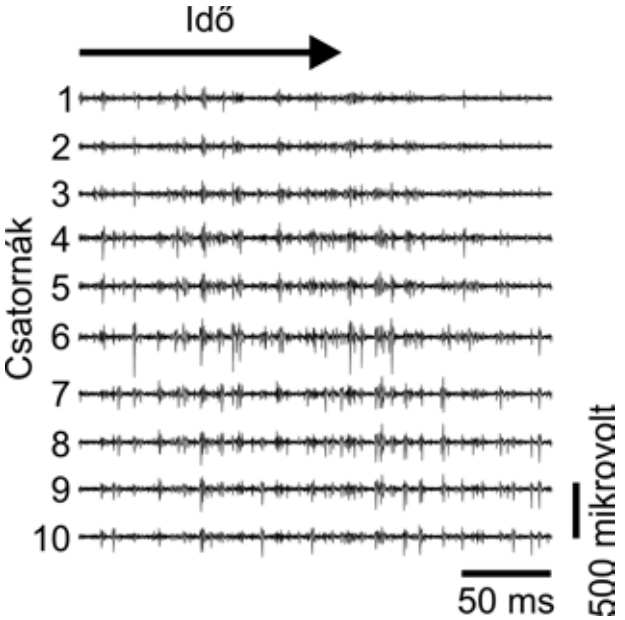
1. ábra: Három különböző, agyszövetbe ültethető implantátum alsó része (balra), valamint az implantátumok szárán található érzékelők (fehér négyzetek és körök) különféle elrendezésben (jobbra). A legalsó implantátum nyolc, egymás mellett elhelyezkedő szárral rendelkezik.



2. ábra: Egy érzékelővel mért, több idegsejttől eredő elektromos impulzusok (akciós potenciálok) öt másodperc hosszú mérésben (fent), és ennek egy rövid szakasza kinagyítva, több időskálán (lent). Az elektromos impulzusok kisebb-nagyobb tüskék formájában jelentkeznek a felvételeken (s – másodperc, ms – millimásodperc).

Egy érzékelő csupán a közvetlen környezetében (maximum nagyjából 50–100 mikrométer távolságra) található neuronok akciós potenciáljait képes észlelni, így az implantátum szárán található különböző érzékelőkkel, amennyiben azok egymástól viszonylag nagy (>100 mikrométer) távolságra helyezkednek el, különböző idegsejtcsoportok (neuronpopulációk) elektromos tevékenységét fogjuk tudni monitorozni. Ez összességében egyszerre akár több száz, egyedileg azonosítható idegsejt aktivitásának mérését jelenti. Az extracelluláris elektrofiziológiára jellemző tipikus mérési adat tehát a következőképpen áll össze a fentiek alapján: az adatunk egy több csatornából álló mérés, ahol minden csatornán egy adott érzékelőtől eredő idősor látható, vagyis a feszültség időbeli ingadozása, változása az agy egy kicsiny, lokális területén (3. ábra). A mérés során a folytonos (analóg) feszültségjeleket speciális elektronikai áramkörökkel előbb felerősítjük, majd digitalizáljuk. Utóbbi esetén, hogy a nagyon rövid időtartamú elektromos impulzusokat megbízhatóan és jó időbeli felbontással tudjuk rögzíteni, egy adott csatornán másodpercenként legalább húszezerszer mintát veszünk a folytonos jelből, vagyis egy adott érzékelő jeléből ennyi adatpontunk (mért feszültségértékünk) keletkezik egy másodperc alatt. Ez a számérték (húszezer) az úgynevezett mintavételezési frekvencia,

mértékegysége a hertz (Hz). Kissé leegyszerűsítve a folyamatot azt mondhatjuk, hogy a digitális nyers mérési adatokat tulajdonképpen egy táblázatos formában mentjük le a számítógép tárolójára: a táblázat sorai tartalmazzák az egyes csatornákon (érzékelőkön) mért feszültségértékeket, oly módon, hogy a táblázat első oszlopa tartalmazza az összes csatornán mért legelső adatpontot (feszültségértéket), a második oszlop az időben következő (második) adatpontot, és így tovább. A humán EEG mérésekből eredő adatokat is hasonló módon tárolják: általában 32 vagy 64 csatornán (a fejbőrre helyezett elektródák számának függvényében), viszont kisebb mintavételezési frekvenciával (pl. 1000 Hz), mivel az EEG-vel egy másik fajta jeltípust (sok idegsejt közös aktivitásából eredő agyi ritmusokat, lásd lentebb) rögzítünk. Ennek röviden az az oka, hogy mivel az EEG elektródák az idegsejtek méretéhez képest túl nagyok és messze is találhatók azoktól, az EEG az idegsejtek által kibocsátott elektromos impulzusok közvetlen rögzítésére nem alkalmas.

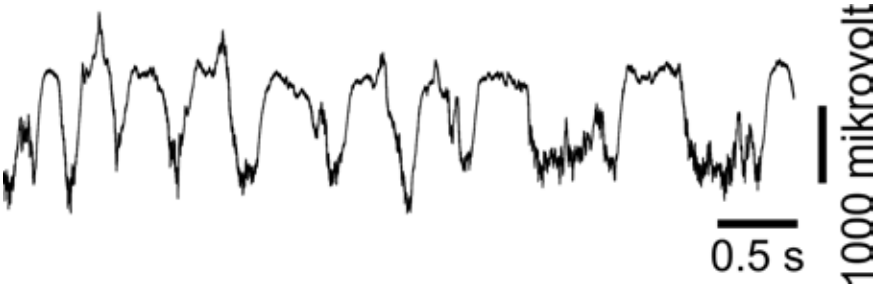


3. ábra: Az ábrán egy 128 érzékelőt tartalmazó agyi implantátummal rögzített agykérgi jel nagyjából negyed másodpercnyi szakasza látható, tíz különböző csatornán (10 érzékelő jele). A felvétel mindegyik csatornáján megfigyelhetők az idegsejtek által kibocsátott, kisebb vagy nagyobb tüskeként jelentkező elektromos impulzusok (ms – millimásodperc).

A sokcsatornás elektrofiziológia területén a nyers kutatási adatok mérete lényegében tehát az alábbiaktól függ: a használt csatornák (érzékelők) száma (minél több, annál nagyobb méretű az adat), a mintavételezési frekvencia (minél többször veszünk mintát egy másodperc alatt, annál nagyobb méretű adatfájl keletkezik), valamint a felvétel hossza (minél hosszabb a mérés, annál több tárterületet foglal). Egy 100 csatornán rögzített, 20 000 Hz-cel mintavételezett, egy óra (3600 mp) hosszúságú mérés például 14.4 GB nagyságú fájlt eredményez, amennyiben minden feszültségértéket két bájt nagyságú, egész számokat ábrázoló adattípusként tárolunk el. A tudományterületen általánosak az ennél hosszabb felvételek is, illetve gyakori a magasabb csatornaszámmal rendelkező eszközök használata. Hosszú távra beültetett implantátumokkal például szabadon mozgó rágcsálókból 3–4 óra hosszúságú mérések is megszokottak, akár több mint 300 csatornán, több száz gigabájt nagyságú fájlokat eredményezve. A jelenleg is gőzerővel zajló technológiai fejlesztések eredményeként a következő generációs agyi implantátumokon található érzékelők száma (ezáltal pedig a csatornák száma is) nagy valószínűséggel tovább fog emelkedni, így ennek folyományaként a kutatási adatok, adatcsomagok méretének jelentős növekedése is várható, hacsak nem kezdenek el időközben olyan tömörítési eljárásokat alkalmazni a területen, melyekkel érdemben csökkenthető a nyers adatok mérete.

Egy adott idegsejt által generált akciós potenciál voltaképpen válasz a hozzá kapcsolódó idegsejtektől rá érkező elektromos impulzusokra. Azonban érdemes megemlíteni egy másik, szintén nagyon fontos jeltípust is, melyet az agyból elektrofiziológiai módszerekkel rögzíteni tudunk: ez pedig az agyi ritmusok (más néven agyhullámok) jelensége. Ahogy fentebb is említettem, a hajas fejbőrre helyezett EEG elektródákkal az egyes neuronok elektromos impulzusait nem tudjuk észlelni, mivel ezek nagyon kis feszültségű és térben gyorsan elhaló események. Az extracelluláris akciós potenciálok csupán néhány tíz mikrométeres távolságra detektálhatóak az érzékelőktől, így tehát az érzékelőnek nagyon közel kell elhelyezkednie egy adott neuronhoz, ha rögzíteni szeretnénk a kibocsátott jeleit. Továbbá az érzékelőknek is nagyon

apróknak kell lenniük, ez nagyjából egy átlagos neuron sejttestének megfelelő méretet (10–20 mikrométer) jelent. Agyunkban egyes idegsejtcsoportok gyakran közel egyszerre, egy időben (szinkron) aktívak, így ezen neuronpopulációk aktivitása (elektromos tere) összeadódik, ami jól elkülöníthető, ritmikusan változó (oszcilláló) jelet eredményez a felvételeken (4. ábra). Ezeket az agyi ritmusokat mind a nagyobb méretű EEG elektródákkal (nagyszámú, több millió idegsejt össztevékenysége), mind pedig a kisebb, agyszövetbe ültethető implantátumokkal (kisebb neuronpopulációk lokálisabb össztevékenysége) rögzíteni tudjuk. Az egyik, talán legismertebb ilyen agyi ritmus az alfa ritmus, melynek ciklusai (másodpercenként 8–12) jól megfigyelhetők az EEG felvételeken csukott szemű, nyugalmi állapotban lévő embereknél. A beültethető implantátumok érzékelőivel észlelt elektromos impulzusok mellett a kutatók általában az agyi ritmusokat is rögzítik, mivel például érdekes felfedezéseket lehet tenni az agyi ritmusok és a neuronok által generált elektromos impulzusok közötti időbeli összefüggéseket tanulmányozva.



4. ábra: Altatott patkányok agykérgéből rögzített ritmikus agyi tevékenység, az úgynevezett agykérgi lassú ritmus, melynek átlagosan egy másodpercenként jelenik meg egy-egy újabb ciklusa. s – másodperc

A kutatók a nyers mérési adatokat a kísérletek, mérések után a tudományterületre jellemző módszerekkel és algoritmusokkal feldolgozzák, majd elemzik. Ezek főként időtartományban történő elemzéseket jelentenek, mint például annak vizsgálata, hogy milyen időbeli kapcsolat (együttjárás vagy korreláció) figyelhető meg egy idegsejt tüzelési időpontjai

(mikor adott ki elektromos impulzusokat a neuron) és a jutalomadás (pl. cukrozott vizet kap az egér) időpontjai között. Egy másik jelentős vizsgálati módszertan a frekvenciatartományban történő spektrális elemzések, melyet többek között az agyi ritmusok vizsgálatára alkalmaznak, például annak tanulmányozására, hogy milyen agyhullámok alakulnak ki az agy egyes területein az alvás különböző fázisai alatt. A mesterséges intelligencia területén használt módszereket, mint például a mélytanuló algoritmusokat, is egyre gyakrabban alkalmazzák az idegtudományi kutatási adatok feldolgozásához és analizálásához. Jelen tanulmány szűkös keretei között csupán az egyik alapvető, de fontos feldolgozási módszert, a sejtválogatást (spike sorting) említem meg, mely gyakorlatilag az egyik kezdő lépése az extracellulárisan rögzített, nagyszámú neuron akciós potenciálját tartalmazó nyers adatok feldolgozásának. A sejtválogatás módszerével tulajdonképpen a sokcsatornás mérésekben található tüskéket rendeljük hozzá egy-egy idegsejthez, vagyis megpróbáljuk meghatározni, hogy melyik elektromos impulzus melyik neurontól eredhetett, mivel erről nincs tudomásunk a mérések során (nem ismerjük az idegsejtek érzékelőkhöz viszonyított helyzetét). Vagy egy másik nézőpontból megközelítve a módszert: annak azonosítása a cél, hogy mely impulzusok származhattak ugyanattól az idegsejttől. Ez az azonosítás általában a különböző idegsejtek elektromos impulzusai közötti kis (időbeli vagy nagyságbeli) különbségek alapján történik: ugyanazon neuron által kibocsátott elektromos impulzusok például nagyjából ugyanolyan hullámformával fognak megjelenni a felvételeken (amennyiben nem mozdítjuk el az érzékelőnket), míg egy szomszédos idegsejt akciós potenciáljai már némileg eltérő hullámformával kerülnek rögzítésre. Ennek a különbségnek a fő oka az, hogy az észlelt akciós potenciál hullámformájának nagysága és alakja nagymértékben függ az idegsejt érzékelőtől való távolságától, illetve az idegsejt típusától is. A sejtválogatás eljárásának végeredménye a kiválogatott idegsejtek listája lesz (ez néhány sejttől néhány száz neuronig terjedhet felvételenként), valamint ezen idegsejttől eredő elektromos impulzusok megjelenésének időpontjai az adott felvételen. Az adatok további elemzése során gyakran ezeknek az egyedi neuronoknak az időbeli tevékenységét viszonyítjuk egymáshoz, vagy a kísérlet egyéb változóihoz, mint például a fentebb említett jutalomadáshoz.

A származtatott kutatási adatok sok esetben kisebb méretűek, mint az eredeti nyers mérési adatok. Erre jó példa a sejtválogatás eredményeként kapott adatok: tegyük fel, hogy a néhány bekezdéssel korábban példaként felhozott egy óra hosszúságú felvételünkben azonosítottunk száz egyedi neuront, és ezek mindegyike átlagosan 5000 elektromos impulzust bocsátott ki az egy órás időintervallum alatt. Mivel mindegyik impulzus egyetlen számértékkel reprezentálható (az akciós potenciál kibocsátásának időpontja a felvétel kezdetéhez képest), a fenti adathalmaz összesen 500 000 adatponttal lesz leírható (ha feltesszük, hogy a mérésünk többi része zaj, és nem tartalmaz számunkra további hasznos információt). Ez így a nyers adathoz képest egy majdnem 15 000-szeres csökkenés az adatpontok számában, vagyis az adat méretében! Ez a méretcsökkenés előnyös lehet, ha például az elérhető szabad tárterület korlátjai miatt a nyers kutatási adatok hosszú távú archiválására, eltárolására nincs lehetőségünk. Ebben az esetben, bár némi információveszteség mellett, de a származtatott adatok eltárolása is egy jó és hasznos alternatíva lehet.

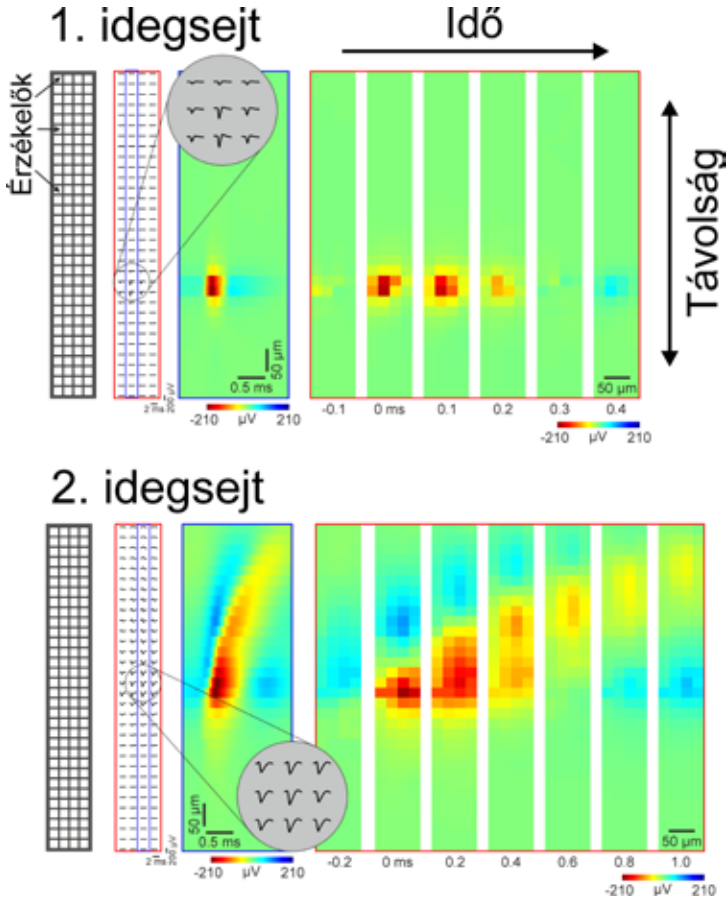
Érdemes még megemlíteni az adatok egy különleges típusát, a metaadatot. A metaadat lényegében a kutatási adatot leíró információ, megadása nélkül az adatok nehezen értelmezhetők, így sokszor használhatatlanok lennének. Esetünkben a metaadatok közé sorolhatók például az elektrofiziológiai mérések paraméterei: hány csatornán, milyen mintavételezési frekvencia mellett, milyen implantátummal történt a mérés. De metaadat a vizsgált agyterület neve, hogy milyen agyi (sztereotaxiás) koordinátákra és milyen agyi mélységbe ültettük be az implantátumot, vagy, amennyiben valamilyen kísérleti protokollt is alkalmaztunk a kísérletek során, akkor ennek a paraméterei is. Utóbbira egy példa különféle vizuális ingerek alkalmazása a látórendszerhez kapcsolható agyi területek vizsgálatára. Ezen ingereket leíró jellemzők, paraméterek is a metaadatokhoz sorolhatók tehát. Véleményem szerint alapvető fontosságú, hogy lehetőségeinkhez mérten minél több releváns metaadatot mellékeljünk kutatási adataink mellé, hogy szükség esetén akár kollégáink, akár más kutatócsoportok (vagy akár a kísérleteket követően mi magunk is) értelmezni, használni tudják az eltárolt és megosztott

adatfájlokat – akár a kutatás eredményeinek reprodukálása céljából, akár új tudományos kérdések vizsgálatára.

3. Az archivált kutatási adatok és az ezekhez kapcsolódó tanulmányok rövid bemutatása

Az Integratív Idegtudományi Kutatócsoportban alkalmazott egyik alapvető kísérleti módszerünk a beültethető agyi implantátumok használata. A kutatócsoport az elmúlt évek során több ilyen típusú, szilícium-alapú implantátum fejlesztésében és tesztelésében is részt vett, illetve rendszeresen használja is ezeket az eszközöket különféle agyi ritmusok (pl. a mélyalvás, altatás során megfigyelhető lassú agykérgi ritmus; lásd a 4. ábrát) vizsgálatára rágcsáló modellen. Jelen tanulmányban az ilyen típusú eszközökkel rögzített sokcsatornás elektrofiziológiai mérések archiválásával és megosztásával kapcsolatos tanulságokat, valamint a szerzett tapasztalataimat szeretném bemutatni. Ezek a kísérleti adatok szélessávú mérések – tehát egyaránt tartalmazzák az agyi ritmusokat és az egyes neuronok elektromos impulzusait is –, általában 30–60 perc hosszúságúak és altatott patkányok agykérgéből (szürkeállományából) kerültek rögzítésre 32–384 csatornás implantátumokkal. Az adatok regisztrálásához használt implantátumok speciális, nagy elektróda sűrűségű eszközök: a beültethető szárukon található nagyszámú érzékelő szorosan egymás mellett helyezkedik el (néhány mikrométer távolságra egymástól), lehetővé téve az agyi elektromos tevékenység nagy téri felbontású mintavételezését. A nagy téri felbontás egyik legfőbb előnye, hogy egy adott idegsejt által generált elektromos impulzust több szomszédos érzékelő is észlel, és – mivel az impulzus nagysága és alakja függ az érzékelő és a neuron távolságától – ez az impulzus eltérő hullámalakkal jelenik meg a szomszédos érzékelők jelén, az adott neuron egyfajta egyedi elektromos „lenyomatát" eredményezve (5. ábra). Ez a többcsatornás elektromos „lenyomat" segíti és megbízhatóbbá, pontosabbá teszi a neuronok azonosítását, vagyis a sejtválogatás folyamatát. A pilot projekt során archivált kutatási adatcsomagok a nyers mérések és a részletes metaadat információ mellett a sejtválogatás eredményeit, valamint a kiválogatott egyedi idegsejtek

különféle térbeli és időbeli jellemzőit, tulajdonságait is tartalmazzák. Ilyen jellemzők például, hogy átlagosan hányszor tüzeltek a neuronok egy másodperc alatt, milyen hosszú és mekkora nagyságú az akciós potenciáljuk, vagy hogy hány csatornán észlelhető egy adott neuron elektromos impulzusa. A nyers adatok és az egyedi idegsejtek minőségére vonatkozó különféle leírók is megtalálhatók a megosztott adatcsomagban (pl. mennyire voltak tiszták/zajosak a rögzített jelek), mely segíti a megfelelő (vagy akár a legjobb) minőségű mérések, kísérletek kiválasztását.



5. ábra: Két idegsejt többcsatornás elektromos „lenyomata". Az adatok rögzítéséhez egy 128 érzékelővel rendelkező implantátumot használtunk (az érzékelők elrendezése a bal oldalon látható). A középső és jobb oldali színes ábrákon markáns különbségek figyelhetők meg a két neuron akciós potenciál hullámformája között, mind térben, mind pedig időben (Horváth és mtsai (2021) alapján), (ms – millimásodperc, µV – mikrovolt, µm – mikrométer).

A projekt időtartama alatt három korábbi, nyílt hozzáférésű tanulmányunk (Fiáth és mtsai, 2019; Fiáth és mtsai, 2021; Horváth és mtsai, 2021) kutatási adatait archiváltuk és tettük szabadon elérhetővé a CONCORDA adatrepozitóriumban a tudományterületen használt egyik szabványos adatformátumban, metaadatokkal kiegészítve, a FAIR (megtalálható, hozzáférhető, átjárható, újrafelhasználható) alapelveknek megfelelően (Wilkinson és mtsai, 2016). A fejezet további részében röviden ismertetem a három tanulmányt és a kísérletsorozatok során keletkezett sokcsatornás mérési adatok főbb jellemzőit.

A kutatócsoport 2019-ben megjelent publikációjában (Fiáth és mtsai, 2019)[3] nagy téri felbontású, szilícium-alapú agyi implantátumokkal azt vizsgáltuk, hogy az implantátum agyszövetbe történő beültetési sebessége milyen hatással van az eszközzel mért agyi elektromos tevékenység minőségére, azaz romlanak vagy javulnak-e a rögzített jelek, attól függően, hogy lassabban vagy gyorsabban juttatjuk be az implantátumot az agyszövetbe. Ehhez altatott patkányok agykérgéből rögzítettük az elektromos aktivitást egy 128 érzékelőt tartalmazó implantátummal, négy különböző beültetési sebességet alkalmazva a kísérlet során. Ily módon négy darab, 45 perc hosszúságú adatfájlunk keletkezett minden kísérletben. Ezt a protokollt összesen tíz kísérletben ismételtük meg, ugyanazt a négyféle beültetési sebességet alkalmazva, csak más-más sorrendben. A leglassabb és leggyorsabb beültetési sebességekkel további kísérleteket végeztünk, egyrészt egy másik típusú, 32 csatornás eszközzel, másrészt szövettani vizsgálatok céljából. Összesen 68 darab nagy méretű (3–13 GB), sokcsatornás adatfájlunk keletkezett a kísérletsorozatban, hozzávetőleg 0.7 TB összmérettel.

A második tanulmányban (Fiáth és mtsai, 2021)[4] azt vizsgáltuk, hogy van-e különbség a rögzített jelek minőségében, ha azokat sokcsatornás, nagy téri felbontású implantátumok szárának különböző részein

3 https://science-data.hu/dataset.xhtml?persistentId=doi:10.5072/FK2/QWSIXM

4 https://science-data.hu/dataset.xhtml?persistentId=doi:10.5072/FK2/CGMYAH

(szélén vagy közepén) elhelyezkedő érzékelőkkel rögzítjük. Ehhez ötféle, különböző tulajdonságokkal rendelkező, szilícium-alapú implantátummal (32–384 közötti csatornaszámmal) mértük altatott patkányok agykérgi elektromos tevékenységét. Néhány esetben más kutatócsoportok által megosztott, nyílt hozzáférésű mérési adatokat is felhasználtunk a vizsgálatokhoz. A 20–60 perces felvételek száma eszközönként eltérő volt (6–21 darab mérés, összesen 54 felvétel), és a különböző időtartamokból, valamint az eltérő csatornaszámokból adódóan ezen felvételek mérete is jelentős különbségeket mutatott (2–28 GB). A kísérletsorozatban keletkezett nyers mérési adatok összmérete nagyjából 0.8 TB-ot tett ki, ebbe beleértve a külső forrásból származó adatokat is.

A harmadik adathalmaz (Horváth és mtsai, 2021)[5] egy olyan kutatási projektben keletkezett, ahol a fentebb már említett 128 érzékelős implantátummal rögzítettük agykérgi idegsejtek aktivitását különböző agykérgi területekről és rétegekből. A 30–60 perc hosszúságú adatfájlok mérete 8–17 GB közötti, az adatcsomag összesen 109 nyers mérési adatot tartalmaz, megközelítőleg 1.2 TB összmérettel. Az adatcsomagot a kutatási adatok bemutatására és megosztására specializálódott Scientific Data folyóiratban publikáltuk, és a publikáció megjelenésével egyidejűleg elérhetővé tettük a német G-Node (German Neuroinformatics Node) adatarchívumában, a GIN-ben (G-Node INfrastructure).[6] Az ennél az adatrepozitóriumnál tapasztalt lassú letöltési sebességek miatt azonban hasznosnak és szükségesnek láttuk ezeknek az adatoknak a duplikálását egy másik adatrepozitóriumban.

Mind a három fenti adatcsomagból kinyertük az egyedi idegsejtek aktivitását a sejtválogatás módszerével (ami összességében több ezer egyedi neuronból álló adathalmazt eredményezett), így a nyers mérési adatok mellett ennek az eredményét is archiváltuk és megosztottuk, a fejezet első bekezdésében felsorolt további jellemzőkkel együtt. Az adatok bemutatására, megjelenítésére az adatcsomagok mellé készítettünk és

5 https://science-data.hu/dataset.xhtml?persistentId=doi:10.5072/FK2/OKPT5U

6 https://doi.gin.g-node.org/10.12751/g-node.arf7ol/

feltöltöttünk egy MATLAB programnyelv alapú programkódot (szkriptet) is, mely alapvető adatvizualizációs funkciókkal rendelkezik, így kiválóan használható a megosztott adatok megtekintésére.

4. Szabványos adatformátumok az idegtudomány területén

Az agyi elektromos tevékenység rögzítésére számtalan különféle elektrofiziológiai mérőrendszer létezik, ezek egy része kereskedelmi forgalomban kapható termék, más részük saját fejlesztésű rendszer. Ezek a mérőrendszerek általában a saját egyedi, sokszor zárt fájlformátumukban mentik el a rögzített adatokat, és gyakran ezeket a fájlformátumokat csupán a gyártó cég, általában költséges, saját fejlesztésű szoftverével lehet használni, illetve az elérhető nyílt forráskódú programok és programkódok csupán egy korlátozott része képes ezeket az adatfájlokat beolvasni. Ebből látható, hogy ha azt a célt szeretnénk elérni, hogy minél szélesebb körben felhasználják a megosztott kutatási adatainkat, törekednünk kell arra, hogy az adatokat olyan formátumban rögzítsük vagy olyan adatformátumra alakítsuk át, melyet az adatfeldolgozásra és elemzésre fejlesztett nyílt és kereskedelmi forgalomban kapható szoftverek többsége képes beolvasni, vagy az adott tudományterületen gyakran használt programnyelveken (pl. Python, MATLAB, R) írt programkódokkal könnyű kezelni.

A kutatási adatok tárolására, megosztására, használatára és feldolgozására fejlesztett szabványos adatformátumok kiválóan alkalmasak lehetnek a fenti feladatra, továbbá segíthetik a kísérleti eredmények reprodukálását, különböző laboratóriumok adatainak és eredményeinek összehasonlítását, az adatok újrafelhasználását, vagy a kutatócsoportok közötti együttműködéseket. A csillagászat területén például már több mint 40 éve kidolgozták az első szabványos adatformátumot (FITS, Flexible Image Transport System), amit azóta is széles körben használnak, rendkívüli mértékben segítve a terület kutatásait, valamint az asztronómiai adatok archiválását és megosztását. Az elektrofiziológia területén a különböző adatformátumok egységesítése, és szabványos adatformátumok kifejlesztése az elmúlt évtized

közepén kezdődött el és jelenleg is tart. A kísérletes elektrofiziológia területén két szabványos adatformátum terjedt el az utóbbi években: a NIX (Neuroscience Information eXchange; Stoewer és mtsai, 2014)[7] és az NWB (Neurodata Without Borders; Rübel és mtsai, 2019; Rübel és mtsai, 2021; Teeters és mtsai, 2015).[8] Előbbit inkább Európában, míg utóbbit főként az Amerikai Egyesült Államokban található laboratóriumokban használják. Mindkét adatformátum a szabványos HDF5 (Hierarchical Data Format version 5) fájlformátumot használja az adatok tárolására, és gazdagon annotálható metaadattal. Jelenleg mind a NIX, mind pedig az NWB formátumot csupán néhány tucat kutatócsoport használja, de ezek a számok folyamatosan emelkednek. Természetesen a szabványos adatformátumok elterjedését jelentősen segíti az is, ha olyan eszközöket és szoftveres környezetet fejlesztenek ki ezek mellé, melyek alkalmasak az ilyen formátumban tárolt adatok közvetlen feldolgozására és elemzésére. Végül érdemes még megemlíteni a BIDS:EEG (Brain Imaging Data Structure – Electroencephalography)[9] szabványos adatformátumot is (Gorgolewski és mtsai, 2016; Pernet és mtsai, 2019), mely a humán EEG regisztrátumok tárolására és megosztására fejlesztett adatstruktúra.

A mi választásunk az NWB adatformátumra esett, pontosabban annak újabb verziójára az NWB:N 2.0-ra (Neurodata Without Borders: Neurophysiology version 2.0), mivel több olyan kiemelkedő, az USA-ban található élettani labor is ezt használja, melyek hasonló területen vagy kísérleti metodikával tevékenykednek, mint kutatócsoportunk (pl. a Seattle-ben található Allen Institute for Brain Science, Buzsáki György és Losonczy Attila laborjai New Yorkban, vagy Soltész Iván laboratóriuma a Stanford Egyetemen). Az NWB formátumot különféle fiziológiás módszerekkel kinyert adatok tárolására fejlesztették, többek között ide tartozik az intracellulárisan (sejten belülről) vagy extracellulárisan (sejtek közötti térből) rögzített elektromos idegi tevékenység, az optikai fiziológiás kísérletek mérései (ilyen például a kétfoton pásztázó lézermikroszkópiás módszerrel végzett kalcium képalkotás), vagy különféle kísérletes ingerlési paradigmák adatainak tárolása.

7 http://g-node.github.io/nix/
8 https://www.nwb.org/
9 https://bids.neuroimaging.io/

5. Hazai és nemzetközi adatrepozitóriumok idegtudományi kutatási adatok befogadására

A közlésre beadott vagy elfogadott tanulmányokhoz felhasznált kutatási adatok megosztása már több rangos tudományos folyóiratnál is kötelező (akár már a szakmai bírálati folyamat során), többek között azért, hogy könnyebben rá lehessen akadni az elemzés során elkövetett esetleges hibákra, ellenőrizni lehessen az eredmények reprodukálhatóságát, vagy, hogy ezzel megakadályozzák a kutatási adatokból kinyert eredmények sajnálatos módon néha előforduló manipulálását. A kutatási adatokat általában az adatok tárolására és megosztására szakosodott intézményi vagy nagyobb hazai és nemzetközi szervezetek által létrehozott repozitóriumokban lehet elhelyezni. Vannak általános adatrepozitóriumok, melyek bármelyik tudományterületről fogadnak adatokat, az ismertebbek közé tartozik például a figshare,[10] a Zenodo[11] vagy az Open Science Framework (OSF).[12] Léteznek tudományterület-specifikus repozitóriumok is, az idegtudomány területén ilyenek például a G-Node,[13] az OpenNeuro,[14] az EBRAINS,[15] a DANDI (Distributed Archives for Neurophysiology Data Integration)[16] vagy a NeuroMorpho,[17] hogy csak néhányat említsünk a legnagyobbak és legjelentősebbek közül. Ezek közül több nem csupán adatrepozitórium, hanem komplex kutatási infrastruktúra különféle kutatást segítő szolgáltatásokkal. A hazai tudományos közösség számára készült első jelentős adatrepozitórium az ELKH Számítástechnikai és Automatizálási Kutatóintézete által fejlesztett CONCORDA,[18] mely 2020-ban kezdte meg működését, de az adatrepozitórium fejlesztése jelenleg is zajlik az ELKH Adatrepozitórium Projekt keretein belül.[19]

10 https://figshare.com/
11 https://zenodo.org/
12 https://osf.io/
13 https://gin.g-node.org/
14 https://openneuro.org/
15 https://ebrains.eu/service/share-data
16 https://www.dandiarchive.org/
17 https://neuromorpho.org/
18 https://science-data.hu/
19 https://science-research-data.hu/

A repozitóriumok között jelentős különbségek lehetnek a feltöltött adatokra vonatkozó mennyiségi- és méretkorlátokban (pl. az egyedileg feltölthető fájl méretében, az adatcsomag méretében vagy a feltölthető fájlok számában), az esetleges költségekben (sok repozitórium ingyenes egy adott méretkorlátig), a felület használatának módjában vagy egyéb követelményekben. Érdemes tehát ezeket figyelembe venni a kutatási adataink számára megfelelő adattár kiválasztásakor. A kiválasztáshoz több hasznos online segédeszközt is találunk, ilyen például a re3data (REgistry of REsearch data REpositories),[20] melynek felületén egyszerűen kereshetünk az elérhető adattárak között, akár tudományterületenkénti, akár országonkénti bontásban. Az oldal hasznos információkat tartalmaz az adatrepozitóriumokról, például, hogy használ-e DOI-t (Digital Object Identifier) a megosztott adatcsomagok azonosításához, milyen metaadat szabványt alkalmaz, vagy milyen típusú adatokat fogad. Hasonló célt szolgál a FAIRSharing[21] weboldal is, melyen azonban a repozitóriumok mellett különféle szabványokat (pl. szabványos adatformátumokat) és kutatásiadat-kezeléshez kapcsolódó irányelveket is lehet keresni.

6. A kutatási adatok archiválása a CONCORDA adatrepozitóriumba, tanulságok és tapasztalatok

Az adatarchiválás folyamatának első lépése a három tanulmányunk nyers és származtatott kutatási adatainak összegyűjtése és rendszerezése volt. Esetünkben a származtatott adatok a sejtválogatás eredményét jelentették. Az alkalmazott kísérleti módszereinkből kifolyólag nagy adattömegekkel dolgozunk, ezért a kutatócsoportunk rendelkezik egy több terabájt kapacitású hálózati adattároló (NAS) eszközzel, mely többszörösen védve van adatvesztés ellen. Az egyes kísérletek elvégzése után közvetlenül erre az eszközre kerül archiválásra minden nyers mérési adat, így ezek összegyűjtése nem okozott problémát. A származtatott adatok összegyűjtése viszont már okozott némi fejtörést. A nyers adatok feldolgozását és elemzését ugyanis sokszor

20 https://www.re3data.org/
21 https://fairsharing.org/

különböző személyek végzik, különböző számítógépeken és más helyszíneken, és ezen vagy egyéb okokból kifolyólag a tanulmány közzététele után a származtatott adatok néha elkallódhatnak, törlésre kerülhetnek. Példának okáért tegyük fel, hogy van egy kutató, aki egy adott projekten dolgozott, majd a projekt végeztével egy másik kutatócsoportba került, és a projekt során használt számítógépét (rajta a származtatott adatokkal) pedig egy másik kutató kapta meg. Utóbbi kutató, mivel szüksége van szabad tárterületre egy másik projekthez, lementi a számítógépen található korábbi projektadatokat külső merevlemezre, ami ezután általában egy fiókba (és ezzel gyakran a feledés homályába) kerül. Néhány esetben ezt a külső tárolót egy harmadik kutató kapja meg, aki viszont már nagy valószínűséggel nem fogja tudni, hogy honnan erednek a tárolón található adatok, és azt feltételezvén, hogy korábban már archiválták ezeket az adatokat, törli azokat.

A származtatott adatok hosszú távú megőrzésére egy kutatócsoportnak gyakran sajnálatos módon nincs elegendő tárterülete, így előfordulhatnak a fentihez hasonló esetek. Különösen a több évvel korábbi kutatási projektek alatt összegyűjtött adatok lehetnek hiányosak a nem megfelelő és nem rendszeres adatarchiválás miatt. Esetünkben az adatok egy részénél a sejtválogatás után kapott adatfájlok már nem voltak fellelhetők a NAS tárolón archivált és egyéb tárolókon található adatok között. Emiatt úgy döntöttünk, hogy a korábban a sejtválogatáshoz használt szoftver egy újabb, jelentősen továbbfejlesztett verziójával ismét megcsináljuk ezt a hiányzó adatfeldolgozási lépést. Habár ennek az extra lépésnek jelentős volt az időigénye, jelentős volt a haszna is: ezáltal értékes információkhoz jutottunk mind az eredményeink reprodukálhatóságát illetően (más kutató végezte a feldolgozást egy más verziójú szoftverrel, mégis hasonló eredményre jutva), mind pedig a tanulmány eredményeinek megbízhatóságát tekintve (a tanulmány eredményeit és konklúzióját nem változtatta meg az új feldolgozás eredménye).

Mivel a projektünk esetében összességében nagy adattömegről volt szó (kb. 2.5 TB), az adatok összegyűjtését követően azokat nagy sebességű külső SSD (solid state drive) tárolókra másoltuk, hogy a szabványos adatformátumba történő átkonvertálást, a különféle jellemzők

kinyerését és az adatcsomagok mozgatását fel tudjuk gyorsítani. Mind a nyers, mind pedig a származtatott adatok, valamint az adatokat leíró és minőségüket bemutató értékek (kb. 20-30 különféle érték kiszámítása) is az NWB fájlokba kerültek becsomagolásra. Az NWB fájlok elkészítéséhez először MATLAB programnyelven elkészítettük a szükséges szkripteket, illetve módosítottuk egy korábbi projektből meglévőket. Az NWB formátumhoz Python (PyNWB)[22] és MATLAB (MatNWB)[23] alapú alkalmazásprogramozási felületet (API) is fejlesztettek, melynek a segítségével, a felület megismerése után és némi programozási tapasztalattal, elkészíthetők az NWB adatfájlok. Mi a MatNWB API-t használtuk erre a feladatra. Az NWB formátum mélyebb megismerése és a MatNWB használatának elsajátítása érdekében a pilot projekt során részt vettünk egy angol nyelvű online kurzuson, de korábbi adatarchiválási és programozási tapasztalataink is segítettek a tanulási folyamatban. Érdemes itt megemlíteni az NWB fájlok egyik előnyét. Mivel a HDF5 adatstruktúra képezi az NWB fájlok alapját, ezért lehetőség van az adatok tömörítésére, akár jelentős méretcsökkenést is elérve a végleges adatállományban, ezzel pedig akár az adatok tárolásának költségeit is csökkenteni tudjuk. Tömörítés alkalmazásával az adataink eredeti összméretét mintegy ötödével, nagyjából 2 terabájtra sikerült lecsökkentenünk. A tömörítés hátránya, hogy az adatok beolvasásának ideje kis mértékben nőhet, mivel azokat előbb ki kell csomagolni.

Az NWB fájlok elkészítése után következett az adatok feltöltése a CONCORDA repozitóriumba. A Dataverse[24] nyílt platformon alapuló CONCORDA adatrepozitórium használata az angol nyelvű leírás alapján könnyen megtanulható. Mind új tárolót, mind pedig új adatcsomagokat nehézségek nélkül létre tudtunk hozni a felületen. Egy tárolóba készíthetünk új tárolókat vagy adatcsomagokat, és utóbbiba helyezhetjük az adatainkat. Az Integratív Idegtudományi Kutatócsoport tárolójában létrehoztunk három adatcsomagot, mind a három tanulmányunkhoz egyet-egyet. Jelenleg mind a három adatcsomag nyilvános

22 https://pynwb.readthedocs.io/
23 https://github.com/NeurodataWithoutBorders/matnwb/
24 https://dataverse.org/

(az elérhetőségüket lásd fentebb a 3. fejezetben). A felhasználó szempontjából a metaadatok szerkesztése (pl. szerzők, kulcsszavak, kapcsolódó közlemények, használt szoftverek vagy egyéb megjegyzések megadása) és a fájlok feltöltése is viszonylag egyszerű. A kutatóintézeten belülről az adatfájlok feltöltése és letöltése is gyorsnak bizonyult, a 10 GB feletti mérettel rendelkező adatfájlok is problémamentesen feltölthetők voltak (jelenleg 100 GB a feltöltési korlát egy fájl esetén). Véleményem szerint hasznos lenne, ha az alapértelmezett „Tábla" nézet mellett a „Fa" nézetben is lehetőségünk lenne egyszerre több fájl kiválasztására és letöltésére (pl. egy adott könyvtárban lévő összes fájl letöltése), mivel utóbbi nézet használata az adataink könyvtárszerkezete miatt praktikusabb. Megpróbálkoztunk curl és Python alapú szkripteket használva a fájlok automatikus, kötegelt feltöltésével is, mely módszer egy programozásban járatos felhasználó számára a platform részletes dokumentációjának köszönhetően viszonylag gyorsan elsajátítható. Próbaképpen kisebb fájlokat sikeresen fel is tudtunk tölteni ezzel a módszerrel. Curl alapú programkóddal nagyobb méretű (>2GB) fájlok feltöltése is lehetséges, azonban mivel összességében viszonylag kevés, de nagy méretű feltöltendő adatfájlunk volt (a nyers és a származtatott adatok is a szabványos fájlformátumokba kerültek becsomagolásra), így végül a webes grafikus felhasználói felületről (GUI) elérhető egyszerű és kényelmes fájlfeltöltési módszer mellett döntöttünk.

Először a G-Node (GIN) nemzetközi adatrepozitóriumban elérhető kutatási adataink duplikálását végeztük el a CONCODA-ba, ezzel jelentősen javítva az adataink elérhetőségét. A körülbelül 0.9 TB összméretű és 112 fájlt tartalmazó adatcsomagot néhány óra alatt fel tudtuk tölteni a webes felületet használva. Ezt követte a két másik adatcsomag feltöltése hasonló módon. Minden adatcsomaghoz készítettünk egy rövid angol nyelvű leírást a fájlok elnevezésének logikájáról, a könyvtárszerkezet hierarchiájáról, a szabványos fájlformátum struktúrájáról, valamint rövid példakódokat is mellékeltünk a fájlok beolvasására MATLAB környezetben, illetve egy Python-alapú célszoftverben. Továbbá minden adatcsomaghoz feltöltöttünk két CSV (Comma Separated Values) fájlt, melyek a kísérletekhez és kísérleti állatokhoz kapcsolódó jellemzőket

(pl. állat súlya, neme, megcélzott agyi terület), valamint a mérési adatokhoz kapcsolódó információkat (pl. fájl mérete, hossza, minősége) tartalmazzák minden mérési adatfájlhoz.

Végül az adatok beolvasásához és megjelenítéséhez készítettünk egy MATLAB-alapú programkódot, melyet minden adatcsomag mellé feltöltöttünk az adatrepozitóriumba. A kód példákat mutat be a nyers mérési adatok, valamint a kinyert idegsejtek és tulajdonságaik beolvasására és megjelenítésére, továbbá alapszintű jelfeldolgozási lépéseket is (pl. jelek szűrése) tartalmaz. A szkriptet dokumentáltuk is, így a programkód könnyen átalakítható úgy, hogy különböző típusú agyi implantátumok adatait, vagy a mérési adatok különböző részeit is tanulmányozni lehessen vele. Az archivált adatcsomagok bibliográfiai leírását végül a Magyar Tudományos Művek Tárába (MTMT) is felvittük kutatási adattípusként.

Hivatkozásjegyzék

Fiáth R, Meszéna D, Somogyvári Z, Boda M, Barthó P, Ruther P, Ulbert I. Recording site placement on planar silicon-based probes affects signal quality in acute neuronal recordings. (2021) SCIENTIFIC REPORTS 11(1): 2028. https://doi.org/10.1038/s41598-021-81127-5

Fiáth R, Márton AL, Mátyás F, Pinke D, Márton G, Tóth K, Ulbert I. Slow insertion of silicon probes improves the quality of acute neuronal recordings. (2019) SCIENTIFIC REPORTS 9 (1): 111. https://doi.org/10.1038/s41598-018-36816-z

Fiáth R, Raducanu BC, Musa S, Andrei A, Lopez CM, van Hoof C, Ruther P, Aarts A, Horváth D, Ulbert I. A silicon-based neural probe with densely-packed low-impedance titanium nitride microelectrodes for ultrahigh-resolution in vivo recordings. (2018) BIOSENSORS & BIOELECTRONICS 106: 86–92. https://doi.org/10.1016/j.bios.2018.01.060

Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, Glatard T, Halchenko YO, Handwerker DA, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols TE, Pellman J, Poline JB, Rokem A, Schaefer G, Sochat V, Triplett W, Turner JA, Varoquaux G, Poldrack RA. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. (2016) SCIENTIFIC DATA 3(1): 160044. https://doi.org/10.1038/sdata.2016.44

Horváth C, Tóth LF, Ulbert I, Fiáth R. Dataset of cortical activity recorded with high spatial resolution from anesthetized rat (2021) SCIENTIFIC DATA 8(1): 180. https://doi.org/10.1038/s41597-021-00970-3

Pernet CR, Appelhoff S, Gorgolewski KJ, Flandin G, Phillips C, Delorme A, Oostenveld R. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. (2019) SCIENTIFIC DATA 6(1): 103. https://doi.org/10.1038/s41597-019-0104-8

Raducanu BC, Yazicioglu RF, Lopez CM, Ballini M, Putzeys J, Wang S, Andrei A, Rochus V, Welkenhuysen M, Van Helleputte N, Musa S, Puers R, Kloosterman F, Van Hoof C, Fiáth R, Ulbert I, Mitra S. Time multiplexed active neural probe with 1356 parallel recording sites. (2017) SENSORS 17(10): 2388. https://doi.org/10.3390/s17102388

Rübel O, Tritt AJ, Ly R, Dichter BK, Ghosh SS, Niu L, Soltesz I, Svoboda K, Frank LM, Bouchard K. The Neurodata Without Borders ecosystem for neurophysiological data science. (2022) BIORXIV 435173 https://doi.org/10.1101/2021.03.13.435173

Rübel O, Tritt A, Dichter B, Braun T, Cain N, Clack N, Davidson TJ, Dougherty M, Fillion-Robin JC, Graddis N, Grauer M, Kiggins JT, Niu L, Ozturk D, Schroeder W, Soltesz I, Sommer FT, Svoboda K, Ng L, Frank LM, Bouchard K. NWB:N 2.0: An Accessible Data Standard for Neurophysiology. (2019) BIORXIV 523035 https://doi.org/10.1101/523035

Steinmetz NA, Aydin C, Lebedeva A, Okun M, Pachitariu M, Bauza M, Beau M, Bhagat J, Böhm C, Broux M, Chen S, Colonell J, Gardner RJ, Karsh B, Kloosterman F, Kostadinov D, Lopez CM, O'Callaghan J, Park J, Putzeys J, Sauerbrei B, van Daal RJJ, Vollan AZ, Wang S, Welkenhuysen M, Ye Z, Dudman JT, Dutta B, Hantman AW, Harris KD, Lee AK, Moser EI, O'Keefe J, Renart A, Svoboda K, Häusser M, Haesler S, Carandini M, Harris TD. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. (2021) SCIENCE 372(6539): eabf4588. https://doi.org/10.1126/science.abf4588

Stoewer A, Kellner CJ, Benda J, Wachtler T and Grewe J. File format and library for neuroscience data and metadata. (2014) FRONTIERS IN NEUROINFORMATICS CONFERENCE ABSTRACTS: NEUROINFORMATICS https://doi.org/10.3389/conf.fninf.2014.18.00027

Teeters, JL, Godfrey K, Young R, Dang C, Friedsam C, Wark B, Asari H, Peron S, Li N, Peyrache A, Denisov, G, Siegle JH, Olsen SR, Martin C, Chun M, Tripathy S, Blanche TJ, Harris K, Buzsáki G, Koch C, Meister M, Svoboda K, Sommer FT Neurodata without borders: creating a common data format for neurophysiology. (2015) NEURON 88(4): 629–634. https://doi.org/10.1016/j.neuron.2015.10.025

Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship. (2016) SCIENTIFIC DATA 3(1): 160018. https://doi.org/10.1038/sdata.2016.18

Kis dózisoknál megfigyelhető hiperszenzitivitással és indukált sugárrezisztenciával kapcsolatos adatok gyűjtése és közzététele

Polgár Szabolcs[1,2]
ORCID: 0000-0003-1133-3890

Madas Balázs Gergely[2,3]
ORCID: 0000-0002-0698-4982

[1]Eötvös Loránd Tudományegyetem, Fizika Doktori Iskola
[2]Energiatudományi Kutatóközpont, Környezetfizikai Laboratórium
[3]Budapesti Műszaki és Gazdaságtudományi Egyetem, Fizikai Kémia és Anyagtudományi Tanszék

Bevezetés és célkitűzés

Ha az ionizáló sugárzás hatását sejtszinten vizsgáljuk, akkor egy sejtkultúra sugárérzékenységét a sejtek osztódási képességének változásával is jellemezhetjük. Ezekben a kísérletekben egy sejtet akkor tekintünk túlélőnek, ha a besugárzást követően legalább 5–6 osztódási cikluson keresztülmegy, azaz létrehoz egy legalább 50 sejtből álló kolóniát.[1]

Bizonyos sejtkultúrák esetén megfigyelhető, hogy kis dózisú besugárzás esetén a túlélő sejtek hányada a dózis növekedésével kezdetben meredekebben csökken, azaz több sejt pusztul el, mint amennyit várnánk. Ezt nevezzük kis dózis hiperszenzitivitásnak. Egy adott dózisnál (jellemzően 0,2–0,8 Gy között) a túlélési görbe elér egy lokális minimumot, amelyet követően a túlélő sejtek aránya a dózis függvényében növekedni kezd egy lokális maximumig (indukált sugárrezisztencia). A jelenség érdekes, hiszen meglepő, hogy van olyan dózistartomány, ahol a sugárdózis növekedésével nő a sejtek túlélési valószínűsége.

Az utóbbi években azt vizsgáltuk, hogy az ionizáló sugárzás milyen módon okoz mutációkat, illetve az élő rendszerek milyen szövetszintű válaszokkal

tudják ezt a mutagén hatást csökkenteni.[2] Megmutattuk, hogy miközben az ionizáló sugárzás mutagén voltát szinte kizárólag a pontatlanul javított DNS-sérülések kialakulásával magyarázzák, a belélegzett radonleányelemek esetén ennél jelentősebb tényező, hogy a sugárzás által elpusztított sejtek pótlása többletosztódásokkal jár, amely a mutációk számának növekedéséhez vezet.[3] Az ionizáló sugárzás tehát kétféleképpen okoz mutációkat: egyfelől növeli a DNS-sérülések számát, másfelől a sejtpusztító hatás révén többletosztódásokat vált ki a szövetben.

Ha az élő rendszer arra törekszik, hogy minimalizálja a mutációk számát, akkor az ionizáló sugárzás e két hatását egyszerre kell kezelnie. A két hatás kezelése azonban egymással ellentétes lépéseket követelhet. A DNS-sérülések miatt kialakuló mutációk számát azzal lehet csökkenteni, hogy elpusztulnak azok a sejtek, amelyekben sok mutagén sérülés van (apoptózis). A többletosztódások miatt kialakuló mutációk számát viszont éppen azzal lehet csökkenteni, ha minél több sejt életben marad, ehhez viszont a DNS-sérülések javításának intenzitását kell növelni. Ez alapján lehetséges, hogy a mutációk számának minimalizálásához különböző dózistartományokban (azaz különböző átlagos sérülésszámoknál) különböző stratégia tartozik.[4,5]

A fentiek vizsgálatára elkészítettünk egy matematikai modellt, amelyben minden sejt információt továbbít a környezetében lévő sejtek felé arról, hogy hány mutagén DNS-sérülés keletkezett bennük. A sejtek akkor pusztulnak el, ha mutagén sérüléseik száma nagyobb mértékben haladja meg a környezeti átlagot, mint amennyi mutációval egy sejtosztódás jár. Ennek a folyamatnak az eredményeként a teljes sejtpopulációban a mutagén sérülések száma csökken.[6]

A matematikai modell teszteléséhez rendkívül fontos, hogy annak jóslatait kísérleti adatokkal vethessük össze. Az összehasonlítás annál értékesebb, minél több és minél jobb minőségű kísérleti adat áll rendelkezésre. Természetesen ezek az adatok más modellek validálására is alkalmazhatóak. Mivel nincsen olyan átfogó adatbázis, ahol a korábbi mérések adatai könnyen elérhetőek lennének, ezért a kutatók

jellemzően a saját kísérleti adataikat használják. Ugyanakkor a nyers adatok közzététele nem jellemző a tudományterületen.[7,8]

A munka során létrehoztunk egy adatbázist, amely olyan túlélési görbék adatait tartalmazza, amelyekben megfigyelhető a kis dózis hiperszenzitivitás jelensége. Ezeket az adatokat kiegészítik a kísérlet jellemzőit leíró adatok, mint például a sejtvonal típusa, a sugárzás fajtája vagy a dózisteljesítmény, illetve a szerzők által illesztett modellek paraméterei és ezek bizonytalanságai is. Az adatbázis elkészítésénél és közzétételénél a FAIR elvekhez igazodtunk, amelyek szakterületünkön való megvalósulása kapcsán már két közleményünk is született.[7,8] A létrehozott adatbázis összesen 101 túlélési görbe adatait (dózis, túlélési hányad, illetve ezek bizonytalansága) tartalmazza, melyhez 46 cikk adatait dolgoztuk fel.

Az adatgyűjtés módszere

A munkánk első lépése a megfelelő publikációk kiválasztása volt. Elsőként a Google Scholar felületét használtuk, a keresési címszavak a „low-dose hyper-radiosensitivity", „low-dose hrs", és „induced radioresistance" voltak, ezen kívül az így talált cikkek referenciáit is vizsgáltuk. A cikkeknek tartalmaznia kellett legalább egy hiper-radioszenzitivitást mutató ábrát. A munkánk során angol nyelven elérhető publikációkat dolgoztunk fel. Összesen 46 tudományos cikk került feldolgozásra az 1993 és 2021 közötti időszakból. Az utolsó keresést 2021. augusztus 2-án végeztük el.

A következő lépés a megfelelő ábrák kiválasztása és az adatok leolvasása volt. Ha ugyanazon az ábrán több adatsor is szerepelt, akkor fontos kiválasztási szempont volt, hogy ezek jól elkülöníthetők legyenek. Emellett csak olyan ábrát elemeztünk, amelyen jól leolvashatóak voltak:

- a tengelyek és a tengelyek beosztásai,
- a feliratok,
- az adatpontok és az adatpontokhoz tartozó bizonytalanságok.

Az adatok leolvasása kézzel történt a WebPlotDigitizer 4.2 (GNU Affero General Public License v3.0) és az OriginPro 2018 (OriginLab Corporation) programok felhasználásával. A leolvasás során az első lépés az x, illetve az y tengely meghatározása volt, majd a skála definiálása (lineáris vagy logaritmikus), majd két-két pont segítségével az egységhossz megadása. Ezek után az adatpontok összegyűjtése egyesével történt meg. Egy adatponthoz az elnyelt dózis nagyságát (Gray-ben megadva), a túlélési hányadot és a túlélési hányadhoz tartozó bizonytalanság nagyságát (alsó és felső érték) jegyeztük fel. A bizonytalanság közlése a különböző közleményekben nem egységes, így ezt kifejezhették a standard hibával, vagy a 90 vagy 95%-os konfidencia intervallummal is. Összesen 101 adatsort olvastunk le.

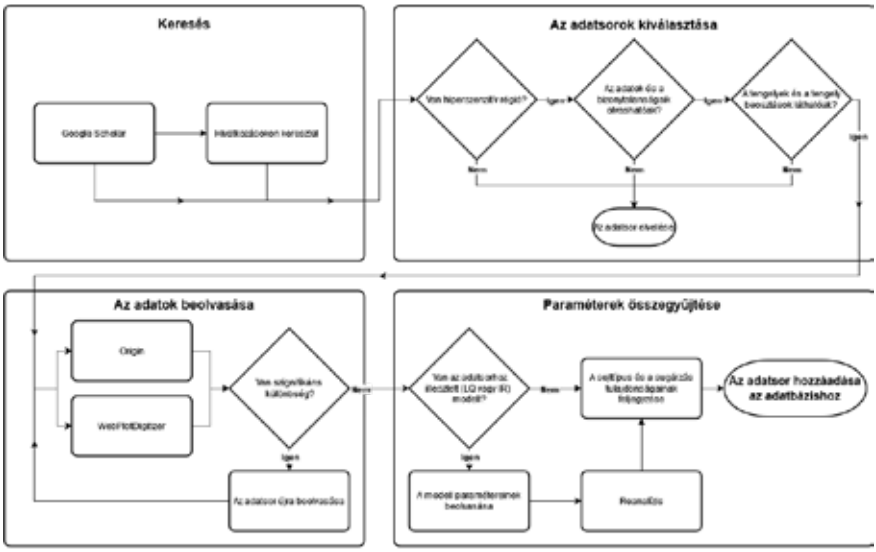
Az adatbázisban minden adatsornál az első oszlopban tüntettük fel, hogy melyik publikációból származik, rögzítve a címet, a szerzőket és a publikációhoz tartozó digitális objektumazonosítót (Digital Object Identifier, DOI), illetve azt is, hogy pontosan melyik ábráról történt a leolvasás.

A feldolgozott cikkekben a sejtkultúrák túlélésének számolási módszere nem volt azonos (történhetett számítógépes programok segítségével vagy azok nélkül), de egységesen azokat a sejteket tekintették túlélőnek, ahol a létrejött kolónia legalább 50 sejtből állt a vizsgálat idején. Három publikációban ugyan nem említették meg ezt, de a szerzők más közleményei alapján feltételezhetjük, hogy ugyanezt a meghatározást alkalmazták. A 'plating efficiency', azaz, hogy a nem besugárzott sejtek hány százaléka hoz létre legalább 50 sejtből álló kolóniát, általában nem volt megadva, de azt közölték, hogy a túlélési hányadot mindig a kontrollcsoport átlagos 'plating efficiency' értékére vonatkoztatva számolták ki.

A leolvasás pontosságát úgy ellenőriztük, hogy minden adatpontot kétszer, különböző programokat (WebPlotDigitizer 4.2 és az OriginPro 2018) alkalmazva olvastunk le. Amennyiben bármelyik adatpont túlélési hányad értékei esetén eltérést tapasztaltunk, azaz a leolvasott adatpontok között nagyobb eltérés volt mint 0,01, akkor egy újbóli,

harmadik leolvasást is elvégeztünk annak megállapítására, hogy hol követtük el a hibát. Az elnyelt dózisok esetén az eljárás hasonló volt, azzal a feltétellel kiegészítve, hogy az értékek 0,05 Gy egész számú többszörösei legyenek, mivel a gyakorlatban ilyen értékeken szokták a méréseket elvégezni.

A sejttípust, a sugárzás fajtáját és tulajdonságait is feljegyeztük. Ezen kívül, ha az eredeti cikkben a szerzők függvényt illesztettek az adatokra az LQ (linear-quadratic, magyarul lineáris-kvadratikus) vagy az IR (induced repair, magyarul indukált javítás) modell alapján, akkor a modell paramétereinek értékeit és ezek bizonytalanságait (szórás vagy 95%-os konfidencia intervallum) szintén feljegyeztük. Összesen 24 adatsornál adtak meg LQ paramétereket (két paraméter) és 59 adatsornál IR paramétereket (négy paraméter).
Az adatgyűjtés menetét az 1. ábrán foglaltuk össze.



1. ábra: Az adatgyűjtés menetének folyamatábrája

Az adatokat Excel táblázatban rögzítettük (Microsoft Excel 2016, Microsoft Corporation). A publikációk időrendi sorrendben helyezkednek el

egymás alatt: a legkorábban publikált cikk található legfelül és lefelé haladva találhatók az újabb közlemények. Minden adatsornál feltüntettünk minden információt akkor is, ha ugyanabból a publikációból több leolvasás is történt, mivel így mindegyik egymástól függetlenül értelmezhető. Az első oszlopban található az adott adatsorhoz tartozó publikáció címe, szerzői, a DOI és a leolvasott ábra sorszáma. Ezt követi maga az adatsor (a túlélési hányad és a hozzá tartozó elnyelt dózis értékek), majd a következő oszlopokban az LQ, illetve IR paraméterek értékei ('X' karakterrel jelölve, ha az adott adatsorhoz nem adták meg az adott paramétert, illetve '–' jellel jelezve, ha az adott illesztésnél a paraméter nem értelmezhető). A két modell paraméterei egymás alatt helyezkednek el. Minden adatsor esetén a modellparaméterek utáni oszlopokban, jobboldalt található a vizsgált sejtkultúra típusa és mellette a besugárzás jellemzői.

Az LQ és IR paramétereket mi is meghatároztuk a nyers adatokra való illesztéssel, hogy teszteljük azok helyességét. Az illesztéshez az OriginPro 2018 (OriginLab Corporation) programot használtuk kiindulásként a Levenberg – Marquardt algoritmus alkalmazásával. Abban az esetben tekintettük az illesztésünket különbözőnek az eredetileg megadott értékektől, ha legalább egy paraméter esetén a két illesztés különbsége nagyobb volt, mint a bizonytalanságok összege.

Mivel az LQ modell nem veszi figyelembe a kis dózis hiperszenzitivitási régiót, ezért ebben az esetben a modellt csak a nagy dózistartományban található adatpontokra illesztettük, azaz olyan adatokra, ahol az elnyelt dózis értéke meghaladta az 1 Gy-t, de minden esetben legalább három adatpontra. Ha ez nem adott megfelelő eredményt, akkor az egész adatsorra illesztettük a függvényt. Az LQ modellnek két paramétere van: α és β, amelyekkel a függvény a következő alakot ölti:

$$SF = e^{-\alpha D - \beta D^2}$$

Ahol *SF* jelöli a túlélési hányadot. Az IR modell esetén ez az összefüggés az alábbiak szerint módosul:

$$SF = e^{-\alpha_r\left(1+\left(\frac{\alpha_s}{\alpha_r}-1\right)e^{-\frac{D}{D_c}}\right)D-\beta D^2}$$

Ez azt jelenti, hogy az IR modell esetén egy négyparaméteres illesztést kell elvégezni, ami azt eredményezi, hogy nagyon erősen függ a paraméterek kezdeti értékeitől, hogy konvergál-e, illetve hová konvergál a függvény. Ezért annak érdekében, hogy megtaláljuk azt az illesztési módszert, ami visszaadja a cikkekben megadott értékeket, az alábbi protokollt használtuk. Minden esetben, ha az illesztett értékek közül legalább egyben szignifikáns különbség volt, akkor megpróbálkoztunk a következő lépéssel:

1) Az α_r és β paraméterek kezdeti értékeit az LQ modell illesztésből számoltuk ki (mivel ez a két paraméter megfelel az ott található α-nak és β-nak). Az α_s és D_c paraméterek kezdeti értékeit pedig 1 Gy^{-1}-nek és 1 Gy-nek hagytuk. Ezekkel a kezdeti értékekkel illesztettük a fenti függvényt, figyelembe véve a nyers adatok bizonytalanságát.
2) A kezdeti értékeket mind a négy paraméter esetén a cikkben megadottra állítottuk és így illesztettünk, figyelembe véve a nyers adatok bizonytalanságát.
3) A kezdeti értékeket az 1) lépésnek megfelelően választottuk meg, de az illesztést az adatok bizonytalanságának figyelembevétele nélkül végeztük el.
4) A kezdeti értékeket a 2) lépésnek megfelelően választottuk meg, de az illesztést az adatok bizonytalanságának figyelembevétele nélkül végeztük el.
5) A kezdeti értékeket a 2) lépésnek megfelelően választottuk meg, de a függvények logaritmusát illesztettük úgy, hogy az adatpontok esetén is a túlélési hányadok logaritmusával számoltunk a bizonytalanságok figyelembevétele nélkül.

6) A Levenberg – Marquardt algoritmus helyett az Orthogonális Távolságregressziós módszerrel illesztettük az előző öt lépésnek megfelelően, mindet végigpróbálva.
7) Végül az előző lépéseket követve egy paraméter rögzítésével próbáltunk illeszteni. Ennek a lépésnek az indoka, hogy három változóra könnyebben találja meg az optimumot a program, mivel kevesebb paramétert kell egyidejűleg illeszteni.
   a) Ha a β paraméter negatív volt az LQ modell alapján, akkor ezt rögzítettük 0-ra.
   b) Egyéb esetekben pedig az α_r paraméter értékét rögzítettük az LQ modellből kapott α értéknek megfelelően.

Az 59 esetből, ahol az IR illesztés adott volt, összesen három olyan adatsort találtunk, ahol az általunk végzett protokoll után legalább egy paraméter értéke szignifikánsan eltért a cikkben megadottétól. Valószínűsíthető, hogy ennek oka elírás lehetett az eredeti cikkben; az egyik esetben egy negatív előjel elhagyása.

Az adatok közzétételének módja

A létrejött adatbázist a STORE^{DB} (https://www.storedb.org/store_v3/) angol nyelvű repozitóriumban helyeztük el, ahol nyilvánosan elérhető. A repozitórium digitális objektumazonosítót rendel a projekthez, és az ahhoz tartozó adatsorokhoz is. A projekt a https://doi.org/10.20348/STOREDB/1163 linken keresztül, az adatsorok pedig a https://doi.org/10.20348/STOREDB/1163/1252 linken keresztül érhetőek el. Az adatbázis leírása a Scientific Data című folyóiratban jelent meg.[9]

A STORE^{DB} repozitóriumban való elhelyezés révén az adatok megfelelnek a FAIR elveknek. Az egyedi azonosítón keresztül megtalálhatóak (findable) és hozzáférhetőek (accessible). Mivel az adatmennyiség nem nagyon nagy, a feldolgozhatóságot (átjárhatóságot, interoperable) a táblázatos forma biztosítja. Az adatok újrafelhasználhatóak (reusable) a CC-Attribution ShareAlike licenc szerint.

Összefoglalás

Az adatbázis létrehozásának célja a hiper-radioszenzitivitást mutató különböző sejttenyészeteket és sugárzásokat felhasználó publikált tudományos munkák kísérleti eredményeinek összegyűjtése volt.

Összesen 46 közleményből 101 kísérlet eredményét dolgoztuk fel és gyűjtöttük össze. Az adatbázis magja a publikációkból kinyert kísérletek eredményei: a vizsgált dózis nagysága a hozzá tartozó sejttúlélési aránnyal és ennek bizonytalanságával, feltüntetve a vizsgált sejtkultúrát és a besugárzási paramétereket. Amennyiben a szerzők készítettek függvényillesztést az eredményeikre, akkor a függvény paramétereit és azok hibáit szintén rögzítettük. Az adatbázis tartalmazza továbbá a feldolgozott cikkek címét, szerzőit, a DOI-t, illetve a leolvasáshoz használt ábra sorszámát.

Az adatbázisban összegyűjtött adatsorok, reményeink szerint, a kis dózis hiperszenzitivitás témakörében jól használhatók lesznek modellek validálására, illetve a sokféle sejtkultúra, és a különböző besugárzások miatt rendkívül sokszínű kísérletek eredményeként jó összehasonlítási alapok lehetnek az újabb kísérleti eredmények értelmezése esetén.

A létrejött adatbázist a STORE^{DB} (https://www.storedb.org/store_v3/) angol nyelvű repozitóriumban helyeztük el.

Köszönetnyilvánítás

A kutatást támogatta a Hungarian Research Data Alliance és a Magyar Tudományos Akadémia Könyvtár és Információs Központja (21-61), az Euratom 2019–2020. évi kutatási és képzési programja (900009, RadoNorm), a Magyar Tudományos Akadémia Bolyai János Kutatási Ösztöndíja (MBG, bo-37-2021) és az Innovációs és Technológiai Minisztérium ÚNKP-21-5 kódszámú Új Nemzeti Kiválóság Programja a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból (MBG, ÚNKP-21-5-BME-387).

Felhasznált irodalom

1 Franken, N. A. P., Rodermond, H. M., Stap, J., Haveman, J. & van Bree, C. Clonogenic assay of cells in vitro. *Nat. Protoc.* **1**, 2315–2319 (2006), https://doi.org/10.1038/nprot.2006.339.

2 Madas, B. G. Radon induced hyperplasia: effective adaptation reducing the local doses in the bronchial epithelium. *J. Radiol. Prot.* **36**, 653–666 (2016), https://doi.org/10.1088/0952-4746/36/3/653.

3 Madas, B. G. & Balásházy, I. Mutation induction by inhaled radon progeny modeled at the tissue level. *Radiat. Environ. Biophys.* **50**, 553–570 (2011), https://doi.org/10.1007/s00411-011-0382-9.

4 Madas, B. G., Hanusovszky, L. & Drozsdik, E. Kis dózis, nagy érzékenység: a sugárvédelmi szabályozás alapfeltevése és a sejtek hiperszenzitivitása. *Magy. Tud.* **177**, 62–67 (2016).

5 Madas, B. G. & Drozsdik, E. J. Computational modelling of low dose hyper-radiosensitivity applying the principle of minimizing mutation rate. *Radiat. Prot. Dosimetry* **183**, 147–150 (2019), https://doi.org/10.1093/rpd/ncy227.

6 Polgár, Sz. Comparison of mathematical models of low-dose hyper-radiosensitivity and induced radioresistance based on experimental data. (Eötvös Loránd Tudományegyetem, 2019).

7 Madas, B. *et al.* FAIRing the radiation science commons. *BIO Web Conf.* **14**, 08002 (2019), https://doi.org/10.1051/bioconf/20191408002.

8 Madas, B. G. & Schofield, P. N. Survey on data management in radiation protection research. *Radiat. Prot. Dosimetry* **183**, 233–236 (2019), https://doi.org/10.1093/rpd/ncy250.

9 Polgár, Sz., Schofield, P. N. & Madas, B. G. Datasets of *in vitro* clonogenic assays showing low dose hyper-radiosensitivity and induced radioresistance. *Sci. Data* **9**, 555 (2022), https://doi.org/10.1038/s41597-022-01653-3.

A BTK Régészeti Intézet rajzgyűjteményének közzététele a Concordában, európai gyakorlatok a régészeti archiválásban

Horváth Friderika
Bölcsészettudományi Kutatóközpont, Régészeti Intézet

Kiss Tünde
ORCID: 0000-0003-1905-0027
Bölcsészettudományi Kutatóközpont, Régészeti Intézet

Az emberiség minduntalan törekszik arra, hogy hagyatékát maradandó formában az utókor számára is megőrizze; a múzeumok, levéltárak, könyvtárak, irattárak, személyes hagyatékok tele vannak olyan tárgyakkal, alkotásokkal, írott anyagokkal, melyek történeti, kulturális vagy jogi okoknál fogva az egyén vagy a közösség számára fontosak, megőrzendő és megosztandó értéket képviselnek.

A Bölcsészettudományi Kutatóközpont (BTK) Régészeti Intézetének rajzgyűjteménye – benne a régészet számára nagy értékű, ma már nem reprodukálható helyszínrajzokkal, ásatási felmérésekkel, továbbá különböző múzeumokban őrzött tárgyakról készült elemző grafikákkal, publikációk illusztrációs anyagával és hatalmas fotóanyaggal – az egyik legjelentősebb, önálló információs tartalommal bíró kutatási adatforrást képviseli.

Az Adattár és a gyűjtemények kezdete a Régészeti Kutató Csoport (1967-től Régészeti Intézet) 1958. évi megalakulásáig nyúlik vissza.[1] A Magyar Tudományos Akadémia által „nagyszabású közös feladatokra" életre hívott kutatócsoport feladatköre kezdetben három jelentősebb kutatási súlypont köré rendeződött: a dunai vízierőmű építésével kapcsolatos megelőző leletmentések elvégzése, amelyhez később Magyarország Régészeti Topográfiájának (MRT), valamint a „magyar föld régészetének

1 A Kutató Csoportot az MTA 13/1958. sz. elnöki utasítása hívta életre.

összefoglalását nyújtó régészeti kézikönyv" elkészítése társult.[2] A kutatások támogatására, irodalmi és tárgyi adatgyűjtés céljából, zárt gyűjteményként megalakult az intézet Adattára, illetve a terepi és feldolgozó munkák dokumentációs anyagának elkészítéséhez a katonai térképész végzettségű Seitl Kornél vezetésével a Műszaki Részleg.[3] Nevéhez kötődik az Intézet nagy horderejű munkáihoz kapcsolódó műszaki dokumentációs eljárások és irányelvek kidolgozása.[4] Az Adattár gyűjtőköre a kezdetektől fogva az interdiszciplináris kutatómunka dokumentációira, a kutatást segítő anyagokra irányul, ezeket hosszú távú megőrzés, kutatásra való rendelkezésre bocsátás, újbóli felhasználás és feldolgozás céljából nyilvántartásba veszi és tárolja. Az állomány kiemelt kulturális és közgyűjteményi értéket képvisel.

Régészeti Intézet Adattár

Az Adattár állománya az Intézet szerteágazó kutatásainak köszönhetően különböző forrástípusok rendkívül gazdag és folyamatosan gyarapodó gyűjteménye, amelyek a terepi munkák dokumentációs eljárása, illetve a feldolgozó kutatási projektek során keletkeznek. Az Adattár nagy mennyiségű hagyományos alapú dokumentumot őriz (papír, fénykép, diafilm, negatív stb.), ezeket kartotékrendszerben archiválja, papíralapú és részben digitális formátumú (Excel táblázat) nyilvántartást vezet. A hagyományos régészeti állományok kezelése kidolgozott archiválási rendszer szerint történik.

Az 1990-es évek végétől az Adattár gyűjteményfejlesztése a technológiai fejlődés részeként az egyre növekvő mennyiségű digitális formátumú állományokra is kiterjed.[5] A „hirtelen jött" technológiai váltás és

2 Ehhez bővebben Castiglione 1967, 87–89.

3 Az idézett közleményben Castiglione László a Műszaki Részleg munkájáról is összefoglaló jelentést közöl, Castiglione 1967, 102–103.

4 Az ásatások felmérését az országos alapponthálózathoz csatlakozva végezte el, ennélfogva a lelőhelyek a terep megváltozása után is pontosan visszaazonosíthatók, Virágh 1988.

5 A dokumentálásban ma már szinte a digitális technológia teljes térhódítása a jellemző, egyedi kérésre az intézet Grafikai Műhelyében azonban ma is készülnek papíralapú kézi rajzok, vázlatok. A terepi rajzos dokumentációk elsődlegesen továbbra is hagyományos formában kerülnek rögzítésre, majd ezeket később digitalizálják. Egyes forrástípusok ezért több különböző formátumban is létezhetnek.

a nagymértékű gyarapodás jelentős teherként nehezedett az Adattárra, a tárolási kapacitás folyamatos bővítésének kényszere került előtérbe, és eleinte kevesebb figyelem fordult a rendkívül sérülékeny és könnyen tünékennyé váló digitális információ hosszú távú megőrzésére. A digitális adat sok esetben technológiafüggő, a kezdetben használt adathordozók mára már elavultak, esetleges fizikai sérülésük, illetve a hardver- és a szoftvertechnológiák hiánya, inkompatibilitása miatt az adatok egy része nehezen vagy csak jelentős költségráfordítással nyerhető vissza. Mindeközben szembe kell néznünk azzal a helyzettel is, hogy a régészeti kutatás egyes területein (térinformatikai felmérések, LiDAR, 3D szkennelés, légirégészet, számítógépes modellezés stb.) jelentős állományt képviselnek a kivételesen nagy adatformátumok, amelyek hosszú távú megőrzése szintén az adatkezelési politika részét kell, hogy képezze.[6]

A hatalmas mennyiségű heterogén forrás digitális archiválása jelentős kihívást jelent koncepcionális és technikai vonatkozásban is.[7] Elsődleges feladatként fogalmazódott meg, hogy felgyülemlett adataink számára, amelyek valódi értéke az újrafelhasználásban rejlik, olyan technológiai lehetőségeket keressünk, amelyek hosszú távú, biztonságos elhelyezést és hatékonyabb tudománykommunikációs felületet biztosítanak.

A Hungarian Research Data Alliance és a Magyar Tudományos Akadémia Könyvtár és Információs Központ, valamint az Eötvös Loránd Kutatási Hálózat Titkársága által kiírt és támogatott „Kutatásiadat-archiválási pilot-projekt" pályázati programja, valamint a Concorda

6 A nagy adatformátumok kezelésére és megőrzésére az *Archaeology Data Service* (ADS) *„Big Data"* projektje nyújt ajánlásokat. A projekt záróbeszámolója ismerteti azokat a szervezeteket, akik hasznos dokumentációkat, referenciamodelleket, szemléltető példákat közölnek; némelyik link sajnos már nem él, vagy a tartalom másik felületre költözött, lásd http://ads.ahds.ac.uk/project/bigdata/final_report/bigdata_final_report_1.3.pdf.

7 „Based on past experience, we can argue that the archiving of digital data represents a task far more complex and costly than the archiving of analogue data. The possibilities of disseminating and harvesting digital data are so desirable and promising, that there is no other obvious solution than to allocate considerable resources to data digitisation and the on-going care for their preservation." Kuna et al. 2017.

felülete biztosította számunkra azt a lehetőséget, hogy rajzgyűjteményünkön keresztül a repozitóriumban való adatelhelyezés munkafolyamatát modellezzük.[8] A rendezetlen adatok kárba vesznek, letétbe helyezésük egyben értékmentés is.

Jogi háttér

A Bölcsészettudományi Kutatóközpont Régészeti Intézet keretei között folyó kutatómunka során létrejövő szellemi alkotásokhoz fűződő jog a Kutatóközpontot illeti meg.[9] Az Intézet jelenleg nem rendelkezik olyan adatkezelési szabályzattal, ami jól dokumentáltan segítené a szabványos adatrögzítés (fájlelnevezési konvenció, fájlformátum, tárolókra vonatkozó döntés, metaadatséma) és a szabályszerű adatnyilvántartás menetét. A stratégia kidolgozásáig a kutatási adatok tárolásáról és módozatairól esetí döntések születnek. Továbbra is érvényesül azonban az intézmény fennállása óta alkalmazott elv, miszerint a kutatási adatok dokumentációjának végső kezelője és nyilvántartója az Intézet adattára.

A Kutatóközpont egyes intézményeiben működő adattárak a Magyar Tudományos Akadémia tulajdonát képezik, melyek a Kutatóközpont használatában állnak. Az adattárak birtokában lévő kulturális javak felhasználásának engedélyeztetési eljárását, kutatási és közlési engedély vonatkozásában, a Kutatóközpont főigazgatójának 6./2021. számú utasítása szabályozza.

Régészeti Rajzgyűjtemény repozitóriuma

A Régészeti Rajzgyűjtemény adatkészlete az Adattár hagyományos technológiával készült rajzainak[10] azon állományrészét adja közre, amelynek digitális, raszteres formátumba való konvertálása (másodlagos

8 MTA KIK, témaszám: 61, projektkód: 21-61.

9 MTA BTK-T/1549/2013.

10 A rajzgyűjtemény hagyományos technikával készült darabjait az Adattár számozott fém rajzszekrényekben, egyedi számozású fiókokban, kiterített állapotban tárolja. Nyilvántartásuk egyedi leltári számos azonosítóval papíralapú leltárkönyvekben, illetve Excel táblázat formájában áll rendelkezésre; továbbá a rajzszekrényekről is nyilvántartást vezet. A hagyományos rajzállomány jelenlegi nagysága közel 47 000 darab.

digitális dokumentáció)[11] állományvédelmi és újrahasznosítási szempontból az elmúlt évtizedben megvalósult.[12]

A régészeti ábrázolások elsődlegesen műszaki rajzok, amelyek alapvetően nem azzal a céllal készülnek, hogy a látottakat művészi vagy realisztikus módon adják közre, hanem, hogy szakmai konvenciókat követve – grafikus információs nyelvezetet alkalmazva – tudományosan értelmezett leképezései legyenek a jelenségeknek.[13]

Az adatkészlet téma, műfaj és az alkalmazott technika tekintetében heterogén, térbeli és időbeli szóródása is jelentős. A rajzok egy része nyomtatott formátumban korábban közlésre került; egy részük azonban eddig még nem képezte tudományos kiértékelés tárgyát, vagy publikációkban még nem jelent meg.

Tartalmi vonatkozásaik alapvetően az Intézet előzményeként működő kutatócsoport tevékenységéhez kötődnek, az őskortól a kora újkorig terjedő teljes időpalettát lefedik. Az őskori kutatások többsége a nagyszabású dunakanyari feltárási kampány része volt, illetve az alföldi lelőhelyekre fókuszált; a kutatások tárgyát a települések melletti temetők és földvárak alkották. A dunakanyari munkálatok során számos római kori lelőhelyet is vizsgáltak, később a tartomány belseje, a Borostyánkő út térsége, egy-egy jelentősebb település is bevonásra került, ami katonai táborok, őrtornyok, villagazdaságok és városok dokumentációs rajzanyagát hívta életre. A középkori adatok zöme világi és egyházi

11 A „Digital Secondary Documentation" fogalmát az utólagosan digitalizált analóg állományra alkalmazzák, Aspöck et al. 2020, 93, fig. 9.

12 A rajzok szkennelése 300 dpi felbontásban, JPEG, esetenként TIFF formátumban a BTK Közös Adatbank és az MTA Infra pályázat terhére készült el. 2015 és 2016 folyamán a rajzok közül azok kerültek kiválogatásra, melyek szkennelése méreteiknél vagy technikájuknál fogva intézményi keretek között nem volt elvégezhető. A munkálatok külsős cégek bevonásával jelenleg is folynak, lehetőségeinkhez mérten a rajzszekrények állományának folyamatos és hiánytalan szkennelésére törekszünk.

13 Az a szempont érvényesül, amit a régész fontosnak tart, Banning 2020, 348, 367, a régészeti rajzokhoz további irodalommal.

központok, városok kutatása során keletkezett, amelyeknek a régészeti mellett jelentős művészettörténeti és építészettörténeti vetülete is van.

A Kutató Csoporthoz kötődik a hazai régészet legnagyobb volumenű vállalkozásának, az MRT-nek az elindítása 1962-ben. A nyomtatásban megjelent kötetek több típusú ábrát közölnek, az eredeti manuálék néhány darabja – leletközlő táblarajzok, összesítő térképek, szintvonalas lelőhelyfelmérések és különböző régészeti objektumok összesítő rajzai – szintén megosztásra kerültek.[14]

A kezdeti években számos előkészület történt „a magyar föld régészetének" összefoglalását nyújtó kézikönyv kapcsán, végül nyomtatásban csak egyetlen kötet jelent meg.[15] A kötethez készült adatgyűjtés és ábraanyag is a repozitórium adatkészletét gyarapítja.

A rajzos adathordozók műfaja szerteágazó, a terepi munkák során keletkezett ásatási helyszínrajzok, terepi vázlatok, geodéziai felmérések mellett az adatok másik forrását a feldolgozás során keletkezett tárgyrajzok, rekonstrukciók, részletrajzok, tematikus gyűjtések alapján összeállított táblarajzok, elterjedési térképek, grafikonok és egyéb illusztrációs anyagok adják.

Az adatállomány térbeli lefedettsége jelentős szóródást mutat. Az intézet kutatásai során keletkezett különböző képi állományok javarészt a mai Magyarország területére eső lelőhelyekről származnak, a tudományos feldolgozások adatgyűjtései azonban a Kárpát-medencére, valamint azon túl, a szomszédos országokra is kiterjednek, továbbá Európa távolabbi régészeti lelőhelyeire, valamint más kontinensre vonatkozó adatokat is tartalmaznak. A Kutató Csoport két jelentős nemzetközi vállalkozása a fenti régiókon is túlmutat. Az UNESCO szervezésében 1964-ben indult nemzetközi régészeti

14 A repozitóriumban jelenleg az MRT 2. kötet (Veszprémi járás), a 4. kötet (Pápai és Zirci járás), a 9. kötet (Szarvasi járás), illetve a 13. kötet (Aszódi és Gödöllői járás) illusztrációs anyagából találhatók rajzok.

15 Vértes 1965.

kampány nubiai munkálataiba a magyar expedíció is bekapcsolódott, ahol Abdallah Nirqi keresztény kori fellegvárát és városát kutatták. A jelentős feltárás rajzi dokumentációja is az adatkészlet részét képezi. A másik távol eső terület Mongóliához kötődik, ahol négy kampányban, hosszabb időintervallumot és több régészeti korszakot átfogó lelőhelyeken folyt feltárás. A dokumentációs- és tárgyrajzok a prákrit nyelvű sziklafelirat mellett hiung-nu és a kárpát-medencei korai avar leletekkel rokon emlékek adatait hordozzák.

Az adatok időbeli lefedettsége alapvetően a korai neolitikumtól a kora újkorig (17. század vége) terjed, a természettudományos vizsgálatokhoz kapcsolódóan azonban modernkori adatgyűjtés is történt, ezek közé tartozik többek között a hagyományos típusú fazekaskemencék, kézi őrlőkövek, fazekaskorongok felmérési rajzainak gyűjteménye, ami a régészeti mellett jelentős néprajzi értéket is képvisel.

Az adattartalom nyelve – feliratok, jelmagyarázatok, egyéb megjegyzések – többnyire magyar, a publikációs rajzok idegen nyelvű közlése esetén főként angol, német vagy orosz.

A rajzokon a grafikusok és geodéták mellett, akik alkotói értelemben a rajzok tényleges szerzői, a kezdeti időszakban a tudományos adattartalom gazdájaként a régészek, természettudósok neve is feltüntetésre került, később ezeket az adatokat csak a leltárkönyvi nyilvántartásba vezették be. A leltárkönyv a készítés időpontját is rögzíti és az időközi revíziók dátumát is bepecsételték vagy beírták.

Repozitálás lépései

A repozitálás első fázisa az adatmodellezés előkészítése, amelyet a raszteres képfájlok és az Excel formátumban tárolt nyilvántartás esetében is, miután korábban ilyen irányú előkészület nem történt, szinte a kezdeti lépésektől kell indítanunk. A fájlok nevezéktanának kidolgozását és átnevezését[16] követően új mappastruktúra kerül kialakításra, amely a hagyományos állomány fizikai tárolási

16 RIA(Régészeti Intézet Adattár)_RK(Rajz kézi)_00000(leltári szám)

rendszerének feleltethető meg.[17] Az adatmodellezést számos tényező nehezíti, az Excel alapú nyilvántartás adatait új szabványkategóriákkal szükséges kiegészítenünk, továbbá gondot jelentenek a meglévő adatoknál előforduló hiányosságok, illetve az adatelőkészítéssel párhuzamosan állományrevízió és adattisztítás zajlik.[18] A metaadatok bevitelére metaadatblokkok szerint felépített Excel-sablont használunk. A folyamat legnagyobb erőforrást igénylő része az egyedi állományok szisztematikus, régészeti szempontú leíró elemzése, a nyilvántartási jellegű leltárkönyvi leírások ugyanis nem felelnek meg a korszerű repozitálás során alkalmazandó szempontoknak.

A hazai repozitóriumok közül választásunk a Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI) által fejlesztett, Harvard Dataverse típusú[19] Concorda repozitóriumra esett. Ennek keretein belül alakítottuk ki a *Régészeti Intézet Adattár,* alatta a *Régészeti Rajzgyűjtemény* tárolóját.[20] Az adatkészlet a repozitóriumban nyílt hozzáféréssel van elhelyezve, amelyhez az Intézet honlapja linken keresztüli hozzáférést biztosít.

A repozitórium adatkészleteket/adatcsomagokat kezel, amelyeknek központi eleme a megőrzendő objektum, ehhez kapcsolódnak a hosszú távú megőrzést szolgáló metaadatok. A rajztár adatállományának heterogeneitásából eredően a fájlok egyedi letétbe helyezése mellett döntöttünk, ezért az adatcsomagok száma kiemelkedően magas, egy-egy adatcsomaghoz azonban csak kisszámú objektum társul. A repozitálás során a tömeges adatcsomag-feltöltéshez és a csatolandó fájlok azonosítón keresztüli hozzárendeléséhez folyamodtunk.[21]

17 RSZ(rajzszekrény)1_F(fiók)1–F25, RSZ2_F1–F25...RSZ6_F1–F25

18 A *Rajznyilvántartás* az alábbi mezőket tartalmazza: ltsz (leltári szám), perszám, anyag, technika, cím, lépték, rajzoló, kutató, negltsz (fotónegatív leltári száma), elhelyezés, megjegyzés.

19 A Dataverse típusú repozitóriumok előnye, hogy nagy mennyiségű, könnyen kereshető adatot képesek tárolni, az összes Creative Commons (https://creativecommons.org/) licencet támogatják, szükség esetén az objektumokhoz egyedi licenc rendelhető, verziókövetést biztosítanak, illetve az adatok a teljes szövegben indexelve vannak, lásd https://dataverse.org/

20 Elérhető: https://science-data.hu/dataverse/ria_rajztar

21 Ahhoz, hogy adatcsomagjainkat API-n keresztül Curl-parancssorral fel tudjuk tölteni, az Excel-sablon soraiból segédprogram segítségével egyedi JSON-fájlokat hoztunk létre, amely külső segédprogramot és informatikai segítséget igényelt.

Az adatkészlet elemeinek ismertetése és összevetése a DANS Data Station Archaeology szerkezetével

A világhálón közzétett nagy mennyiségű adat kezelésének jelenlegi leghatékonyabb módja a gépi értelmezhetőséget is biztosító metaadatolás. A metaadatok feltárják számunkra annak az információhalmaznak a belső szerkezetét, amit az adatokhoz való közvetlen hozzáférés nélkül is tudni érdemes.[22] A metaadatok nélküli adatokról nem tudjuk, hogy mit tartalmaznak, honnan és kitől származnak, hogyan lehet hozzájuk férni és melyek az újrafelhasználás feltételei.

A régészeti rajzgyűjtemény metaadat-struktúrájába való rövid betekintés mellett elengedhetetlenné vált, hogy az általunk felépített szerkezetet a *Data Archiving and Networked Services* által frissen ismertetett és újonnan kialakított *Data Station Archaeology* (DSA) tárolóban alkalmazott rendszerével összevessük,[23] amely szintén a Dataverse szoftverét használja.[24]

A Régészeti Adattárban őrzött állományaink archiválását illetően gazdag fájlszintű dokumentálás mellett döntöttünk, metaadat-sémáinkat a különböző adattípusainkhoz igazítjuk.[25]

A jelenlegi adatcsomagjainkhoz kialakított metaadat-struktúra a rendszeren belül létrehozott három metaadatblokk mezőiből – *Citation, Geospatial* és a *Social Science and Humanities* – a számunkra releváns elemeket tartalmazza; illetve tervbe vettük egy, a régészet leíró nyelvezetének

22 A metaadat-struktúra kidolgozásához a TSV-fájl felépítéséről a *Dataverse* felhasználói útmutatójában tájékozódtunk: https://docs.google.com/spreadsheets/d/13HP-jI_cwLDHBetn9UKTREPJ_F4iHdAvhjmlvmYdSSw/edit#gid=0

23 A tároló a https://archaeology.datastations.nl/ címen érhető el.

24 DANS honlapján 2022. június 24-én tette közzé Helga Hollander a *Data Station Archaeology* tároló ismertetését, bővebben https://dans.knaw.nl/nl/nieuws/dans-data-station-archaeology-is-nu-live/. A szervezet régészeti adatkezelési elveire részletesebben a bevált gyakorlatoknál fogunk kitérni. A DSA felületére az adatokat maguk a kutatók helyezik el, amelyben az adatfeldolgozási csapat segítségükre van.

25 Ez az eljárás megfeleltethető az ADS által alkalmazott gyakorlatnak: https://archaeologydataservice.ac.uk/advice/Downloads.xhtml

megfelelő egyedi metaadatblokk kialakítását és rendszerszintű fejlesztését. A DSA a *Citation* mellett öt másik blokk mezőit alkalmazza, melyek saját fejlesztései: *Rights, Relation, Archaeology-Specific, Temporal and Spatial Coverage,* illetve *Data Vault Metadata.* A két rendszerben sok a hasonlóság, néhány elem esetében azonban eltérő kategóriákra esett a választásunk.

A Dataverse rendszerében az alapmezők kivételével, mint a *Title, Notes, Language, Depositor,* a többi elem annyiszor ismételhető, ahányszor azok az adatkészlet leírásához szükségesek.

Az egyes kategóriák az alábbi információkat hordozzák:

Citation_title – a mező az adatkészlet rövid megnevezését, címét a lelőhely neve[26] – lelőhely típusa – dokumentáció műfaja sorrendben közli.[27] A lelőhelyek típusainak megadásakor a Magyar Nemzeti Múzeum Régészeti *Archeodatabase* adatbázisának szabványos szószedetére támaszkodtunk.[28] A DSA felületén a lelőhely jellegének jelölésére az *Archaeology-Specific_subject* kategóriát használják, amely a Kulturális Örökségvédelmi Ügynökség ABR nevezéktanán alapul,[29] a dokumentáció műfaji jellege az ő esetükben is a címben szerepel.

Citation_otherIdAgency és *Citation_otherIdValue* mezőkben a tartalomgazda intézmény és szervezeti egység, valamint a rajzok egyedi azonosítója (leltári száma) szerepel.

Citation_dsDescriptionValue – a leíró mezőben három külön elem jelenik meg, az adatkészlet összefoglaló, leíró jellegű ismertetése, a kiegészítő

26 Abban a formában, ahogy az adattári nyilvántartásunkban szerepel.

27 A címek egyelőre magyar nyelvűek, az angol nyelvű címeket hamarosan feltöltjük, a magyar nyelvű címet a *title_hu* mező hordozza.

28 A magyarországi régészeti lelőhelyek legfontosabb adatait tartalmazó adatbázis szabványos szószedetre vonatkozó ajánlása az útmutató mellékletében található: https://archeodatabase.hnm.hu/sites/default/files/attachments/2016/06/Lelohely_feltoltesi_utmutato_2016.06.21.pdf

29 *Archeologisch Basisregister* (ABR), lásd: https://data.cultureelerfgoed.nl/term/id/abr.html; továbbá Willems –Brandt 2004.

adatok és a kormegjelölés.[30] A kiegészítő adatok az adatfelhasználó számára adnak tájékoztatást arról, hogy az adatkészlet milyen mértékben alkalmas újrafelhasználásra, például van-e a rajznak léptéke vagy mércéje, tartalmaz-e magasság- vagy mélységadatokat, helymeghatározáshoz szükséges földrajzi koordinátákat, felszíni rajzoknál északi irány jelölést, jelmagyarázatot, feliratokat és egyéb megjegyzéseket. Hosszú távon a kormegjelölésre önálló kategóriát kívánunk alkalmazni, amelyre a szabványos szószedetek alapján több megoldás is kínálkozik: *Coverage* (Dublin Core, idő és tér is), *contentReferenceTime* (schema.org).[31] A DSA-ban a tér és az idő adatainak bevitelére egy önálló blokk szolgál *(Temporal and Spatial Coverage)*, emellett a régészetspecifikus metaadatok között is található egy *Temporal* mező.

Citation_keywordValue – a mező kulcsszavai az adatkészlethez tartozó régészeti jelenségeket veszik sorra és támogatják az adatkészletben való hatékony keresést.

Citation_notesText – ebben a mezőben olyan járulékos információk találhatók, mint a méretadatok: rajzok kiterjedése és méretaránya, technikai jellemzők: hordozóanyag és az alkalmazott technika. Továbbá itt kaptak helyet az adatelhelyezésre vonatkozó tudnivalók is, amelyek megadják az újrafelhasználás és a nyomdai követelmények feltételeinek megfelelő, teljes méretű képfájl mappájának azonosítóját.[32]

A pilot-projekt során felépített rendszerben kettéválasztottuk az adatfájl szerzőjét (adatfeldolgozó), ez a *Citation_producerName* és *Citation_producerDate*, aki lehet grafikus, geodéta vagy akár régész is, illetve a kutatót *(Social Science and Humanities_dataCollector)*, akihez a tartalmi adatok felgyűjtése, értelmezése vagy újraalkotása kötődik. A nemzetközi gyakorlat az alkotási folyamat szereplői között többnyire nem tesz ilyesfajta megkülönböztetést, az *Author or creator* mező

30 A kormegjelölésnél ugyancsak az *Archeodatabase* adatbázis szószedetét használjuk, korszak és alkorszak nem került feltüntetésre.

31 A *Citation* metaadatblokk *TimePeriodCovered* mezője számunkra nem alkalmas, mert év–hónap–nap formátumban van definiálva, amely a régészeti koroknál nem használható.

32 Mérlegeljük, hogy a mappaazonosítót *otherID* mezőre módosítjuk.

reprodukálásával rögzíti mindkét adatot.[33] A DSA esetében a tartalomgazda személyének megjelölésére a *Citation_authorName*, az intézmény esetében a *Citation_distributorName* mezők szolgálnak.

Az adatgyűjtés típusát[34] a *Social Science and Humanities_samplingProcedure* mezőnél adtuk meg, ez esetleg a *Citation_kindOfData* mezővel kiváltható, vagy követni lehetne a DSA által választott metódust, az *Archaeology-Specific_methodsOfRecovery* mezővel.

A *Citation_relatedMaterial* kategória alá helyeztük el a rajzok repró felvételeinek nyilvántartási számát, illetve itt hivatkoztunk arra, ha két rajz között kontextuális kapcsolat áll fenn. Az adatkészlet minden egyes eleméhez kigyűjtésre került az *Archeodatabase*, mint a legnagyobb hazai lelőhelyadatbázis, lelőhelyre mutató linkje,[35] amely lehetővé teszi, hogy a jövőben közvetlen kapcsolatot létesítsünk a két független adatkészlet között. Ez az adat az előbbi mező ismétlésével hozható nyilvánosságra. A DSA tárolóban a *Relation Metadata Relation* mezője hordozza ezt az információt, illetve a blokk kiegészül az *Audience* kategóriával,[36] amelyet a saját adatkészletünk szempontjából is hasznosnak találunk.

Az adatkészlet lényegi elemét képezi a földrajzi adatokra vonatkozó rész, amelyhez a *Geospatial Metadata* mezőit használjuk. A *geographicCoverage* egyesíti az ország, megye,[37] járás,[38] település adatokat. A *geographicUnit* kategóriában, amely az adatállomány által lefedett földrajzi egység legalacsonyabb szintjére, vagyis esetünkben a lelőhelyre

33 Lásd https://intarch.ac.uk/journal/issue2/wise/part2.html
34 Feltárási adat, feldolgozási adat, mérési eredmény stb.
35 A linkek hozzárendelése folyamatban van.
36 A tudományos közeg megjelölése, amely számára az adat releváns információkat hordoz.
37 Átmeneti jelleggel *state* kategóriába helyeztük el a magyarországi lelőhelyek megyék szerinti besorolását, ez azonban ellentmondáshoz vezet, ezért át kell gondolnunk.
38 Némileg zavaró, hogy a közigazgatási hierarchiában a járás magasabb rangú a településnél, a rendszerben az *otherGegrapohicCoverage* kategória, amely lehetőséget ad arra, hogy az adatot hozzáfűzzük a helyadatokhoz, azonban hátrébb szerepel.

vonatkozik, két adat szerepel: a lelőhely megnevezése,[39] illetve a lelőhely különféle dokumentációkban, publikációkban és forrásokban szereplő névváltozatai.

Az adminisztratív metaadatok tartalmazzák az adatrögzítőt, a tartalomgazdát, az adathoz való hozzáférés módját és feltételeit.

Az adatkészlet adatait közlő publikációk összegyűjtését *(Citation_publication)* a projekt keretein belül nem tudtuk lezárni, a hiányzó adatok kiegészítése időigényes, a hátramaradt feladatok között kiemelt helyen szerepel. Az adatpótlás a közlemény-repozitóriumokhoz való kapcsolódási lehetőség mellett azért is fontos, mert a közzétett állományok újrafelhasználási engedélye semmiféle jogi akadályba nem ütközik, a közöletlen adatok esetében azonban körültekintően kell eljárnunk. Intézményünkben a nyomtatott formában még meg nem jelentetett, de feldolgozás alatt lévő állományok esetében párhuzamos kutatási engedély kiadása nem támogatott.

Az adatkészletekhez tartozó képfájlokat, az MTA kulturális javainak felhasználására vonatkozó rendelkezésekkel összhangban, vízjellel ellátva, csökkentett méretben helyeztük el a tárolóban. A raszteres nézőképek ugyan újrafelhasználásra, nyomdai előkészítésre nem alkalmasak, a tartalmat illetően azonban megfelelő tájékoztatást nyújtanak. Érvényes felhasználási engedély birtokában az Adattár a jó felbontású, teljes méretű állományokat a kutatók rendelkezésére bocsátja. A szerzői jogra és felhasználásra vonatkozó információs sablont a *Feltételek* mezőben helyeztük el.[40]

A repozitóriumban alkalmazott metaadatelemek tehát a közzétett objektumok tartalmi vonatkozásai mellett, a létrehozáshoz, kezeléshez és használathoz kötődő szempontokat írják le és feltárják a

39 Az információkhoz az *Archeodatabase* (https://archeodatabase.hnm.hu) nyújt hozzáférést, a lelőhelyek adatai részben a Miniszterelnökség központi, közhiteles hatósági nyilvántartásából származnak.

40 A DSA-ban a *Rights Metadata* tartalmazza a *Rights Holder*t.

lehetséges adatkapcsolódási pontokat.[41] A leíró metaadattípusok esetében szabványos szószedetet a régészeti kor és lelőhely típusának megjelölésére használtunk.

Bevált gyakorlatok az európai régészetben

A folyamatosan növekvő információk hatékony kezelése, legyen az analóg vagy digitális, mindenütt jelentős kihívás elé állítja a megőrzéssel és tudásközvetítéssel foglalkozó intézményeket.[42] Számos nemzetközi intézet és szervezet rendelkezik egyénileg jól felépített, szakmaspecifikus archiválási rendszerrel, melyekről iránymutatásokat és ismertetéseket tesznek közzé. Nincs és objektív okoknál fogva nem is lehet általánosan alkalmazandó európai standardot bevezetni, az egyetemes elvek azonban megfogalmazásra és lefektetésre kerültek.[43] A néhány európai állam régészeti archívumaiban, hatósági nyilvántartásaiban alkalmazott gyakorlatokba való betekintéssel az a célunk, hogy tanulságokat és mérlegelendő szempontokat gyűjtsünk a megőrzési politika fejlesztése, az adatkezelési szabványok kidolgozása terén.[44]

A yorki székhelyű *Archaeology Data Service* a régészeti örökség digitális archiválását, megőrzését és terjesztését illetően irányadó szerepet

41 Ehhez bővebben https://www.getty.edu/publications/intrometadata/setting-the-stage/

42 A digitális archiválással kapcsolatban részletes útmutatás található az OAIS (nyílt archívumi információs rendszer) referenciamodellben, ami 2022. március 1-től a Magyar Szabványügyi Testület honlapján keresztül magyar nyelven is elérhető (MSZ ISO 14721:2022), lásd https://ugyintezes.mszt.hu/Publications/Details/178190

43 Az Európai Bizottság által támogatott Culture Programme keretében 2007 és 2013 között végzett ARCHES projekt eredményeként lefektetésre kerültek az általános irányelvek, lásd lentebb.

44 Az Internet Archaeology 2021. évi különszáma a régészetben alkalmazható digitális archiválás kérdéskörét járja körül a SEADDA (*Saving European Archaeology from the Digital Dark Age*) workshop résztvevőinek különböző országokra vonatkozó tanulmányainak segítségével, https://intarch.ac.uk/journal/issue58/index.html. A részletes tájékoztatást nyújtó közlésekből csak azokat a legfőbb elemeket emeljük ki, amelyek a további tájékozódást segíthetik.

tölt be.[45] A szervezet a kutatási adatok repozitóriumi tárolása mellett gyakorlati segítséget nyújt a régészet legkülönbözőbb területein alkalmazható helyes gyakorlatokhoz. A gazdag metaadatolás gyakorlatát követik, metaadatsémáik nyilvánosak és szabadon felhasználhatók. Jelenlegi fejlesztésük arra irányul, hogy az adatkezelési folyamat teljes fázisát – a tervezéstől, az adatgyűjtésen és az archiváláson át az adatgondozásig – metaadatokkal lássák el,[46] ehhez iránymutatásokat és stratégiai dokumentumokat tesznek közzé.[47] A stratégia fontos elemét képezi a használt metaadatkészletek és a leírásokban alkalmazott szabványos szószedetek és tezauruszok közzététele.[48] Az ADS a különböző adattípusoknál önálló metaadatsablonokkal dolgozik,[49] az alapkategóriák mellé az adattípustól függően egyedi szabványokat rendelnek hozzá.[50] Az ADS rendszere rendkívül sokféle régészeti adatrepozitóriumot, adatbázist egyesít, a kutatási adatok célirányos felgyűjtéséhez a metaadatrekordok indexelése révén nyílt hozzáférésű tartalmai között az *Archsearch* felületen egyablakos keresztkeresést biztosít.[51] Az ADS *Library* oldala számos, nyílt hozzáférésű publikációt tesz közzé, melyek között a helyes adatkezeléssel behatóan foglalkozó ismertetők és javaslatok is helyet kapnak.[52]

45 1996-ban nyolc angol egyetem Régészeti Tanszéke és a *Council of British Archaeology* által felállított non-profit szervezet, elérhető: https://archaeologydataservice.ac.uk/

46 Bővebben: https://archaeologydataservice.ac.uk/about/strategyStandards.xhtml.

47 https://archaeologydataservice.ac.uk/about/endpoints.xhtml.

48 https://archaeologydataservice.ac.uk/about/strategyStandards.xhtml.

49 Elérhető: https://archaeologydataservice.ac.uk/advice/Downloads.xhtml.

50 Önálló metaadatsablont használnak a raszteres és a vektoros képfájlokhoz, a vektoros állományok esetében a dokumentáció a *Supporting documentation file name(s)* kategóriával egészül ki, ahol a képhez köthető kódokat, rövidítéseket és terminológiát rögzítő dokumentumokat tüntetik fel.

51 Elérhető: https://archaeologydataservice.ac.uk/archsearch/

52 Szabadon elérhető a Duncan H. Brown által összeállított kötet, ami a régészeti adattárak különböző állományainak kezelésére vonatkozó ajánlásokat a legnagyobb részletességgel tárgyalja. Különösen hasznos a kötet irodalomjegyzéke, amiben állományok szerinti csoportosításban szerepelnek a releváns publikációk, https://archaeologydataservice.ac.uk/library/browse/issue.xhtml?recordId=1137506&recordType=MonographSeries

Az angliai *Historic Environment Records* (HER) régészettel és történelmi épített környezettel kapcsolatos hatósági nyilvántartás és információs szolgáltatás.[53] Az Egyesült Királyságban az örökségekkel kapcsolatos adatok dokumentálását központi szabvány alapján végzik,[54] amit a *MIDAS Heritage* rögzít.[55] Az irányelveket több örökségvédelmi szervezet együttműködésében fejlesztették ki, legkorábbi verziója 1998-ban jelent meg, a jelenlegi 2012 óta hatályos. Névtérhasználatuk a *UK e-Government Metadata Standard* szabványának felel meg,[56] ami a *Dublin Core*-on alapul. A HER az indexelésre kifejlesztett FISH *(Forum on Information Standards in Heritage)* tezaurusz hierarchikus szószedetére támaszkodik, amit az ADS is alkalmaz.

Az Európai Bizottság által támogatva és több ország – Belgium, Csehország, Hollandia, Izland, Nagy Britannia, Németország, Svájc és Svédország – tapasztalatait felhasználva az *Europae Archaeologiae Consilium* (EAC) gondozásában 2014-ben jelent meg a kilenc különböző nyelven elérhető régészeti archiválásra vonatkozó európai útmutató.[57] A Bizottság irányzata nem ír elő kötelező szabványokat, sokkal inkább elveket fogalmaz meg,[58] mivel az egyes európai államok hatósági és kutatási struktúrája, valamint annak törvényi háttere jelentősen eltérhet egymástól. Útmutatójukban az adatok teljes életciklusát végigkövetik, gyakorlati segítséget nyújtanak a projekttervezéstől a letétbe

53 https://historicengland.org.uk/advice/technical-advice/information-management/hers/

54 Bővebben: https://historicengland.org.uk/advice/technical-advice/information-management/data-standards-terminology/

55 Lásd: https://historicengland.org.uk/images-books/publications/midas-heritage/midas-heritage-2012-v1_1/

56 Elérhető: https://cdn.nationalarchives.gov.uk/documents/information-management/egms-metadata-standard.pdf

57 Az útmutató az ARCHES projekt eredményeként született meg, https://www.europae-archaeologiae-consilium.org/_files/ugd/881a59_dc8871c3c9d84100a17ac3b763a7f407.pdf, a projektről részletes ismertetés található az ADS oldalán: https://archaeologydataservice.ac.uk/arches/

58 Az európai régészeti archiválási szabványokhoz és útmutatókhoz lásd: WP7 - Ensuring the Sustainability.

helyezésig.[59] A digitális archiválásra vonatkozó szabványokat behatóan nem tárgyalják, de megadják azokat a forráshelyeket, ahol ezek elérhetők.[60] A digitális régészeti adatkezeléssel kapcsolatban az ARCHES projekt oldalán, országonkénti bontásban, részletes bibliográfiai gyűjtéssel szolgálnak.[61]

Adatkészletét és személyi állományát tekintve is Európa egyik vezető intézménye a Holland Tudományos és Művészeti Akadémia (KNAW) és a Holland Örökségvédelmi Intézet *(Rijksdienst voor het Cultureel Erfgoed)* tudományos repozitóriumát, a holland kutatási adatok tárházát kezelő *Data Archiving and Networked Services* (DANS) szervezet.[62] A tudományos adatok egységesítése, közzététele mellett széles körű adattudományi, archiválási, gyűjteménygondozási tanácsadással is foglalkoznak. Az általuk alkalmazott DataverseNL, a Concordához hasonlóan, a Harvard Dataverse szoftverét használja. A repozitóriumos tárolás előfeltételeként a különböző típusú állományok esetében pontosan megadják a preferált és a nem preferált formátumokat, ezekhez bőséges tájékoztatást nyújtanak.[63] Hollandiában a régészeti munkafo-

59 Az adatkezelési szabályzat kidolgozása során a fenti útmutató számunkra is mérvadó sorvezetőként szolgál, különösen hasznos kiindulási alapot jelent az általuk kidolgozott ellenőrzési lista, ami az archiválási folyamat lépéseit, a felelősöket és a hozzájuk kapcsolódó feladatköröket rögzíti. A lista az ARCHES projekt ADS oldalán: https://archaeologydataservice.ac.uk/arches/Wiki.jsp?page=CHECKLIST%20OF%20ARCHAEOLOGICAL%20ARCHIVING%20TASKS%20AND%20ROLES

60 Perrin et al. 2014, 39. további hivatkozásokkal.

61 https://archaeologydataservice.ac.uk/arches/Wiki.jsp?page=BIBLIOGRAPHY

62 https://dans.knaw.nl/nl/

63 Azokat a formátumokat részesítik előnyben, amelyeket sokan használnak, nyitottak, továbbá platform- és fejlesztői környezet függetlenek, lásd: https://dans.knaw.nl/nl/bestandsformaten/. Az általunk használt ISO szabványon alapuló JPEG formátum a támogatottak között van, ami elsősorban fényképes állományok tárolására szolgál. A formátum 32 bites színmélységet és hatékony tömörítési algoritmust kínál a színminőség romlásával. A JPEG lehetővé teszi az EXIF metaadatok (a képfájlban tárolt, a kép rögzítésének körülményeit leíró metaadatok) integrálását a fájlba. Amennyiben a színhűség elvárás, a JPEG2000 formátum „veszteségmentes" tömörítését javasolják. A JPEG2000 XML-t használ a metaadatok tárolására, bővebben lásd: https://dans.knaw.nl/nl/bestandsformaten/afbeeldingen-raster/jpeg/

lyamatokra a BRL SIKB 4000 minőségbiztosítási irányelvei vonatkoznak,[64] az ásatási adatokat a terepmunkát követő 2 éven belül a nemzeti tárhelyként funkcionáló DANS rendszerében letétbe kell helyezni. A DANS keretein belül új programként fut a *Data Station Archaeology*, amit kifejezetten a régészet rendkívül heterogén adattípusainak kezelésére fejlesztettek ki, a tárolóban elhelyezett régészeti adatok 99%-a közvetlenül és nyilvánosan hozzáférhető.[65]

Az Osztrák Tudományos Akadémia (ÖAW) keretein belül 2015 és 2020 között végzett *A Puzzle in 4D* című projekt az egyiptomi Tell el-Daba (Avaris) lelőhelyen 1966 óta folyó ásatások dokumentációs anyagát (analóg és digitális fényképek, hagyományos és digitális rajzok, írott feljegyzések) archiválási szempontból dolgozta fel, ami egyben esettanulmányként szolgált az ÖAW régészeti tárolójának kialakításához.[66] A projekt honlapja közzéteszi a digitalizálási munkafolyamatot, a megőrzési szabályzatot, ismerteti az adatkészletet, az adattípusok szerinti metaadatsablonokat és a repozitálásra használt ARCHE archívum dokumentációját. Az általuk kezelt forrásállomány a mi adattári erőforrásainkhoz nagyon hasonló, analóg és digitalizált, valamint digitális adatokat, illetve több változatban létező dokumentumokat egyaránt gondoznak. Az angol nyelvű Dokumentációs Archívum a kutatási adatokat, köztük a terepi rajzokat, adattípusok szerinti csoportosításban kezeli. Adataik modellezésére a CIDOC CRM referenciamodellt használják, ami a kulturális örökség fogalmaival és információival, valamint a múzeumi dokumentációval kapcsolatban bővíthető ontológiát kínál (ISO 21127:2014).[67] Az összetett adatkapcsolatokat kezelni képes,

64 https://www.sikb.nl/archeologie/richtlijnen/brl-sikb-4000

65 A korábban alkalmazott EASY régészeti gyűjteménye az új tárolóba beépítésre került, https://dans.knaw.nl/nl/data-stations/archaeology/

66 Aspöck et al. 2020, 79–100; a projekt honlapja a https://4dpuzzle.orea.oeaw.ac.at/ címen érhető el.

67 További részekkel egészült ki CRMarcheo, CRMsci (tudományos megfigyelések) és a CRMdig (digitális eredet), melyek révén képes azokat az összetett kapcsolatokat ábrázolni, amelyek a távoli múltban végzett tevékenységek maradványainak feltárásakor és dokumentálásakor, valamint a dokumentáció digitalizálásakor keletkeznek, Aspöck et al. 2020, 93; lásd: http://www.cidoc-crm.org/collaborations

adatbázisokból építkező rendszer az ÖAW Fedora alapú ARCHE repozitóriumában kapott helyet.

A Cseh Köztársaságban a tömeges digitalizálást két jelentős, magát a régészetet is rendkívül hátrányosan érintő katasztrófa idézte elő.[68] A régészeti adatkezelés adatbázisokon keresztül történik, a prágai régészeti intézethez kapcsolódik a *Archaeological Database of Bohemia* (ADB).[69] A 2000-es évek végén valósult meg az *Internet Database of Archaeological Fieldwork*, ami olyan további jelentős adatbázisokkal egészült ki, mint például a brnoi intézeté, a *Digital Archive and Evidence of Archaeological Excavations in Moravia and Silesia (DAEAE)*, majd ezeket továbbiak követték. A különböző technikai paraméterű adatbázisokat egységesítési célzattal 2012 óta az AMCR *(Archaeological Map of the Czech Republic)*[70] integrálja, amit az *Archaeological Information System of the Czech Republic* rendszere üzemeltet.[71] A platform funkciója kettős, adminisztratív nyilvántartás mellett kutatási adatszolgáltatást végez, az adatokat metaadatok és szabványosított szószedetek strukturálják.[72] Az AIS CR felület gyakorlatban is jól hasznosítható eleme a webes felületű cseh, angol és német nyelvű tezaurusz, a TEATER *(Thesaurus of Archaeological Terminology)*, ahol a bejegyzések teljes szöveges kereséssel és hierarchikus csoportosításban is lekereshetők és JSON formátumba importálhatók.[73] A Prágai Régészeti Intézet adattárában található dokumentumok metaadatokkal leírt formában digitalizálva vannak, illetve a Digitális Adattárba a brnoi intézet anyagának is jelentős része bekerült.[74] A két digitális

68 A 2002-es prágai árvíz, majd 2007-ben a mikulčicei bázison pusztított tűz után tudatosult igazából, hogy a régészeti leletek és adatok olyan kulturális értéket képviselnek, amit a jövő számára meg kell őrizni, a digitalizációs folyamat részletes elemzéséhez, Novák – Kuna – Lečbychová 2021.

69 Kuna et al. 2017.

70 https://www.aiscr.cz/en/

71 https://www.aiscr.cz/en/#smooth-scroll-top, bővebben Kuna et al. 2017.

72 Általánosan a zárolt szószedetet részesítik előnyben, amit az ADB rendszeréből örökítettek át, csak kellően indokolt esetben módosítják, https://intarch.ac.uk/journal/issue43/10/table1.html

73 http://teater.aiscr.cz/

74 A Digitális Archívumhoz részletes felhasználói kézikönyvet tesznek közzé, https://digiarchiv.aiscr.cz/napoveda

adattár AMCR keretrendszerbe való integrációjával a dokumentumok indexelésére, illetve a szerteágazó adatkapcsolati háló feltérképezésére törekednek.[75] Az AMCR a cseh régészet sikeres és látványos vállalkozása, az egyéb régészeti adattárakat kezelő szervezetek vonatkozásában azonban már messze nem ennyire jó az általános helyzet.[76] A hatalmasra duzzadt adattömeget egyre nehezebb kezelni és csak a szervezetek 2%-a hoz létre metaadatok szerint strukturált leírásokat. A szervezetek közel felénél az analóg dokumentáció tárolására léteznek szabványok, valamivel kevesebb intézmény rendelkezik a digitális adatok tárolására vonatkozó irányelvekkel is, azonban 40%-nak egyik területre sincs szabványa. A digitális adatkezelés általános elfogadtatásának nehézségei ellenére folyamatos erőfeszítéseket tesznek annak érdekében, hogy a kutatói közösséggel elfogadtassák a digitális adatok szabványos gondozásának szükségességét.

Szlovákiában két intézmény kezeli a régészeti archívumot a Szlovák Tudományos Akadémia Régészeti Intézete (IA SAS) és a Szlovák Köztársaság Műemléki Tanácsa (MB SR).[77] Az utóbbi szervezet államigazgatási célból archiválja a régészeti tevékenységekhez kapcsolódó jelentéseket, amit a Szlovák Köztársaság Régészeti Lelőhelyeinek Központi Nyilvántartása (CEANS) számára is átadnak, ennek kezelője az akadémia Régészeti Intézete. Az intézet kifelé téradatokat és metaadatokat szolgáltat. A digitális állományok a Központi Adatarchívum (*Centrálny dátový archív,* Pozsonyi Egyetem Könyvtár) repozitóriumában is elhelyezésre kerülnek. 2019-ben indult a Műemléki Információs Rendszer (PAMIS), ami az Európai Kohéziós Alapból támogatott GIS alapú projekt, célja a műemléki és régészeti nyilvántartások közös online platformjának kialakítása az államigazgatás, a kutatás és a nagyközönség számára.

75 https://digiarchiv.aiscr.cz/home

76 Az AIS CR csapata 2020-ban 169 szervezetet keresett meg kérdőíves felmérésével, amelyhez 114 intézmény szolgáltatott adatot, részletesen lásd: Novák – Kuna – Lečbychová 2021.

77 Bisták et al. 2021.

Szlovéniában a 2000-es évek elejétől az autópálya-régészet tömeges adattermeléssel járt, digitális adattárak sora született.[78] A legutóbbi helyzetelemzés a digitális adatkezelés szabványosítását azonban elégtelennek ítéli meg,[79] a hatályos utasítások nem írják elő az archiválási módszereket, hiányoznak a részletes iránymutatások és az adatgondozásra vonatkozó szabványok. Az előállított digitális archívumok állandó mappastruktúrán alapulnak, Access és Excel formátumú nyilvántartásokkal dolgoznak. A Múzeumok nem rendelkeznek digitális adatkezelési szabályzattal. A jelenlegi legnagyobb adatkészletet, Szlovénia Régészeti Kataszterét (ARKAS), egy 1993-ban felépített, azóta változatlan struktúrájú adatbázis kezeli, az adattartalom 2004-től GIS adatokkal egészült ki.[80] Emellett több, adatbázisokon alapuló projekt is fut, melyek közül van néhány, amelynek adatkészlete webes felületen is elérhető. A régészeti adatkészleteknek egy országos online nyilvántartása létezik (*Register nepremične kulturne dediščine Republike Slovenije,* RNKD), az 1997 óta élő projekt platformja interaktív térképet és böngészőt nyújt a keresésekhez.[81]

Végül nézzük a hazai gyakorlatot. A Miniszterelnökség Építészeti, Építésügyi és Örökségvédelmi Helyettes Államtitkárságán működő Régészeti Főosztály egyik legfontosabb feladatának a Magyarország területén található és nyilvántartott régészeti lelőhelyek folyamatos, minden részletre kiterjedő védelmét és felügyeletét tekinti, ellátja a régészeti lelőhely központi, közhiteles nyilvántartásának kezelését. A hatóság kérelemre adatot szolgáltat, a nyilvántartás a szakma számára előzetes jóváhagyást követően hozzáférhető.[82] A régészeti tevékenységek elsődleges dokumentációs eljárása jól szabályozott, a jelentéseket a kijelölt hivatalokhoz és múzeumokhoz meghatározott

78 A 2002-ben Krško Poljében lezajlott több nagy ásatásra készülve létrehozták a szlovéniai régészeti ásatások digitális dokumentációs rendszerének gerincét, amelynek jelentős fejlesztései 2008 körül zajlottak. Ennek kezdeti fejlesztését a Ljubljanai Egyetem Régészeti Tanszéke végezte.

79 Štular 2021.

80 http://arkas.zrc-sazu.si/index.php

81 https://www.gov.si/teme/register-kulturne-dediscine/

82 https://www.e-epites.hu/regeszet

időn belül be kell nyújtani.[83] A közhiteles nyilvántartás a kötelezően benyújtandó jelentések metaadatait tartalmazza, a dokumentumok azonban nincsenek velük összekötve, illetve a régészeti tevékenységekhez kötődő egyéb állományok adatkészletében metaadat szintjén sem jelennek meg. Magyarországon jelenleg nincs olyan központosított adattár vagy szervezet, ahol a régészeti tevékenységekkel kapcsolatban minden információs anyagot tárolnának és azt közzé is tennék.[84] A múzeumok és intézmények önálló adattárakat kezelnek, a digitális adatkezelésnek nincs szabványa, mindössze egy 2019-ben megjelent útmutatás áll a szakma rendelkezésére.[85] A Magyar Nemzeti Múzeum lelőhelyadatbázisa az Archeodatabase megpróbálja ezt az űrt kitölteni,[86] lehetővé teszi a dokumentumok méretkorlátozás nélküli, metaadatokkal szabványosított formában való tárolását és online elérését.[87] A felület metaadatai ellenőrzött módon bővíthetők, emellett új lelőhelyek rögzítését is megengedi, illetve a szakma számára biztosítja az adatjavítás lehetőségét. A tárhelyen való adatelhelyezés önkéntes alapú, törvényi kötelezettség nem írja elő.[88] Régészek számára, intézményi regisztráción keresztül, a teljes felület és dokumentumtartalom, nem regisztrált felhasználók számára a metaadatok érhetők el. A szakmai szemlélet megváltozása sokat tehet azért, hogy az Archeodatabase adatai mind pontosabbá váljanak. A közvetlen adatkapcsolat megteremtését a Régészeti Intézet repozitóriuma és a Magyar Nemzeti Múzeum Régészeti Adatbázisa között mi magunk is fontosnak tartjuk.

A kutatási adatok korszerű kezelése terén kritikus tényező a szabványos névterek, a leíró metaadatokban az ellenőrzött szókincsek használata; továbbá az adatok teljes életciklusának nyomon követése, adatmodellek

83 A jelenlegi szabályozás szerint az Építésügyi és Örökségvédelmi Hivatalhoz, a Miniszterelnökség Régészeti Főosztályához, a Magyar Nemzeti Múzeumhoz, az illetékes megyei múzeumhoz és a leleteket befogadó múzeumhoz, ha az nem a megyei múzeum.

84 Bővebben Kreiter 2021.

85 Kómár – Bánki 2019, http://www.oszk.hu/kds-k/feher-konyv

86 https://archeodatabase.hnm.hu/hu

87 A projektet az ARIADNE és az ARIADNEplus támogatta, lásd Kreiter 2019.

88 A szemléletmód lassú átalakulására vannak biztató jelek, Kreiter 2021.

felállítása és az adatkezelési elvek lefektetése sem odázható sokáig. A digitális erőforrások kezeléséhez új kompetenciákat kell elsajátítanunk és ez nem csupán a néhány, közvetlenül az adatokat kezelő kollégákat érinti, hanem a széles szakmai közeget is, miután az adatok előkészítésében jelentős szerep hárul rájuk. A programok sikere sokban függ attól is, hogy a szakmai intézmények hosszú távon milyen jelentőséget tulajdonítanak ezeknek a feladatoknak.

A régészeti tárgyak elkallódhatnak, sérülhetnek, előfordul, hogy egy tárgy rajza vagy fotója az egyetlen olyan hiteles dokumentum, ami a további kutatás számára fennmarad. A feltárások során nyert régészeti bizonyítékok eleve a primer adatok roncsolásából, megsemmisítéséből születnek, a terepi jelenségeket nem lehet újból megalkotni, az ásatásokon rögzített rajzi, fényképes és szöveges dokumentumok az elveszett adatforrás egyedüli reprodukciói. Archiválásuk központi kérdés. Az intézet alapítói tudatában voltak annak, hogy a dokumentációk jelentősége a tárgyakéval vetekszik, ezért is létesült az Adattár az Intézettel egyidőben. A kezdetektől fogva létező kutatószolgálat a mindenkori tudásmegosztást szolgálta, a módszerek a közel 65 év alatt azonban sokat változtak, a digitális technológiák korszerűsödése új lehetőségeket kínál a jelenkori tudástárházak kialakítása terén.

Irodalom

Aspöck et al. 2020
Edeltraud Aspöck, Gerald Hiebel, Karin Kopetzky, Matej Ďurčo, *A Puzzle in 4D: Archiving Digital and Analogue Resources of the Austrian Excavations at Tell el-Daba, Egypt* = E. Aspöck, S. Štuhec, K. Kopetzky, and M. Kucera (eds.), *Old Excavation Data. What Can We Do? Proceedings of the Workshop held at 10th ICAANE in Vienna, April 2016.* Oriental and European Archaeology Series 16 (2020), 79–100.
https://doi:10.1553/0x003bca0e

Banning 2020
Edward Bruce Banning, *Archaeological Illustration and Publication* = E. B. Banning, *The Archaeologist's Laboratory. The Analysis of Archaeological Evidence. Interdisciplinary Contributions to Archaeology*. Cham, 2020. https://doi: 10.1007/978-3-030-47992-3_21

Bisták et al. 2021
Peter Bisták, Ján Zachar, Alexandra Rášová, Tibor Lieskovský, Ivica Kravjanská, Martina Orosová, Kristína Kročková and Michal Felcan, *Archaeological Digital Archiving in Heritage Management in Slovakia.* Internet Archaeology 58. https://doi.org/10.11141/ia.58.16

Brown 2011
Duncan H. Brown, Archaeological Archives: *A guide to best practice in creation, compilation, transfer and curation*, 2011.

Castiglione 1967
Castiglione László, *A Magyar Tudományos Akadémia Régészeti Kutató Csoportjának munkájáról (1958–1965),* Az MTA Filozófiai és Történettudományi Osztályának Közleményei 15 (1966–1967) [1967], 87–110.

Kómár – Bánki 2019
Kómár Éva – Bánki Zsolt (szerk.): *Fehér Könyv. Módszertani útmutató a közgyűjteményi kulturális örökség digitalizálásához és közzétételéhez,* Budapest, 2019.

Kreiter 2019
Kreiter, Attila, *The Hungarian archaeology database in the light of ARIADNE* = Julian D. Richards, Franco Niccolucci (eds), *The ARIADNE Impact,* Budapest, 2019, 63–68. https://doi.org/10.5281/zenodo.3476712

Kreiter 2021
Kreiter, Attila, *The Hungarian Archaeology Database*, Internet Archaeology 58.
https://doi.org/10.11141/ia.58.9

Kuna et al. 2017
Martin Kuna, David Novák, Jan Hasil, Dana Křivánková, *Archaeological Map of the Czech Republic. Current state and future visions of virtual research tools in the Czech Republic*, Internet Archaeology 43.
https://doi.org/10.11141/ia.43.10

Novák – Kuna – Lečbychová 2021
David Novák, Martin Kuna and Olga Lečbychová: *Taming the Beast. Approaches to Digital Archiving in Czech Archaeology.* Internet Archaeology 58.
https://doi.org/10.11141/ia.58.5

Perrin et al. 2014
Kathy Perrin, Duncan H. Brown, Guus Lange, David Bibby, Annika Carlsson, Ann Degrave, Martin Kuna, Ylva Larsson, Sólborg Una Pálsdottír, Battina Stoll-Tucker, Cynthia Dunning, Auréle Rogalla von Bieberstein, *A standard and guide to best practice for archaeological archiving in Europe.* EAC Guidelines 1, Namur 2014.

Štular 2021
Benjamin Štular, *Archiving of Archaeological Digital Datasets in Slovenia: historic context and current practice*, Internet Archaeology 58.
https://doi.org/10.11141/ia.58.17

Virágh 1988
Virágh Dénes, *Régészetünk térképei*, Térképvilág 1988 (oldalszámok nélkül).

Vértes 1965
Vértes László, *Az őskor és az átmeneti kőkor emlékei Magyarországon. A magyar régészet kézikönyve 1,* Budapest 1965.

Willems – Brandt 2004
Willem J.H. Willems, Roel W. Brandt, *Dutch Archaeology Quality Standard.* Den Haag, 2004.

Audio kazettáról mesterséges intelligencián alapuló algoritmusba

Veszélyben lévő kutatási adatok megóvása – beszámoló egy pilot projektről és az eredmények további sorsáról

Egyed-Gergely Júlia[1]
ORCID: 0000-0003-1905-0027

Jakab Miklós[1]

Meiszterics Enikő[1]

[1] ELKH Társadalomtudományi Kutatóközpont
Kutatási Dokumentációs Központ

Absztrakt

A Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központjában (TK KDK) a folyó kutatások adatkezelési igényeinek és az adatok másodfelhasználóinak kiszolgálása mellett intenzív archiválási munka is zajlik, hiszen a KDK és a keretei között működő 20. Század Hangja Archívum és Kutatóműhely az elmúlt 50 év társadalomtudományos kutatási anyagait gyűjti. A Hungarian Research Data Alliance (HRDA) és a Magyar Tudományos Akadémia Könyvtár és Információs Központ (MTA KIK) által szervezett, az Eötvös Loránd Kutatási Hálózat (ELKH) Titkársága által támogatott adatarchiválási pályázatnak köszönhetően az intézmény munkatársainak lehetősége nyílt a KDK digitális tárházát bővíteni. A projekt keretében két értékes, a digitális korszak előtt készült, a kutatói- és a nagyközönség számára eddig nehezen elérhető társadalomtudományos kutatási anyag vált – a FAIR alapelveket követve – kutathatóvá. Ezzel, a két vizsgálat publikált eredményein túl, most már azok háttéranyagai, az azokban gyűjtött adatok, készített interjúk is megtekinthetővé, újrafelhasználhatóvá, másodelemezhetővé, az újabb kutatási eredményekkel összevethetővé, közkinccsé váltak.

A két archivált anyag ráadásul, immár digitalizált formában, a KDK és a SZTAKI közös MILAB „Computational Archival Science" projektjébe is bekerült. A projekt célja bővíteni a KDK-ban elérhető nagy mennyiségű kutatási dokumentum metaadatkészletét a digitálisan olvasható interjúk gépi tanulás és mesterséges intelligencia segítségével történő, tárgyszavakkal való ellátásával.

Az alábbi tanulmány a HRDA pilot projekt megvalósulását és a TK KDK archívumába bekerült új kutatási anyagok további sorsát mutatja be.

Bevezető

A Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központja két adatrepozitóriumot működtet, amelyek kutatási adatok és dokumentációk gondozásával, elérhetővé tételével foglalkoznak. A KDK-repozitórium a TK négy kutatóintézetének kvalitatív és kvantitatív módszerekkel készült anyagait tárolja (pl. interjúfelvételeket, leiratokat, vezérfonalakat, kérdőíves felmérések kérdőíveit, módszertani leírásokat, adatbázisokat, terepnaplókat, jegyzőkönyveket) különféle formátumokban (pl. szöveg, kép, video). A 20. Század Hangja Archívum és Kutatóműhely az elmúlt hatvan év kvalitatív módszerekkel készült, szociológiatörténetileg is meghatározó kutatásainak anyagait gyűjti írott, hangzó és képes dokumentumok formájában. A két repozitórium anyagait leíró metaadatokat bárki szabadon böngészheti, magukhoz a letétbe helyezett kutatási adatokhoz, dokumentációkhoz az elhelyező döntésének függvényében a KDK repozitóriumában szabadon, regisztrációval vagy egyedi kutatói engedéllyel, a 20. Század Hangja Archívumban regisztrációt követően lehet hozzáférni.

Az archívumok állománya folyamatosan bővül – új és évtizedekkel ezelőtt befejezett kutatások anyagaival egyaránt. Utóbbiakra azért is helyezünk különös hangsúlyt, mert az idő előrehaladtával ezek az anyagok egyre veszélyeztetettebbekké váltak, válnak, folyamatosan veszítenek eredeti minőségükből, ráadásul a technológiai fejlődés miatt a használatukat lehetővé tévő technikai eszközök, berendezések is egyre elérhetetlenebbek, tönkremennek, megsemmisülnek.

A HRDA és az MTA KIK pályázati kiírása kiváló lehetőség volt az ilyen régi anyagok digitalizálásra és közkinccsé tételére. A projekt keretében a KDK két, a múlt század 80-as, 90-es éveiben végzett kutatás audio kazettán lévő interjús hanganyagát tette hozzáférhetővé repozitóriumából – digitalizálás, leiratozás és metaadatokkal való ellátás után.

A két kutatásról

Kovách Imre Kuczi Tiborral közösen folytatott *A helyi társadalom és a mezőgazdaság átalakulása a rendszerváltás idején* című kutatásának témája a termelőszövetkezetek felbomlása a 80-as, 90-es években, valamint a társadalmi-gazdasági változásokra adott válaszlehetőségek, az érintett magánemberek és „háztájizók" megoldási kísérletei. Kovách és Kuczi a vizsgálathoz interjúkat készítettek mezőgazdasági termelőkkel és kistermelőkkel két hullámban és két térségben. Az adatfelvételek első hulláma a 80-as évek közepén, második a 90-es évek elején zajlott, a helyszínek a Dunántúl (Bajna és Epöl), illetve Hajdú-Bihar megye (Hajdúnánás és Hajdúböszörmény) voltak. A kutatás eredményei megjelentek Kovách Imre *A jelenkori magyar vidéki társadalom szerkezeti és hatalmi változásai* című MTA doktori értekezésében (Kovách 2010), valamint a Kovách Imre és Megyesi Boldizsár által közösen jegyzett *A vidék harminc éve. A magyar vidék alakulása az erőforrások, a társadalmi tőke és fejlesztéspolitikai változásainak tükrében* című tanulmányban (Kovách-Megyesi 2018) is. A háttéranyagok eddig nem voltak könnyen elérhetőek az érdeklődők számára, a kutatás során készített interjúk audio kazettán és hagyományos írógéppel készített leiratok formájában várták további sorsukat.

Kovács Éva 1995 és 1998 között szlovákiai vegyes házasságokat vizsgált a *Kulturális csere és etnikai identitásváltozások a vegyesházasságokban a XX. századi kassai népesség példáján* című, Fejős Zoltán által vezetett kutatásban Gyurgyík László, Vasik János, Kádek Kata és Németh Szilvia kutatókkal közösen. A vizsgálatban adatbázisok segítségével, valamint interjúk készítésével és feldolgozásával elemezték a vegyes házasságok demográfiai vonatkozásait 1920-tól kezdődően. Az 1920 és 1991 közötti

időszakra adatbázist hoztak létre a Csehszlovák Statisztikai Hivatal, majd a Szlovák Statisztikai Hivatal által közzétett kiadványok vegyes házasságokra vonatkozó (részben a házasságkötések és válások nemzetiségi bontása, részben a születési adatsorok szülők nemzetiségi bontása szerinti) adataiból, valamint a népszámlálások családi nemzetiségi összetételre vonatkozó adataiból (Gyurgyík 1999, Kovács 2003). Az ezt követő időszakra nem állt rendelkezésre statisztikai adatbázis, a kutatók így részben feltevéseikből indulhattak ki, részben más módszerekben bízhattak. A vizsgálat – egyetemi hallgatókat is bevonva – a házasság és a családi élet témájára összpontosító narratív interjúk készítésével folytatódott. Az interjúk alapján később esettanulmányok készültek a vegyes házasságok életrajzi, társadalmi és családi hátteréről, a családi „magtörténet" narratíváiról, kulturális cseréről, családi traumákról, az identitáspolitika mechanizmusairól és stratégiáiról.[1] A Komáromban készült interjúkat – a kutatás idejének megfelelő technikával – audio kazettán rögzítették.

Kapott anyagok

Kovách Imre vizsgálatából 27 darab 60 perces és 5 darab 90 perces magnókazettán a két hullám adatfelvételeinek hanganyagát és a papíralapú interjúleiratokat kaptuk meg. A projekt indításakor nem lehetett tudni, hogy van-e átfedés a hangzó és az írott anyagok között.

Kovács Éva kutatásából a Komáromban készített interjúk hanganyagát tartalmazó 5 darab 90 perces és 16 darab 60 perces audio kazettát kaptuk meg, valamint ezek mellett papír alapú statisztikai táblákat, rövid elemzést, digitálisan hozzáférhető kutatási jelentést, valamint interjúleiratokat és biográfiákat. Az átvételkor ebben az esetben sem állt rendelkezésre információ arról, hogy van-e, és amennyiben van, milyen mértékű az átfedés a hanganyagok és a leiratok között.

1 Forrás: Társadalomtudományi Kutatóközpont, Kutatási Dokumentációs Központ repozitórium, Fejős-Kovács-Gyurgyík-Vasik-Kádek-Németh kutatási gyűjtemény, absztrakt, https://openarchive.tk.mta.hu/199/

Projektcélok

A projekt célkitűzése a fenti kutatások audio kazettán lévő hanganyagainak és papír alapú dokumentációinak digitális formába történő átírása, a teljes anyag FAIR alapelveknek megfelelő archiválása, valamint a KDK repozitóriumában történő elhelyezése és kutathatóvá tétele volt.

A megvalósítás fázisai

Az audio kazetták vonatkozásában a feladat a két kutatás interjúinak a lehető legjobb kondíciókkal történő hangalapú digitalizálása és leiratozása volt. A hangrögzítés régi 60, illetve 90 perces magnókazettákon történt, a rögzítés mikéntjét, körülményeit, az átvételi jellemzőket nem ismertük. Szakértőkkel egyeztetve a hang digitalizálására és szerkesztésére az Audacity[2] ingyenes és minden platformra elérhető hangszerkesztő programot választottuk. Az Audacity használata egyszerű, menüje, eszköztára átlátható és magyar nyelvű. Az interneten sok, a program használatát a gyakorlatban bemutató videó található. A digitalizált hanganyag leiratozása a Régens Zrt. Alrite[3] programjával történt.

A papíralapú gépelt leiratokat és egyéb papíralapú kutatási dokumentációkat szkennelés után OCR (optikai karakterfelismerés) technológiával, ABBYY FineReader[4] programmal véglegesítettük.

A hanganyag Audacityvel történő digitalizálása

A digitalizálás állandó jelenlétet kívánt, előfordult ugyanis, hogy megszakadt a felvétel, más volt a szalagon, esetleg nem volt rajta semmi. A hangszerkesztő programmal a felvétel közben erősíthettük vagy éppen halkíthattuk a hangot, az optimális hangerősséget digitális kijelzőn követtük. A felvételt WAV tömörítetlen fájlformátumba, majd

2 Audacity hangszerkesztő program, https://www.audacityteam.org/
3 Alrite beszédfelismerő és -leiratozó program, https://alrite.io/ai/hu/
4 ABBYY FineReader karakterfelismerő program, https://pdf.abbyy.com/

– mivel a WAV formátum nagyméretű fájlokban tárolja a digitalizált hanghullámképet, és az Alrite leiratozó programnál szét kellett volna darabolni emiatt az interjúkat – MP3 tömörített fájlba mentettük el. Az Audacityvel való utómunka során a következő korrekciókat végeztük:

- zajcsökkentés (szalagzaj, motorzajok stb.),
- lemezpattogás, torzítások kiszűrése, javítása,
- hangbalesetek (pl. kutyaugatás, mikrofon ütögetése) nyomtalan eltávolítása,
- érthetőség optimalizálása.

A legfontosabb a zajcsökkentés mértékének megfelelő meghatározása volt, ugyanis – egy bizonyos határ után – miközben a zaj csökken, a hasznos jel egyre növekvő mértékű torzulásnak indul. A munkálatok közben sajnos kiderült az is, hogy a felvételek spektrálisan sérültek, ami adódhatott a kazetta minőségéből, a felvétel módjából, eszközéből, idejéből (ami a legvalószínűbb, hiszen a kazetták közel 40 évesek), a tárolásból vagy a többszöri lejátszásból.

A meghallgatás során a felvételeket beazonosítottuk, az egy interjúhoz tartozókat összefűztük, minden interjút metaadatokkal láttunk el. Ezzel lehetővé vált a hangzó interjúk és az átvételkor kapott leiratok összevetése is. A vidék átalakulását vizsgáló kutatás interjúi egy részénél találtunk egyezést, ezek esetében a hanganyagot összekötöttük a papíralapú gépelt leirattal. A komáromi hanganyagok és a kapott digitális leiratok között nem volt átfedés.

A digitalizált interjúk leiratozása az Alrite beszédfelismerő programmal

Az Alrite mesterséges intelligenciára épülő, magyar nyelvre optimalizált beszédfelismerő megoldás, amely napjaink korszerű technikákkal készült hangfelvételeit akár 95%-os pontossággal képes leiratozni. A kapott régi kazetták leiratozása közel sem érte el ezt a technikai szintet, mindössze 10-15%-os pontosságú volt. Szakértő bevonása

céljából felvettük a kapcsolatot a Magyar Rádió egyik hangmérnökével, aki kérésünkre speciális szoftverekkel (MAGIX SEQUOIA[5] 16 és CEDAR Audio[6]) két rövid – ötperces – tesztanyagot készített. A javított hanganyagokon már hallás után azt tapasztaltuk, hogy bár az egyébként is hallható, érthető szövegrészek minősége valóban javult, a nehezen érthető vagy érthetetlen szövegrészek továbbra is nehezen érthetőek, illetve érthetetlenek maradtak, az Alrite használata után pedig továbbra is jelentős mértékben javításra szoruló leiratváltozatokat kaptunk.

A gépi leiratokat ezért csupán támpontként használhattuk, a szöveg nagy részét hallás alapján egészítettük ki, illetve gépeltük be.

A program dolgát nehezítette továbbá, hogy az alanyok sokszor hadartak, tájszólással beszéltek, illetve előfordult, hogy többen (házastársak, szomszédok) egyszerre szólaltak meg, amiket a program nem tudott szétbontani. Gyors beszéd esetén próbáltuk lassítani a felvételt, de ez sem hozott kielégítő eredményt. A tapasztalat az, hogy annál pontosabb a leirat, minél jobb az interjúalany artikulációja – ilyen felvételeknél a régi kazetták esetében is jobb lett a leiratozás minősége, találkoztunk olyan interjúval, amelynél 60%-ban tudta a program hibátlan szöveggé alakítani a hanganyagot.

A hallás alapján való javításnál további nehézségbe is ütköztünk. Amikor az alanyok szakszöveget használtak, szótár segítségét kellett igénybe venni, nemcsak a mesterséges intelligencia nem értette a szöveget, mi sem. Utána kellett nézni az olyan, agronómus interjúalany által használt kifejezéseknek például, mint a *meliorációs munkák*, vagy a *szilázs–szenázs* szópár. Előbbi a talajjavító munkákat, utóbbi a silózásnál a takarmány tartósításának különböző nedvességtartalmát jelenti. Más esetben, amikor a válaszadó az adott város utcaneveit sorolta, a Google-térkép volt a segítségünkre.

5 MAGIX SEQUOIA hangszerkesztő program: https://www.magix.com/int/
6 CEDAR Audio hangszerkesztő program: https://www.cedar-audio.com/

A projektben egy 1 órás régi hanganyag leiratának kiegészítése, javítása jó esetben 3, rossz esetben 6-7 órába telt, az átlag 4 óra volt.

A gépelt leiratok, egyéb kutatási dokumentációk digitalizálása OCR technológiával, ABBYY FineReader programmal

A kapott gépelt leiratokat és egyéb kutatási dokumentációkat OCR technológiával, az ABBYY FineReader program segítségével digitalizáltuk. A szövegek viszonylag jó minőségűek és „tiszták" (kézzel írt bejegyzésektől mentesek) voltak, így az OCR technológiával készített digitális változat esetében a kapott szöveg pontossága – az Alrite mai technológiával felvett interjúk leiratozásának pontosságához hasonlóan – 90% feletti volt. Ez nagy könnyebbséget jelentett azoknál a hangzó interjúknál, amelyeknél rendelkezésünkre állt a leirat gépelt változata is.

Megmentett kutatási anyagok a jelen és a jövő számára

A fent bemutatott technológiák és technikák alkalmazásával a két kiválasztott vizsgálat anyagainak digitalizálása megvalósult, ezzel azok megmenekültek az elkallódás és az olvashatatlanná, hallgathatatlanná válás veszélyétől. Az anyagok gyűjteményekbe rendezve, DOI-val ellátva bekerültek a KDK repozitóriumába, ahonnan a kutatásokat és kutatási adatokat leíró metaadatok, az archiválás módjának leírása, a kutatásokhoz kapcsolódó publikációk, illetve a statisztikai táblák és a módszertani leírások minden érdeklődő számára korlátozások nélkül hozzáférhetőek.

Kovách Imre gyűjteménye a https://openarchive.tk.mta.hu/496/ címen,[7] Kovács Éva gyűjteménye pedig a https://openarchive.tk.mta.hu/199/ címen[8] tekinthető meg.

7 KOVÁCH Imre, KUCZI Tibor: A helyi társadalom és mezőgazdaság átalakulása a rendszerváltás idején. [Kutatási gyűjtemény], https://www.doi.org/10.17203/KDK496.

8 FEJŐS Zoltán, KOVÁCS Éva, GYURGYÍK László, VASIK János, KÁDEK Kata, NÉMETH Szilvia: Kulturális csere és etnikai identitásváltozások a vegyesházasságokban a XX. századi kassai népesség példáján. [Kutatási gyűjtemény], https://www.doi.org/10.17203/KDK199.

Az interjúk leiratait csak a teljes nevek szintjén anonimizáltuk, meglátásunk szerint a települések, városrészek, foglalkozások anonimizálása olyan információvesztést okozna, amely ennyi idővel az adatfelvételek után már nem indokolt. Részben emiatt, részben pedig az interjúkban felmerülő szenzitív témák (pl. a településeken élő szlovák vagy zsidó közösséggel való viszony leírása) miatt magukat az interjúkat nem tettük mindenki számára olvashatóvá, azok kutatói hozzáféréssel érhetőek el. Ez azt jelenti, hogy a Társadalomtudományi Kutatóközpont kutatói korlátozás nélkül, szabadon férnek hozzá az anyagokhoz, külsős kutatók, egyetemi hallgatók pedig igényük jelzése után, regisztrációs folyamatot követően válhatnak felhasználókká.

A digitalizált anyagok máris új munkába állnak

A KDK-repozitóriumok fejlesztésének részeként a KDK 2020-ban elkezdett foglalkozni a mesterséges intelligencia kínálta eszközök társadalomtudományos archívumokban történő hasznosíthatóságával. A lehetőségek feltérképezéséből fejlesztési irány lett, amelynek megvalósításában az újonnan digitalizált interjúk is szerepet kapnak. A projekt célja a TK KDK két archívuma, a KDK és a 20. Század Hangja Archívum és Kutatóműhely állományaiban a nagyobb fokú áttekinthetőség és a komplexebb kereshetőség biztosítása a kutatási dokumentumok, elsősorban az interjúk metaadatainak gazdagításával. Az új metaadatok egységes tárgyszólistából származó tárgyszavak társításával, illetve a dokumentumokban szereplő névelemek (személynevek, földrajzi nevek, intézménynevek) és időelemek (dátumok, korszakok, ünnepek) azonosításával kerülnek a korábbiak mellé – mesterséges intelligenciát is használó, tanuláson alapuló algoritmusok segítségével.

A KDK a célok eléréséhez a Számítástechnikai és Automatizálási Kutatóintézettel (SZTAKI) együttműködve a Mesterséges Intelligencia Nemzeti Laboratórium (MILAB) kutatási programjának finanszírozásával, 2020-ban pilot projektet indított, majd 2021-ben, azt folytatva, komoly előkészítő és fejlesztő munkába fogott. A munka során a KDK és a SZTAKI munkatársai a manuális és a gépi szövegfeldolgozás

módszertanának kidolgozásával, egységes társadalomtudományos tárgyszólista meghatározásával, többféle kulcs- és tárgyszavazási módszer kipróbálásával, tesztelésével és validálásával, annotáló program segítségével történő manuális interjúkódolással, majd gépi tárgyszavazási módszerek alkalmazásával végezték a fejlesztőmunkát.

A feladathoz több különböző kutatási gyűjteményünk anyagát is bevontuk, részben a tanításhoz, részben a kapott eredmények ellenőrzéséhez. A folyamat során kiválasztott (legmegfelelőbbre értékelt) algoritmussal a tanulóhalmaz kézi kódolása alapján valamennyi digitálisan elérhető interjúnk anyagához tárgyszavakat rendeltetünk, illetve azokban névelemeket, időelemeket jelöltetünk ki, ezzel fejlesztve és egységesítve repozitóriumaink metaadatkészletét és bővítve a bennük való keresési lehetőségeket.

A módszer alkalmazása annál hatékonyabb, minél nagyobb és minél változatosabb szövegkorpuszon tudjuk tanítani és tesztelni a SZTAKI kutatói által fejlesztett algoritmust, így a munkába több típusú, különböző módszerrel készült, eltérő témájú és mélységű interjúkat vonunk be. A HRDA pilot projekt keretein belül digitalizált két kutatási anyag nagymértékben hozzájárul a MILAB-fejlesztés sikeréhez, speciális témáikkal, nyelvezetükkel, az azokban használt szakkifejezésekkel, a csak azokban előforduló név- és időelemekkel.

A KDK MILAB projektje 2022 őszén érkezett ötödik fázisába. Az eredmények, az új tárgyszó-hozzárendelések, a névelem- és időelem-kiemelések hamarosan a KDK repozitóriumainak keresőfelületein is megjelennek majd (a többi metaadatot kiegészítve), ezzel támogatva a kutatók és a nagyközönség kutatómunkáját, az archívumokban való kiigazodást, az adott kutatási kérdések szempontjából releváns szövegek, szövegrészek megtalálását.

Összegzés

A HRDA és az MTA KIK pályázatának köszönhetően két értékes kutatási anyaggal bővült a TK KDK archívuma, amelyek így elérhetőek és kutathatóak lettek. A projekt a KDK munkatársai számára lehetővé tette különböző új, korábban nem használt technikák kipróbálását, amelyek egy részét azóta is alkalmazzuk archívumaink napi gyakorlatában.

A munkának köszönhetően nehezen hozzáférhető analóg kutatási anyagok váltak néhány hónap alatt mesterséges intelligenciát alkalmazó kutatás szövegkorpuszának részévé.

Irodalomjegyzék

Gyurgyík 1999
Gyurgyík László, *A szlovákiai vegyes házasságok demográfiai vonatkozásai 1949-től napjainkig*, In.: Fórum Társadalomtudományi Szemle 1991/1 pp. 5–18.
https://epa.oszk.hu/00000/00033/00001/gyurgyik.htm

Kovách 2010
Kovách Imre, *A jelenkori magyar vidéki társadalom szerkezeti és hatalmi változásai*. Akadémiai nagydoktori thesis, MTA Politikai Tudományok Intézete
http://real-d.mtak.hu/296/

Kovách-Megyesi 2018
Kovách Imre – Megyesi Boldizsár, *A vidék harminc éve. A magyar vidék alakulása az erőforrások, a társadalmi tőke és fejlesztéspolitikai változásainak tükrében*, In: Erdélyi Társadalom 16 (1), 2018,
https://doi.org/10.17177/77171.209.

Kovács 2003
Kovács Éva, *A „házassági piac" alakulása Komáromban (1900–1940)*, In.: K. Horváth Zsolt – Lugosi András – Sohajda Ferenc (szerk.): Léptékváltó társadalomtörténet, Hermész Kör – Osiris: Budapest, pp. 366–394.
https://www.academia.edu/7769693/A_h%C3%A1zass%C3%A-1gi_piac_alakul%C3%A1sa_Kom%C3%A1romban_1900_1940_

A mellékletek mellékesek? Digitális mellékletek vizsgálata az Akadémiai Kiadó folyóirataiban

Smid Dávid[1]
ORCID: 0000-0003-4484-8591

Böhm Gabriella[1]

[1]Akadémiai Kiadó

1. Bevezetés

Az elmúlt években a tudományos világ egyik legmarkánsabb mozgatórugója kétségkívül az *open science*, azaz a *nyílt tudomány* szerteágazó, de minden ágában ugyanazt a világos célt, a tudományos eredmények szabad áramlását szolgáló eszméje (vö. OECD, 2015). Világszerte egyre több tudományos kiadó igyekszik megfelelni a nyílt tudomány kihívásainak, működésének átláthatóbbá tételével és minél gazdagabb információforrások szolgáltatásával.

A nyílt tudomány kezdeményezések legismertebb összetevője talán a kutatási eredmények *nyílt hozzáférésű* (*open access*) megjelentetése (Chan et al., 2002), ami többet jelent, mint az olvasók korlátozás nélküli hozzáférése a megjelent cikkekhez. Fontos eleme, hogy a közölt eredmények szabadon, külön engedély nélkül legyenek felhasználhatók további kutatásokhoz, közleményekhez (természetesen az eredeti közleményre való pontos hivatkozással).

A közlésre alkalmas tudományos dolgozatok kiválasztásának mára egyedül elfogadott módja a kutatótársak szakértő bírálatán alapszik. A nyílt tudomány a bírálati folyamatok áttekinthetőségéről is megfogalmaz irányelveket. Ma még nem elterjedt, de egyre több folyóirat alkalmazza a *nyílt bírálati* formát, amikor a cikkekkel együtt az elfogadásuk alapjául szolgáló bírálatokat is közlik, akár a bíráló aláírásával, akár anélkül.

A tudományos publikálás átláthatóságának része a folyóiratok működésére vonatkozó széles körű tudománymetriai adatok közzététele is. A *nyílt metrika* (San Francisco Declaration on Research Assessment, n. d.) nemcsak egy folyóirat cikkeinek idézettségéről tájékoztat, hanem online (interneten) megjelent cikkek esetében megtekintésük, letöltésük gyakoriságáról is, éppúgy, mint például a folyóirathoz benyújtott cikkek elbírálására és megjelentetésére fordított átlagos időről vagy a közlésre elfogadott és elutasított dolgozatok arányáról.

A kutatási eredményeket bemutató tudományos cikkeknek hagyományosan nem része a vizsgálatokhoz felhasznált adatok közzététele. A nyílt tudománynak azonban lényeges alappillére a kutatási adatok megosztása, amelyre a tudományos cikkek online megjelenése technikai lehetőséget is biztosít. Az Open Knowledge Foundation kézikönyvének (Open Data Handbook, n. d.) definíciójával *nyílt adat* az, amit bárki szabadon használhat, újrahasználhat és terjeszthet – legfeljebb csak hivatkozási és megoszthatósági követelményeknek eleget téve. A nyílt adatok segítik a felhasználásukkal végzett kutatások független ellenőrzését és reprodukálását, ezzel a tudomány megbízhatóságát és átláthatóságát (vö. OECD, 2015). A kutatási adatok szabad áramlása pedig gyorsítja és megkönnyíti az új tudományos felfedezéseket. A kiadók ébredő válaszát erre a kihívásra jól mutatja Rousi és Laakso (2020) friss kutatási eredménye, amely szerint egyes tudományos kiadók a kutatási adatok megosztására vonatkozó szabályozást is egyre inkább feladatuknak tekintik.

Kutatási adatok többféle (internetes) platformon is megoszthatók. Léteznek külön erre a célra szolgáló repozitóriumok, ahol elhelyezhetők. Gyakran önálló tudományos publikációként teszik őket közzé saját DOI (*Digital Object Identifier*) hozzárendelésével. Megosztásukra alkalmas platform lehet az adatokra épülő kutatást bemutató cikkhez kapcsolt ún. *digitális melléklet* is. Az amerikai National Information Standards Organization és National Federation of Advanced Information Services (NISO & NFAIS, 2013) definícióját követve tanulmányunkban digitális mellékletnek tekintünk minden olyan elektronikusan elérhető anyagot, amely különálló fájlként kapcsolódik egy adott publikációhoz.

Mivel nincsenek elérhető adataink arról, hogy a magyarországi tudományos publikálásban jelenleg milyen szerepet kapnak a digitális mellékletek, itt bemutatott elemzésünk céljának azt tűztük ki, hogy megvizsgáljuk jelentőségüket Magyarország legnagyobb tudományos folyóirat-kiadással foglalkozó kiadójának, az Akadémiai Kiadónak a gyakorlatában. Az alábbiakban egy rövid szakirodalmi áttekintést követően bemutatjuk az alkalmazott kutatási módszereket és eredményeinket, majd a megvitatás után felvázolunk néhány fejlődési útvonalat is.

2. Szakirodalmi áttekintés
2.1 Szabályozások

Kutatásunk elméleti megalapozására először a tudományos kutatási adatok digitális kezelésére vonatkozó nemzetközi szabályozást tekintjük át. Két dokumentumot találtunk relevánsnak a témához.

Wilkinson et al. (2016) munkája általánosságban vett tudományos adatok digitális menedzseléséhez nyújt támpontokat. Eszerint a tudományos kiadóknak és adatszolgáltatóknak biztosítaniuk kell, hogy a digitális objektumok

- megtalálhatók (*findable*),
- elérhetők (*accessible*),
- feldolgozhatók (*interoperable*) és
- újrahasznosíthatók (*reusable*)

legyenek emberi és számítógépes feldolgozással egyaránt. (Angol kezdőbetűikből alkotott betűszóval és kedves szójátékkal ezen elvárásokat együtt FAIR adatkezelési követelményeknek nevezik.) Ahogy a szerzők fogalmaznak, ezek az alapelvek létfontosságúak egy adott publikáció színvonalának, illetve hatásának biztosításához, és egyben a tudomány és az innováció előmozdításához (Wilkinson et al., 2016).

A NISO és NFAIS (2013) útmutatója kimondottan a tudományos folyóiratcikkek digitális mellékleteinek kezelésével kapcsolatban fogalmaz meg javaslatokat. Ez nemcsak általános alapelveket közöl, hanem üzleti és technikai kérdésekben is állást foglal, megköny-

nyítve ezzel a tudományos kiadói közösség munkáját világszerte. Jelen kutatásunk szempontjából legfontosabb pontjai az alábbiak:

- a folyóirat (szerkesztőbizottságának) felelőssége eldönteni, mely anyagok kerülhetnek digitális mellékletbe;
- a szerkesztők és kiadók kötelesek felügyelni, hogy a digitális mellékletek tartalmilag hasznosak és relevánsak legyenek; az elfogadásukra vonatkozó kritériumokat ajánlott útmutatóba foglalni;
- a kiadóknak a digitális mellékletekhez ugyanolyan szintű elérhetőséget és megtalálhatóságot kell biztosítaniuk, mint a kapcsolódó cikkekhez;
- a kapcsolódó cikkben ajánlott megfelelően hivatkozni a digitális mellékletre.

Kutatásunk egyik fontos célja volt annak ellenőrzése, hogy a vizsgált digitális mellékletek az Akadémiai Kiadó folyóiratportálján (akjournals.com) milyen mértékben felelnek meg a fenti elvárásoknak.

2.2 Empirikus kutatások

A nemzetközi szakirodalomban kutatásunk előzményének tekinthető Schaffer és Jackson (2004) dolgozata. Ennek célja, hasonlóan a miénkhez, tudományos folyóiratok digitális mellékleteinek vizsgálata 94 nagy hatású, sokat hivatkozott, azaz magas impakttal rendelkező, a tiszta és alkalmazott tudományokat képviselő lap példáján, tartalmuk, előfordulási gyakoriságuk szerint. Mind a tudományterületek szerint, mind a melléklet tartalmának jellegében nagy változatosságot figyeltek meg. A kutatásnak talán az a legnagyobb érdeme, hogy a kapott eredmények alapján Schaffer és Jackson javaslatokat tett a folyóiratok digitális mellékleteinek megjelenési, technikai, és minőségi szabályozására, amellyel utat nyitottak a NISO és NFAIS (2013) erre vonatkozó dokumentumának. Ugyanakkor hangsúlyoznunk kell, hogy a kutatásnak korlátai is vannak, nevezetesen, hogy az eredményei nem általánosíthatók se az összes tudományterületre, se a tiszta és alkalmazott tudományok egészére, mivel a mintában csak magas impakttal rendelkező lapok szerepeltek.

Schaffer és Jackson (2004) általunk is követett objektív méréseivel szemben Price et al. (2018) kérdőíves felméréssel vizsgálta a különböző tartalmú digitális mellékletek hasznosságát, elfogadottságát a szerzők, bírálók és olvasók körében. Eredményük szerint a táblázatokat és ábrákat tartalmazó mellékleteket díjazzák legjobban.

Kutatásunk szempontjából relevánsak azok a kiterjedt vizsgálatok, melyek – folyóiratok más-más körét választva mintául – a kutatási adatok megosztására vonatkozó szabályozást, útmutatásokat elemzik. A témában megjelent friss dolgozat Rousi és Laakso (2020) munkája, amely amellett, hogy 120 magas impakttal rendelkező lap szabályzatait vizsgálja három tudományterületen (idegtudomány, fizika, operációkutatás), széles áttekintést ad a korábbi, hasonló jellegű kutatásokról [lásd a Rousi és Laakso (2020) által hivatkozott dolgozatokat]. A részletek megismétlése nélkül összefoglalóan elmondható, hogy noha egyenetlenség figyelhető meg a tudományterületek között, egyetlen erre irányuló kutatás sem találta kielégítőnek a szabályozást. Érdekes, hogy az útmutatókkal rendelkező folyóiratok többsége nem digitális mellékletben, hanem általános vagy szakspecifikus repozitóriumokban javasolta a kutatási adatok megosztását. Bár az összes ilyen vizsgálatot szükségképpen valamilyen (tematikus vagy hozzáférésen alapuló) szempont szerint kiválasztott mintán végezték, így önmagában egyik kutatás eredménye sem általánosítható messzemenően, az összességükből kirajzolódó kép azt üzeni, hogy a szabályozás a legtöbb folyóiratnál komoly fejlesztésre szorul.

Pop és Salzberg (2015) a digitális mellékletek – részben a szabályozás hiányosságaiból, részben a cikkek terjedelmének korlátozásából adódó – helytelen alkalmazásának veszélyeit, káros következményeit vizsgálta. Fő kockázatnak a bírálat megnehezítését, és a mellékletekben az indexelő adatbázisok elől rejtve maradó hivatkozások – Seeber (2008) által már korábban megfigyelt – elvesztését találta. Borowski (2011) mindezek mellett a szerző energiaráfordításának indokoltsága miatt felveti a mi kutatásunkban is központi kérdést, vajon kapnak-e a digitális mellékletek megfelelő figyelmet, kiváltják-e a kívánt hatást.

3. Kutatási módszerek

Kutatásunk feltáró volta miatt, illetve a teljesebb kép kedvéért úgy döntöttünk, hogy vizsgálatunkban kvantitatív és kvalitatív adatokat is alkalmazunk. Előbbi alatt a digitális mellékletek metaadatait, olvasottsági és hivatkozottsági számadatait értjük, míg kvalitatív adat alatt az egyes lapok szerzői és bírálói útmutatóját.

Adatainkat a következő négy forrásból gyűjtöttük össze: Google Analytics, Crossref, az akjournals.com, és az Akadémiai Kiadó Business Intelligence (BI) adatbázisa. Elemzésünkhöz a Microsoft Excel táblázatkezelő szoftvert és a Publish or Perish[1] akadémiai adatelemző szoftvert használtuk.

Az Akadémiai Kiadó akjournals.com publikációs felületén 48 aktív és 40 archív folyóirat érhető el szinte minden tudományterületről. Ezek közül a Google Analytics segítségével 22 folyóirat 155 cikkéhez találtunk digitális mellékletet. Az érintett folyóiratok mind idegen nyelvűek és változatos tudományterületekhez tartoznak. Bővebb bemutatásuk a Függelékben található.

4. Kutatási eredmények

Ebben a fejezetben az összegyűjtött digitális mellékleteket elemezzük változatos szempontok szerint. Szót ejtünk a technikai hátterükről, jogi és adminisztratív szabályozásukról, eloszlásukról, olvasottságukról, és a kapcsolódó cikkek hivatkozottságáról is.

4.1 A digitális mellékletek technikai háttere

Az Akadémiai Kiadó jelenlegi publikációs felületén, az akjournals.com-on 2008 óta megjelent digitális mellékletek találhatók. A jelenlegi technikai feltételek lehetőséget adnak tetszőlegesen nagy adatfájlok megjelentetésére szinte tetszőleges fájlformátumban (bár technikailag könnyebben

1 Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish

kezelhetők a futtatható programkódot nem tartalmazó fájlok). A digitális mellékletek szerkesztés, korrektúra és áttördelés nélkül, a szerző által benyújtott formában jelennek meg.

A digitális mellékletek a cikk saját internetes oldalán, külön fülön érhetők el. Így saját URL-lel rendelkeznek (amelynek eleje azonban a hozzájuk tartozó cikk főoldalának URL-je), míg a kapcsolódó cikkel azonos DOI tartozik hozzájuk.

4.2 A digitális mellékletek jogi, adminisztratív szabályozása

A digitális mellékletekre vonatkozó felhasználási jog (*copyright*) azonos a kapcsolódó cikkével. Azaz open access cikk esetén mind a cikk, mind a mellékletek felhasználási joga a szerzőnél marad, míg zárt (előfizetés alapú) hozzáférésnél egyaránt átszállnak a kiadóra.

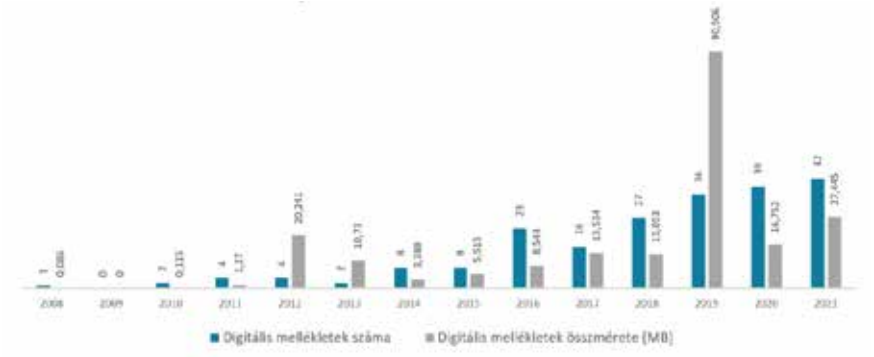
Szintén azonosak a digitális mellékletek és a kapcsolódó cikkek licencfeltételei. Open access cikkek mellékleteire a CC-BY 4.0 és a CC-BY-NC 4.0 licencek közül ugyanaz vonatkozik, mint magára a cikkre.
Azoknál az open access folyóiratoknál, ahol szerzői díj (*Article Processing Charge*) van, az APC összegét nem befolyásolja a digitális mellékletek léte vagy mérete.

4.3 Digitális mellékletek eloszlása

Ebben az alfejezetben különböző szempontok szerint csoportosítjuk a vizsgált mellékleteket.

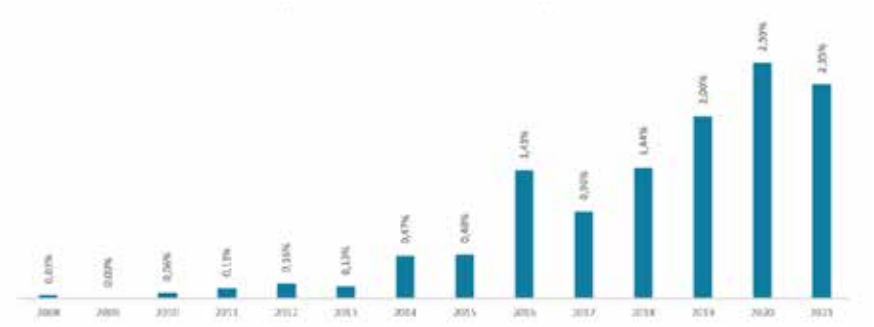
4.3.1 Időbeli eloszlás

A legkorábbi digitális melléklet 2008-ban jelent meg. Azóta az évente megjelenő mellékletek számában és összméretében is lassú növekedés figyelhető meg (1. ábra). (A 2019-ben megfigyelhető, kiugróan nagy összméret egyetlen 62 MB-os szkennelt ábrának köszönhető, így a tendenciák szempontjából figyelmen kívül hagyható.)



1. ábra. Digitális mellékletek száma és összmérete évente.

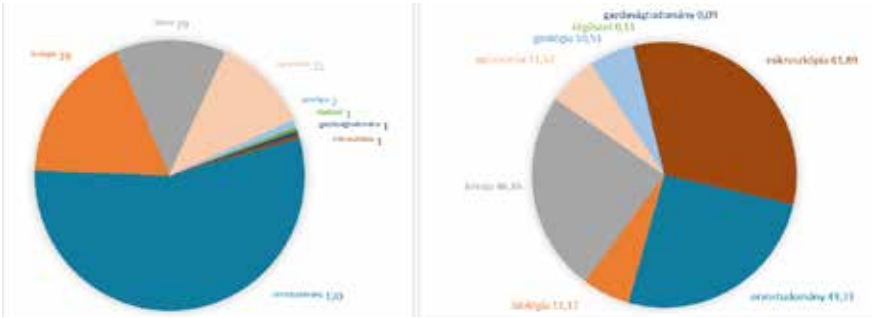
Az időbeli változás vizsgálatakor figyelembe kell vennünk, hogy sem a folyóiratok, sem az évente megjelent cikkek száma nem állandó. Folyóiratok megszűntek vagy átkerültek más publikációs felületre, helyettük újak indultak. Teljesebb képet kapunk, ha azt is megvizsgáljuk, hogy az adott évben megjelent cikkek milyen hányadához tartozik digitális melléklet (2. ábra):



2. ábra. A digitális melléklettel rendelkező cikkek aránya évente.

4.3.2 Tudományterületek szerinti eloszlás

Természetes módon más az adatok, táblázatok szerepe az egyes tudományterületeken, ahogy eltérnek a mellékletek jellemző méretei is (3. ábra):



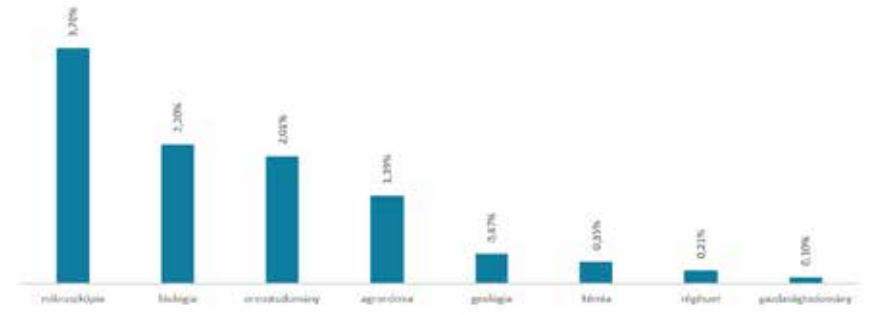
3. ábra. Digitális mellékletek száma és összmérete (MB) tudományterületenként.

Az egyetlen kiugróan nagy mikroszkópia tematikájú, vagy kiugróan kicsi gazdaságtudományi melléklet méretéből nem vonható le általános következtetés, de világosan látszik, hogy például a térképeket tartalmazó geológiai mellékletek tipikusan nagyobbak, mint a jellemzően adattáblákat közlő biológiai mellékletek (1. tábla):

1. tábla. Digitális mellékletek átlagos mérete (MB) tudományterületenként.

tudományterület	**melléklet átlagos mérete (MB)**
mikroszkópia	61,89
geológia	5,26
kémia	1,62
agronómia	0,46
orvostudomány	0,41
biológia	0,29
régészet	0,11
gazdaságtudomány	0,09

Itt is érdekes lehet a mellékletek abszolút száma helyett az arányok vizsgálata (4. ábra):

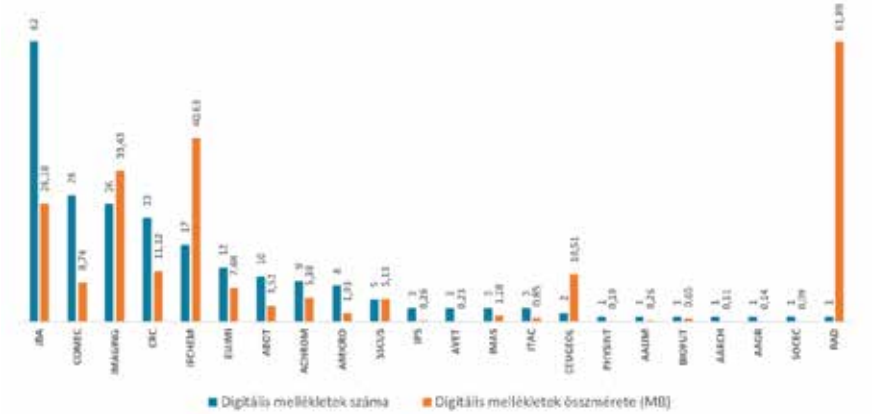


4. ábra. Digitális melléklettel rendelkező cikkek aránya az egyes tudományágakban 2008–2021 között.

Az adatok hitelessége szempontjából fontos hozzátennünk, hogy mikroszkópia témában jelent meg a legkevesebb cikk, mindössze 27, míg a többi tudományterületen legalább 400 cikk alapján következtethetünk.

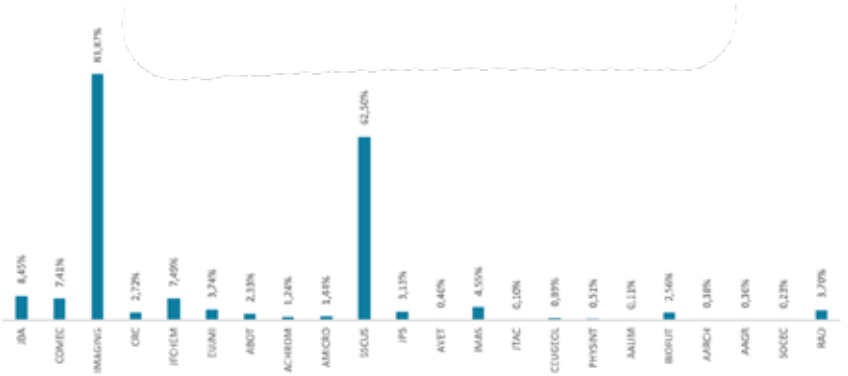
4.3.3 Folyóiratok szerinti eloszlás

A digitális mellékletek gyakoriságát és jellemző méretét egy-egy folyóiratban alapvetően befolyásolja a tudományterületük. Jelentős hatása van azonban számos további, az alábbiakban elemzett tényezőnek is (5. ábra).



5. ábra. Digitális mellékletek megoszlása folyóiratonként [darab és méret (MB) szerint]. Megjegyzés. A rövidítések feloldása a Függelékben található.

Ahogy az időbeli eloszlás vizsgálatánál, itt is torzítja a képet, hogy a tekintett folyóiratok közül nem mind jelent meg az akjournals.com felületen a teljes vizsgálati időszakban. Volt, amelyik később indult, volt, amelyik időközben megszűnt vagy átkerült más publikációs platformra. Itt is teljesebb képet kapunk, ha a digitális melléklettel rendelkező cikkek arányát tekintjük:



6. ábra. A digitális melléklettel rendelkező cikkek aránya az egyes folyóiratokban 2008–2021 között.
A korábban is használt rövidítések feloldása a Függelékben található.

Ahogy a tudományterületek szerinti eloszlásnál is azt láttuk, hogy az orvostudományi folyóiratokban található a legtöbb digitális melléklet (bár arányuk nem itt a legmagasabb), erről az ábráról is az olvasható le, hogy két orvosi lapban a legmagasabb a melléklettel ellátott cikkek aránya. Ugyanakkor az IMAGING az orvosi folyóiratok között is kitüntetett amiatt, hogy témája az orvosi képalkotás, így természetes, hogy szinte minden cikkéhez kapcsolódik fénykép- vagy videómelléklet.

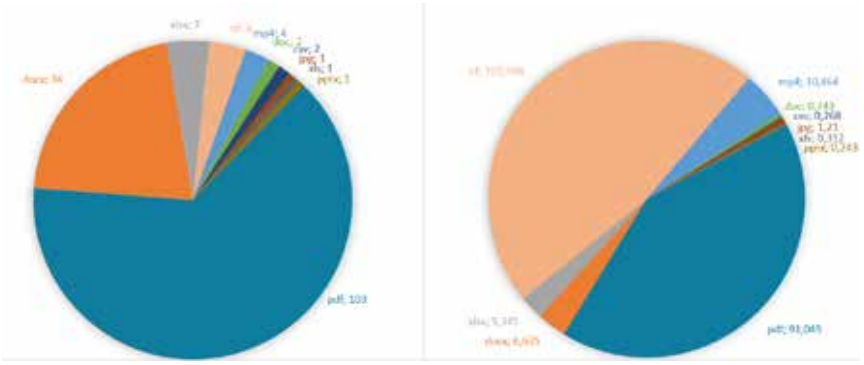
A tekintett folyóiratok közül 10 tisztán open access és nem jelenik meg nyomtatásban, 12 pedig hibrid, azaz egyaránt tartalmaz előfizetéses hozzáférésű és open access cikkeket, továbbá megjelenik nyomtatásban is (lásd a Függeléket). Jelentős különbséget látunk, ha a két csoportban külön-külön vizsgáljuk meg, hogy a cikkek milyen hányadához tarto-

zik digitális melléklet. Ez az arány a tisztán open access folyóiratoknál 2,89%, míg a hibrid lapoknál csupán 1,06%. Az esetleges várakozással szemben tehát nem a nyomtatás költsége terel a mellékletbe egyes részeket.

Ha nem is ekkora, de észrevehető különbség látszik akkor is, ha azokat az aktív, érvényes szerzői útmutatóval bíró folyóiratokat hasonlítjuk össze, ahol van a cikkekre terjedelmi korlát és azokat, ahol nincs. Előbbiek cikkeinek 1,85%-ához tartozik digitális melléklet, az utóbbiak cikkeinek mindössze 0,84%-ához.

4.3.4 Fájltípus szerinti eloszlás

Annak ellenére, hogy a kiadó nem korlátozza a mellékletfájlok lehetséges típusát, mindössze néhány típus fordul elő, és a mellékletek döntő többsége pdf vagy docx formátumú (7. ábra):

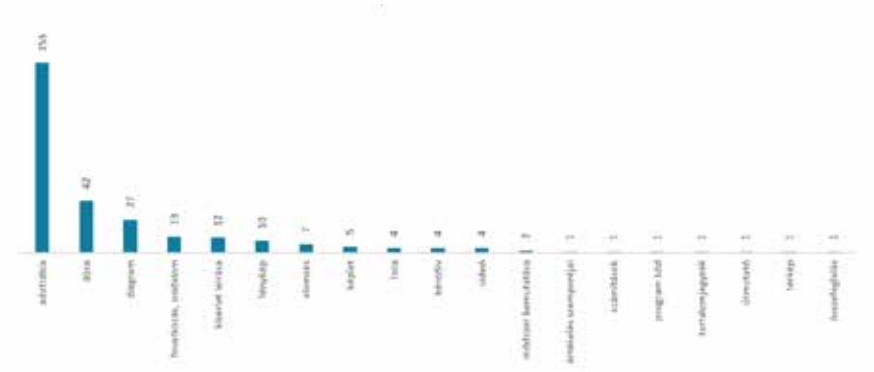


7. ábra. Digitális mellékletek száma és összmérete (MB) fájltípusonként.

Az adott fájltípusú mellékletek összméretét elsősorban az adott fájltípus mérethatékonysága határozza meg.

4.3.5 Tartalom szerinti eloszlás

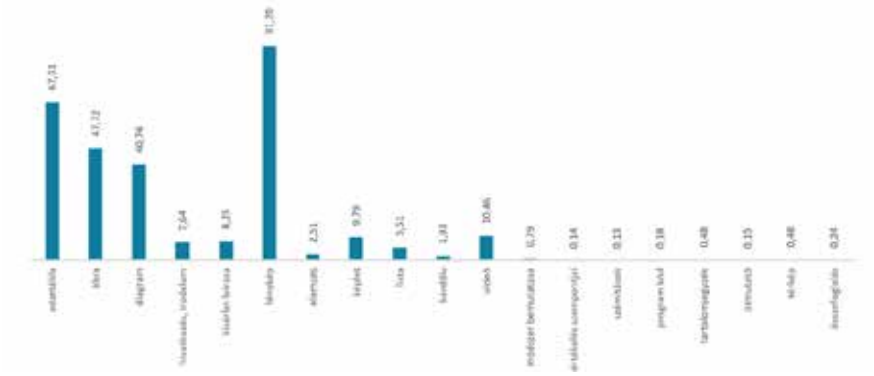
A digitális mellékletek tartalma nagy változatosságot mutat, de a legtöbb mellékletben adattáblákat találunk (8. ábra):



8. ábra. Adott tartalomtípust tartalmazó mellékletek száma.

Megjegyzés. Diagram = kimondottan adatokat ábrázol, ábra = adaton kívül bármi mást ábrázolhat.

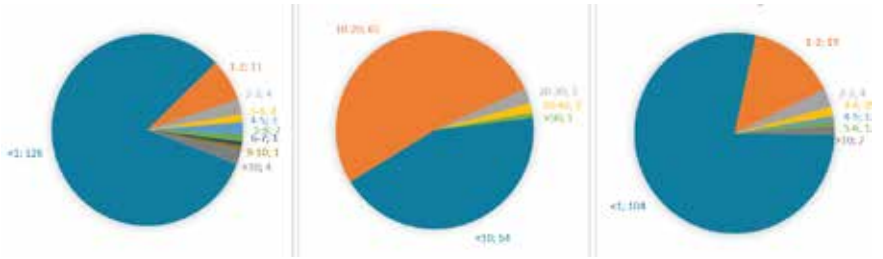
A legnagyobb összmérettel mindazonáltal a képeket tartalmazó mellékletek rendelkeznek (9. ábra):



9. ábra. Adott tartalomtípust tartalmazó mellékletek összmérete (MB).

4.3.6 Méret szerinti eloszlás

A mellékletek méretét mérhetjük megabájtban, oldalban, vagy a kapcsolódó cikk oldalszámának arányában (10. ábra):



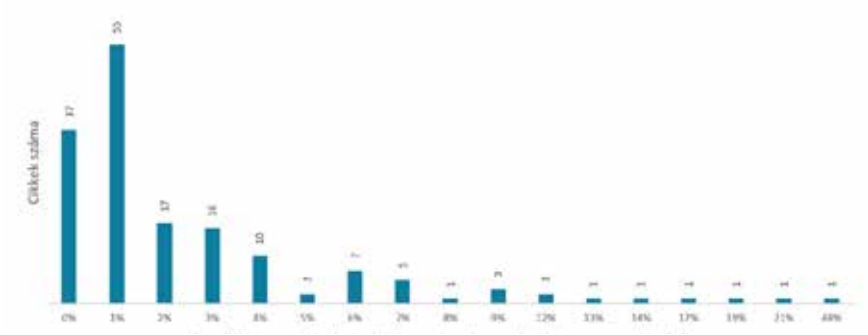
10. ábra. Eloszlás a melléklet MB-ban megadott mérete, oldalszáma, valamint oldalszámának és a hozzátartozó cikk oldalszámának hányadosa szerint.

Leolvasható, hogy a mellékletek jellemzően nem nagyok, a szereplő információ vélhetően nem elsősorban helytakarékosságból kerülhetett mellékletbe, hanem – remélhetően – valós tudományos szempont miatt.

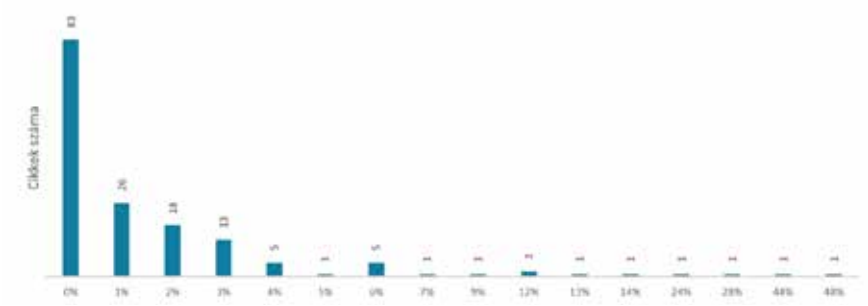
4.4 Digitális mellékletek olvasottsága

A 4.3 fejezet különböző szempontok szerinti leltározása után a jelen fejezetben mélyebbre kezdünk ásni és a mellékletek különböző hatásait vizsgáljuk. Borowski (2011) veti fel a kérdést, arányban van-e a mellékletek visszhangja az elkészítésükre fordított energiával. Erre a kérdésre keresünk választ ebben és a következő fejezetben.

Első lépésként összehasonlítjuk a mellékletek és a hozzájuk kapcsolódó cikkek megtekintéseinek számát (11. ábra), és az online olvasásukkal töltött összes időt (12. ábra). Adataink a Google Analytics-ből származnak és megtekintések alatt az ottani *Page View* számait értjük.



11. ábra. A melléklet megtekintéseinek aránya a hozzá tartozó cikk összes megtekintésében.



12. ábra. A melléklet olvasásával töltött összes idő aránya a hozzá tartozó cikk olvasásával töltött összes időben.

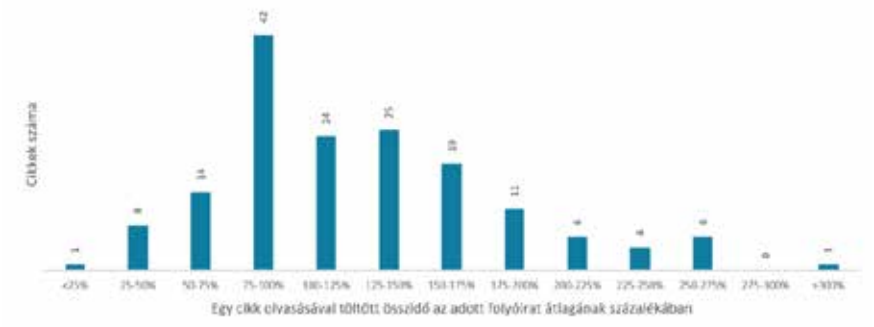
Ezekből az adatokból az az elkedvetlenítő kép rajzolódik ki, hogy az olvasók közül kevesen nyitják meg a mellékletet és a cikkre fordított idejük elenyésző hányadát töltik a melléklettel. Ettől még gondolhatjuk, hogy ha nem is tanulmányozza egy-egy olvasó részletesen a mellékletet, annak léte erősítheti bizalmát a cikk eredményeiben, motiválhatja felhasználásukat.

4.5 Digitális mellékletek hatása a kapcsolódó cikk olvasottságára és hivatkozásaira

Ebben a fejezetben a 4.4 fejezet végén felvetett kérdésnek járunk utána: olvasottabbak, hivatkozottabbak-e a mellékletekkel rendelkező cikkek, mint ugyanabban az évben ugyanabban a folyóiratban megjelent társaik.

A megtekintések számáról (13. ábra) és az összes olvasó által a cikk olvasásával töltött teljes időről (14. ábra) a következő adatokat mértük (adataink ismét a Google Analytics-ből származnak és megtekintések alatt továbbra is az ottani *Page View* számait értjük):

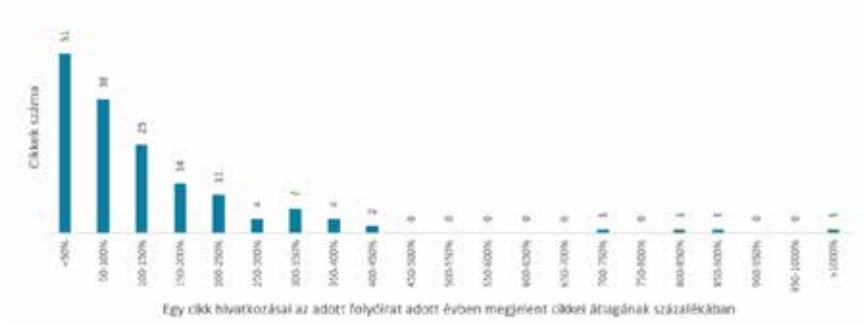
13. ábra. Digitális melléklettel rendelkező cikkek megtekintései az adott folyóirat átlagához képest.



14. ábra. Digitális melléklettel rendelkező cikkek olvasási összideje az adott folyóirat átlagához képest.

Az adatok igazolják azt a feltételezést, hogy a mellékletekkel rendelkező cikkeket az átlagnál többen olvassák. Olvasásukkal a legtöbb esetben az átlaghoz közeli időt töltenek, de inkább többet, mint kevesebbet.

Adataink egyáltalán nem utalnak viszont arra, hogy (a Crossref nyilvántartása szerint) többször hivatkoznak a melléklettel rendelkező cikkeket, inkább ellenkezőleg (15. ábra):



15. ábra. Digitális melléklettel rendelkező cikkek Crossref szerinti hivatkozásai az adott folyóirat adott évben megjelent cikkeinek átlagához képest.

4.6 Digitális mellékletekre vonatkozó útmutatások

Az előző fejezetekben vizsgált 22 folyóirat közül 14 jelenik meg kutatásunk idején is az akjournals.com felületen. Ezek szerzői és bírálói tájékoztatóit ellenőriztük, tartalmaznak-e a digitális mellékletekre vonatkozó útmutatásokat.

Egyetlen szerzői tájékoztatóban, az IMAGING-ében volt említés a mellékletekre vonatkozó elvárásokról. A bírálók számára összeállított útmutatók közül (ha volt a folyóiratnak egyáltalán ilyen dokumentuma), három foglalkozott az ábrák minőségével és szükségességével, de nem kimondottan a mellékletekre vonatkozóan. Nem túlzunk sokat, ha azt mondjuk, a vizsgált folyóiratok nem rendelkeznek a mellékletekre vonatkozó útmutatással sem a szerzők, sem a bírálók számára.

5. Következtetések

Jelen tanulmányunkban azt vizsgáltuk, hogy az Akadémiai Kiadó által kiadott folyóiratok gyakorlatában milyen szerepet kapnak a digitális mellékletek. Motivációnkat az adta, hogy a nyílt tudomány fokozatos térhódítása ellenére a hasonló, magyarországi tudományos publikálással kapcsolatos kutatások eddig hiányoztak. Eredményeink alapján elmondhatjuk, hogy a digitális mellékletek jelen vannak az Akadémiai Kiadó gyakorlatában és bár szerepük fokozatosan nőni látszik, egyelőre nem tekinthető jelentősnek. Tudományterületek szerinti eloszlásukról megállapíthatjuk, hogy nem egyenletes, ami korábbi kutatások (pl. Rousi & Laakso, 2020; Schaffer & Jackson, 2004) alapján természetes jelenségnek tekinthető. A számos érintett tudományterület tükrében ugyanígy természetesnek vélhető a megvizsgált mellékletek tartalmi változatossága. Elemzésünkből kiolvasható a mellékletek valamelyes hatása a kapcsolódó cikkek olvasottságára, nem figyelhető meg azonban határozott befolyásuk a hivatkozottságra.

Kutatásunk talán legfontosabb gyakorlati következtetése a nemzetközi útmutatóknak (NISO & NFAIS, 2013; Wilkinson et al., 2016) való megfelelés erősítése. Egyfelől bátran kijelenthető, hogy a FAIR adatkezelés megvalósul. Másfelől, mivel az egyes lapok digitális mellékleteire vonatkozó útmutatásai erősen hiányosnak bizonyultak, elengedhetetlen ezek fejlesztése a közeljövőben a tudományos szerkesztőbizottságok bevonásával. Ez az eredmény összhangban van a Bevezetésben bemutatott nemzetközi kutatások következtetéseivel.

Hivatkozások

Borowski, C. (2011). Enough is enough. *Journal of Experimental Medicine*, *208*(7), 1337. https://doi.org/10.1084/jem.20111061

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., Hagemann, M., Harnad, S., Johnson, R., Kupryte, R., La Manna, M., Rév, I., Segbert, M., de Souza, S., Suber, P., Velterop, J. (2002, February 14). *Budapest Open Access Initiative*. https://www.budapestopenaccessinitiative.org/read/

NISO (National Information Standards Organization), & NFAIS (National Federation of Advanced Information Services). (2013). *Recommended practices for online supplemental journal article materials* (NISO RP-15-2013). NISO. https://www.niso.org/publications/niso-rp-15-2013-recommended-practices-online-supplemental-journal-article-materials

OECD (Organization for Economic Co-operation and Development). (2015). *Making open science a reality* (OECD Science, Technology, and Industry Policy Papers, No. 25). OECD. http://doi.org/10.1787/5jrs2f963zs1-en

Open Data Handbook. (n.d.). Opendatahandbook. http://opendatahandbook.org/

Pop, M., & Salzberg, S. L. (2015). Use and mis-use of supplementary material in science publications. *BMC Bioinformatics*, *16*(237). https://doi.org/10.1186/s12859-015-0668-z

Price, A., Schroter, S., Clarke, M., & McAneney, H. (2018). Role of supplementary material in biomedical journal articles: Surveys of authors, reviewers and readers. *BMJ Open*, 8(e021753). http://doi.org/10.1136/bmjopen-2018-021753

Rousi, A. M., & Laakso, M. (2020). Journal research data sharing policies: A study of highly-cited journals in neuroscience, physics, and operations research. *Scientometrics*, *124*, 131–152. https://doi.org/10.1007/s11192-020-03467-9

San Francisco Declaration on Research Assessment. (n.d.). Sfdora. https://sfdora.org/read/

Schaffer, T., & Jackson, K. M. (2004). The use of online supplementary material in high-impact scientific journals. *Science & Technology Libraries*, 25(1–2), 73–85. https://doi.org/10.1300/J122v25n01_06

Seeber, F. (2008). Citations in supplementary information are invisible. *Nature*, *451*(887). https://doi.org/10.1038/451887d

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and scholarship. *Scientific Data*, 3(160018). https://doi.org/10.1038/sdata.2016.18

cím	rövidítés	tudomány-terület	kb. éves cikkszám	open access	nyomtatott megjelenés	megjegyzés
Acta Agronomica Hungarica	AAGR	mezőgazdasági	50	nem	igen	2014-ben megszűnt
Acta Alimentaria	AALIM	kémiai	60	hibrid	igen	
Acta Archaeologica Hungarica	AARCH	régészet	20	hibrid	igen	
Acta Botanica Hungarica	ABOT	biológia	25	hibrid	igen	
Acta Chromatographica	ACHROM	kémia	60	igen	nem	
Acta Microbiologica Hungarica	AMICRO	orvosi	40	hibrid	igen	
Acta Veterinaria Hungarica	AVET	orvosi	60	hibrid	nem	
Biologia Futura	BIOFUT	biológia	40	hibrid	igen	2018 előtt más címen jelent meg 2013 óta közös kiadású a Springer Nature-rel
Central European Geology	CEUGEOL	földtudomány	10	igen	nem	
Cereal Research Communications	CRC	mezőgazdasági	60	hibrid	igen	2021 óta közös kiadású a Springer Nature-rel
Community and Ecology	COMEC	biológia	30	hibrid	igen	2021 óta közös kiadású a Springer Nature-rel
European Journal of Microbiology and Immunology	EUJMI	orvosi	25	igen	nem	
IMAGING	IMAGING	orvosi	20	igen	nem	2020-ban indult
Interventional Medicine and Applied Science	IMAS	orvosi	10	igen	nem	2021-ben megszűnt
Journal of Behavioral Addictions	JBA	pszichológia	100	igen	nem	
Journal of Flow Chemistry	JFCHEM	kémia	40	hibrid	igen	2018 óta közös kiadású a Springer Nature-rel
Journal of Psychedelic Studies	JPS	multidisz-ciplináris	20	igen	nem	
Journal of Thermal Analysis and Calorimetry	JTAC	kémia	700	hibrid	igen	2013 óta közös kiadású a Springer Nature-rel
Physiology International	PHYSINT	orvosi	35	hibrid	igen	
Resolution and Discovery	RAD	mikroszkópia	3	igen	nem	
Sleep Spindles and Cortical Up States	SSCUS	orvosi	3	igen	nem	2019-ben megszűnt
Society and Economy	SOCEC	gazdaság	30	igen	nem	

Kutatási adatok a bölcsészettudományban

Maróthy Szilvia

ORCID: 0000-0003-2558-9504
Bölcsészettudományi Kutatóközpont, ELKH

Adat

A kulturálisörökség-digitalizálási törekvések széles körűvé válásával, valamint a személyi számítógépek megjelenésével a bölcsészettudományi kutatásokhoz kapcsolódóan is egyre láthatóbb mennyiségben keletkeznek digitálisan létrehozott, illetve digitalizált kutatási adatok. A kutatáshoz kapcsolódó anyagok, zömmel szövegesek lévén, nem nagy méretűek, a kutatásiadat-kezelés szempontjából inkább a strukturálás, szabványosítás, dokumentálás terén okoznak kihívásokat.

A bölcsészeti kutatások esetében többnyire összegyűjtött, nem generált adatokról beszélünk. A kutatási adatok részint a forrásdokumentumok digitalizálása révén állnak elő, részint az adott kutatási terület által vizsgált jelenségek, tárgyak leírása, osztályozása, elemzése által. Példa kutatási adatokra:

- történeti (irodalom, művészet, zene stb.) adatok, bibliográfiák: Excel táblázattól a webes adatbázisokig
- jelölőnyelvvel (pl. HTML, XML) kódolt tudományos szövegkiadások, forrásközlések és a rájuk épülő webes szolgáltatások
- nyelvi elemzőszoftverrel annotált szövegek (pl. XML, JSON, TSV)
- szövegszerkesztőben rögzített szövegátiratok, adatgyűjtések (pl. DOCX, ODT)
- digitális kották, zenei notációk (pl. SMF/XMF, XML)
- audiovizuális tartalmak
- adat- és korpuszelemzések eredményei (pl. statisztikák, diagramok, interaktív tartalmak)
- programkódok
- megjelenítés, interface

A kutatási adatok strukturáltságának mértéke igen eltérő lehet, melynek csak részben oka az információ vagy a technikai jártasság hiánya. A bölcsészettudomány szöveg- és hagyományközpontú, több értelmezést megengedő, interpretáción alapuló, így az osztályozási rendszerek következetes alkalmazása nem mindig lehetséges, emellett gyakoriak a bizonytalan adatok és a hiányok is. A kutatási adatok összegyűjtése pedig önmagában jelentős befektetéssel jár, az adatkezelés, dokumentálás inkább adminisztratív jellegű terheire a kutató egymagában nem áll készen.

Általános problémát jelentenek a projektalapú működésből adódóan félbemaradt, eltűnt vagy elfeledett, illetve az elavult adatbázisok is. Ezek archiválása és hozzáférhetővé tétele a strukturáltság és a dokumentáció említett hiányai miatt nehézségekbe ütközik. A Bölcsészettudományi Kutatóközpontban például közel 150–200 kutatásokhoz kapcsolódó honlap van, s emellett számos olyan (online vagy offline, a kutatók saját adathordozóin tárolt) dokumentummal is számolnunk kell, melyek biztonsági mentése, archiválása nem feltétlenül megoldott. Hosszú távon ezekkel a kihívásokkal akkor tud egy intézmény megküzdeni, ha már a kutatás tervezési szakaszától fogva, annak teljes futamideje alatt szakmai és infrastrukturális támogatást nyújt a kutató(csoport)nak. Noha vitán felül áll, mennyi előnnyel jár a kutatási adatok digitális létrehozása és elemzése, a publikálásuk terén a bölcsészettudomány képviselői lemaradásban vannak. Ez az átmeneti időszak – melynek során már nem nyomtatják ki, de még nem publikálják a weben a kutatási eredmények ezen típusát – pedig jelentős adatvesztéssel jár a jövő kutatónemzedékeire nézve, s efölött permanens eltékozlása is a kutatásra fordított közpénznek. (Lásd az esettanulmányokat: Maróthy 2020.)

Azt, hogy az adatkezelés a kutatási folyamat szerves részévé váljék, nagyban támogatják azon EU-s, és most már magyarországi közfinanszírozású pályázatok (elsősorban az NKFIH által kiírtak) is, melyek 1) elvárják adatkezelési terv (data management plan, DMP) benyújtását pályázáskor, 2) minél szélesebb körű nyílt publikálást irányoznak elő.

A DMP nem pusztán a kutatót segíti projektje megvalósításában, hanem befogadó intézményét is az adatok archiválásához és közzétételéhez megfelelő környezet biztosításában.

Az alábbiakban a kutatásiadat-kezelés kapcsán egyre általánosabban alkalmazott FAIR alapelvek bölcsészettudományi vonatkozásaival, a repozitóiumi archiválás lehetőségeivel, valamint a kutatási adatok kezelésének, publikálásának kutatásértékelésben elfoglalt – jelenleg nem túl előkelő – helyzetével foglalkozom.

FAIR

A kutatásiadat-kezelési irányelveket összefogó FAIR (Findable, Accessible, Interoperable, Reuseable) alapelvek (Wilkinson és mtsai. 2016) a publikálásuk óta eltelt hat évben a legtöbb tudományterületen ismertté váltak, mára bölcsészettudományi alkalmazásukról is rendelkezésünkre állnak tapasztalatok. A FAIR bölcsészeti implementálását szorgalmazzák, segítik többek között az ALLEA (All European Academies) ajánlása (Harrower és mtsai. 2020), a DARIAH (Digital Research Infrastructure for the Art and Humanities) oktatási anyagai és a CO-OPERAS (Open access in the European research area through scholarly communication) jelentései,[1] vagy a Library Carpentry rövid, tudományterületekre fókuszáló útmutatói (Top 10 FAIR Data & Software Things).[2]

Vannak azonban tudományterületből fakadó nehézségek is az adatok közzétételében. A megtalálhatóság és hivatkozhatóság feltétele az állandó azonosítók (pl. DOI) megléte, ezzel azonban még kevés kutatásiadat-gyűjtemény/adatbázis büszkélkedhet ezen a területen. Részint azért, mert nem is teszik közzé a kutatási adatokat, az nem része a

1 Például Elena Giglia, Arnaud Gingold, Iraklis Katsaloulis, Lottie Provost and Francesca Di Donato (2021). FAIR Data in Social Sciences and Humanities. DARIAH-Campus, https://campus.dariah.eu/id/3fOvyNYHb2Hhq4sQCbtsT; CO-OPERAS reports on FAIRification efforts in the SSH, https://www.go-fair.org/2020/08/28/co-operas-publishes-a-variety-of-workshop-reports-on-fairification-efforts-in-the-ssh/.

2 https://librarycarpentry.org/Top-10-FAIR/

publikálási gyakorlatnak, részint amiatt, hogy az azonosítóval való ellátás infrastruktúrája, finanszírozása (pl. DataCite előfizetés formájában) nem általánosan megoldott. A kutatási adatok dokumentálása, archiválása data stewardok híján a kutatókra hárul, akikre ez egyfajta láthatatlan munkaként ró plusz terheket.

Fontos szempont a nyílt közzététellel kapcsolatban, hogy a kutatások sok esetben közgyűjtemények dokumentumaihoz kapcsolódnak, melyek részint a lassan változó közgyűjteményi közzétételi gyakorlatok, részint a szerzői jogok miatt nem publikálhatók, s ezáltal nehezítik a FAIR-esítést a kutatók számára.

> „Europeana survey reveals that only one third (thirty-four percent) of digitised cultural heritage resources are currently available online, with barely three percent of these works suitable for real creative reuse; meaning, only this three percent has the chance to fulfil the discipline-specific measures of being FAIR." (Tóth-Czifra 2020)

A szerzői jog tekintetében nagy előrelépés az EU 2019/790-as irányelve a digitális egységes piacon alkalmazandó szerzői és szomszédos jogokról, mely kivételt képez a szerzői jog érvényesítésében a következő felhasználási esetekben: szöveg- és adatbányászat; a művek oktatási szemléltetés céljából történő digitális felhasználása; a kulturális örökség megőrzése.[3] A szabályozás talán nem kapott még kellő figyelmet a kulturális örökséggel és a kutatással foglalkozó intézmények részéről, noha éppen a kutatási adatok FAIR-esítésében, valamint a digitális kulturális örökség közzétételében kulcsszerepe lehet az új szabályozás adta lehetőségek kiaknázásának.[4]

3 Lásd részletesen: Szerzői és szomszédos jogok a digitális egységes piacon, https://eur-lex.europa.eu/legal-content/HU/LSU/?uri=CELEX:32019L0790.

4 A közgyűjtemények viszonyáról és az új szerzői jogi környezetről jó áttekintést ad a Networkshop 2021 műhelybeszélgetése, különösen Lábody Péter előadása: „Könyvtári (közgyűjteményi) digitális tartalmak újrahasznosításának lehetőségei, feltételei a hálózatban", https://kifu.videotorium.hu/hu/recordings/42177/konyvtari-kozgy-jtemenyi-digitalis-tartalmak-ujrahasznositasanak-lehetosegei-feltetelei-a-halozatban.

A kutatási adatok feldolgozására számos általános és területspecifikus szabvány rendelkezésre áll, mely az átjárhatóságot, újrafelhasználást segíti.[5] A szövegkódolásra a Text Encoding Initiative XML alapú, egyre kiterjedtebbé váló kódolási rendszere széles körűen használt. A dokumentumokat leíró metaadatok terén a könyvtári szabványok állnak rendelkezésre, igaz, ennek ellenére aránylag ritkán alkalmazzák azokat. A kutatások során a legtöbb esetben egyedi metaadatkészletekre van szükség (a már említett forrásközeliség és az értelmezési hagyományok jelenléte okán), ezek „kinyitása", FAIR-esítése is nehézségeket okoz – ezért vagy több szabványt együttesen alkalmaznak, vagy gyakrabban egyet sem. A különféle biográfiai forrásokat feldolgozó és azokat prozopográfiai adatbázisba rendező Norssi High School Alumni projekt[6] például úgy hidalta át a heterogén adatforrások és az érvényben lévő metaadatszabványok közötti szakadékot, hogy saját leíró rendszerét zömmel más szabványok elemeiből állította össze, együttesen felhasználva a schema.org, a SKOS és a CIDOC-CRM sémáit (Leskinen, Hyvönen, és Tuominen 2018).

A bölcsészettudományi kutatási adatok archiválására és közzétételére is számos nemzetközi gyűjtőkörű repozitórium áll rendelkezésre, ezek többsége nem szakterület-specifikus. Szélesebb körben ismert, bölcsészeti kutatásokban is használt általános repozitóriumok például a Zenodo, a Figshare, illetve az elsősorban programkódok kezelésére használt GitHub.[7]

A Re3data (Registry of Research Data Repositories) adatbázisa szerint 337 bölcsészet- és társadalomtudományi területet is felölelő adatbázis/-repozitórium van, köztük néhány olyan, amely kifejezetten bölcsészeti kutatásokat támogat – ilyen például az ARCHE (A Resource

5 A bölcsészettudományban alkalmazott szabványokhoz lásd a Research Data Alliance által kezdeményezett, újabban kibővített Metadata Standards Catalog tudományterület-specifikus gyűjtését: https://rdamsc.bath.ac.uk/subject/Arts%20and%20humanities.

6 https://www.ldf.fi/dataset/norssit

7 https://zenodo.org/, https://figshare.com/, https://b2share.eudat.eu/, https://github.com/

Centre for Humanities Related Research in Austria) vagy a DARIAH-DE Repository.[8] Az ARCHE az osztrák kutatási infrastruktúra része, archiválási irányelveiben az OAIS (Open Archival Information System) ajánlásait követi, annak mentén ad részletes tájékoztatást az archiválási folyamatról, emellett számos archiváláshoz kapcsolódó szabványt támogat (pl. DCMI, OWL, OAI PMH, FAIR). A DARIAH-DE Repository ugyan a DARIAH németországi szervezetének fejlesztése, azonban nyitott minden kutató és kutatócsoport számára. A szolgáltatás számos más DARIAH-DE által fejlesztett és fenntartott digitális bölcsészeti eszközzel kapcsolatban áll (pl. TextGrid, Geo-Browser), ezen infrastruktúra szerves részét képezi.

A DARIAH DDRS (Data Deposit Recommendation Service) szolgáltatása a Re3datához hasonlóan segítséget nyújt bölcsészettudományi kutatóknak a megfelelő repozitórium kiválasztásában, igaz a rendszer csupán két szempontot vesz figyelembe: a kutató országát és kutatási területét. Magyarországi kutatóként kutatási területre nem szűkítve a következő négy javaslatot adja a kereső: B2SHARE, Zenodo, Figshare és a Debreceni Egyetem Adattára – azaz más hazai repozitóriumot a kereső jelenleg nem ismer.[9]

Ha a magyarországi körképet nézzük, számos intézményi repozitórium áll a kutatók rendelkezésére, melyek elsősorban dokumentumok archiválására jöttek létre (pl. MTA Könyvtár: REAL, Eötvös Loránd Tudományegyetem: EDIT, Debreceni Egyetem: DEA, Szegedi Tudományegyetem: Contenta). Amennyiben a kutatási adat publikáció vagy diplomamunka/disszertáció mellékletét képezi, általában lehetőség van a közleménnyel együtt azok archiválására is, de ez nem része a gyakorlatnak, a dokumentum-repozitórium funkciói sem felelnek meg ennek a célnak.

8 https://www.re3data.org/, https://arche.acdh.oeaw.ac.at/, https://de.dariah.eu/en/web/guest/repository

9 https://ddrs-dev.dariah.eu/ddrs/

Üzemelő adatrepozitóriuma jelenleg a Társadalomtudományi Kutatóközpontnak van (Micsik és Gárdos 2014). Adatrepozitórium létrehozásával (a Dataverse szoftver implementálásával) jelenleg a SZTAKI, a Társadalomtudományi Kutatóközpont, a Wigner Fizikai Kutatóközpont az ELKH Adatrepozitórium Platform (ARP) projekt keretében, valamint a Debreceni Egyetem foglalkozik. Mindkét fejlesztés alatt álló repozitórium fogad (alapvetően az adott intézményhez kapcsolódó) kutatási adatokat. A rendszerek fejlesztés alatt állnak, DOI szolgáltatást egyelőre csak a debreceni Adattár biztosít. Utóbbinak további előnye, hogy elérhető egy-egy rövid magyar nyelvű tájékoztató az adatmegosztás menetéről (RDA ajánlások nyomán), valamint a FAIR közzététel elveiről. A projektek kezdeti állapotát mutatja, hogy egyik platform se tartalmaz leírást önmagáról, céljairól, az általa biztosított szolgáltatásokról stb. – csak a fenntartó intézmény megnevezése és egy általános kapcsolattartó email található az oldalon. Összehasonlítás végett, ahhoz, hogy milyen információkat érdemes repozitóriumi honlapon feltüntetni, két példa: Digital Repository of Ireland, Repository of Open Data/RepOD.[10]

A (számítógépes) bölcsészeti projektek gyakran saját infrastruktúrát építenek, melynek része az archiválás is. Ilyen rendszert tervez az előbb a PIM, majd az OSZK intézményéhez tartozó Digitális Bölcsészet Központ,[11] és ilyen lett volna az időközben eltűnt/félbemaradt ELTE Digitális Bölcsészet Központ nagyszabásúnak indult repozitóriuma is, melyet Islandora CLAW és Drupal összekapcsolásával fejlesztettek. Noha korábban a Magyar Filozófiai Tudástár (vagy MAFITUD, nem összetévesztendő a Magyar Fiatal Tudósok Társaságával) volt a

10 https://www.dri.ie/, https://repod.icm.edu.pl/

11 Erről legújabban: Mihály Eszter, „Mi az a dHUpla?" Networkshop 2022 konferenciakötet, Budapest: Hungarnet Egyesület, megjelenés alatt.

pilot projektje ennek a fejlesztésnek,[12] mára a repozitóriumban nem található meg ez az anyag, ahogy a többi, ezen a platformon létrehozott gyűjtemény is üres.[13] A projektalapon működő kutatás/digitalizálás eredményeképp létrejövő kutatási adatoknak (2-3 évnél) hosszabb távú archiválása és szolgáltatása a tapasztalatok szerint nem valósítható meg projekten belül, ahhoz szélesebb körű összefogásra, magasabb szintű intézményi háttérre, hosszabb távú stratégiára van szükség. Jelenleg egy olyan magyar bölcsészeti műhely sincs, mely kutatásiadat-szerű kimeneteit a FAIR elvek hozzáférési és újrafelhasználhatósági kritériumainak megfelelően közreadta volna, az elmúlt évek, évtizedek infrastrukturális fejlesztései ellenére.[14]

Az említett Concorda a HRDA kutatásiadat-kezelési pályázatának köszönhetően már rendelkezik egy jelentősebb bölcsészettudományi adatgyűjteménnyel, ez a Bölcsészettudományi Intézet Régészeti Intézetének rajzgyűjteménye, mely több mint 1100 rekordot számlál. Ezen gyűjtemény repozitóriumi archiválásáról számol be a jelen kötetben Horváth Friderika és Kiss Tünde.

Kutatásértékelés és tudománymetria

Van egy eddig nem említett tényező is, amely a kutatásiadat-kezelési gyakorlatok sikerességéhez jelentősen hozzá tudna járulni a bölcsészettudományok terén is, mégpedig a kutatási adatok létrehozásának,

12 Palkó Gábor és Smrcz Ádám, „A Magyar Filozófiai Tudástár bemutatása," Networkshop 2018 konferencia, http://kifu.videotorium.hu/hu/recordings/21153. Az ELTE DH projektjei részint egy másik infrastruktúrába kerültek át, melynek keretében a WikiBase szoftvert használják (nem közösségi) adatbázis-építésre, részint újabb repozitórium fejlesztésébe fogtak InvenioRDM szoftverrel. Vö. újabban Kiss Tamás, Palkó Gábor, „Adatrepozitórium digitális bölcsészeti funkciókkal," 2022. április 6., http://mtabtk.videotorium.hu/hu/recordings/45893.

13 http://repository.elte-dh.hu/s/magyar-filozofiai-tudastar-hu/page/about, gyűjtemények: http://repository.elte-dh.hu/s/eltedh-hu/page/home

14 Egyedi, szigetszerű példák kutatási adatok közzétételére természetesen a bölcsészettudományban is vannak, itt a rendszerszintű hiányon, a kutatóintézmények, műhelyek reprezentációján van a hangsúly.

kezelésének és publikálásának elismerése kutatási tevékenységként. Jelenleg a kutatási adatok gyűjtése, dokumentálása, közzététele és archiválása a kutató „láthatatlan munkája". Nem képezi szerves részét a kutatói munkának, a kutatásértékelési rendszerek és a tudománymetria sem igen foglalkozik vele, a kutatók magukra maradnak ezzel a feladattal a tájékozódástól a megvalósításig.

Szemmel látható a bölcsészettudományok lemaradása a kutatási adatok publikálása terén, s ez elsősorban nem az egyéni kutatók felelőssége. Az irányító, döntéshozó szereplők feladata, hogy ezt elősegítsék és előirányozzák, megteremtsék a kutatásiadat-kezeléshez szükséges feltételeket, valamint jutalmazzák a kutatási adatok szakszerű és széles körű közzétételét a kutatásértékelési rendszerekben.

Hogy általános, a bölcsészettudományokat érintő problémakörről van szó, jól mutatják azok az újabban közreadott, illetve készülő jelentések, ajánlások, melyek a kutatási terület egyediségeire hívják fel a figyelmet kifejezetten a kutatásértékelés, illetve a tudománymetria vonatkozásában. Ilyen az OPERAS (Open Scholarly Communication in the European Research Area for SSH) átfogó jelentése, mely többek között a sokszínűség és soknyelvűség, az új (webes) publikációs és kollaborációs formák, műfajok, valamint a minőségellenőrzés és a kutatásértékelés témakörével foglalkozik. Néhány példa az ajánlásból arra, hogy milyen problémákat azonosítottak:

- Az írás [értsd tudományos közlés] innovatív formái jelenleg nem kellően elismertek az akadémiai közegben.
- Akadályozzák az innovációt a minőségértékelési rendszerek, a presztízs és a kompetenciák hatásai, valamint az új [közlési] formák hivatkozási gyakorlatának, szabványainak hiánya.
- A hatalmi struktúrák blokkolják az innovációt.
- A kompetenciahiány visszafogja az új eszközök alkalmazását.

(saját fordítás)

Az ALLEA E-Humanities munkacsoportjának készülő jelentése (munkacím: Recommendations on Recognising Digital Scholarly Outputs in the Humanities) is a kutatásértékelést helyezi középpontba, a bölcsészettudományi kutatás jellemző kimeneteit, publikációs műfajait, s azok lehetséges minőség-ellenőrzési és értékelési szempontjait vonultatja fel. Esettanulmányai többek között a digitális szövegkiadással, a történeti adatbázisokkal, az adatvizualizációkkal, valamint a szoftverekkel, programkódokkal is foglalkoznak.

Összegzés

A bölcsészettudomány, noha nem jár élen a kutatási adatok közzététele terén, viszont egyre nagyobb mennyiségben termeli a digitális adatokat. A lemaradás okai többek között a kutatási terület történeti előzményei, az elmaradottabb infrastruktúra, az egyre csökkenő finanszírozás, valamint a digitalizálásban és kutatásban érdekelt szereplők mérsékelt összetartása. A jövőben nagy lehetőségeket rejthet a kutatási adatok archiválása, közzététele és újrafelhasználása terén az Európai Unió szerzői jogi környezetének közgyűjteményekre és kutatókra nézve előnyös változása, mely lehetővé teszi a digitálisan keletkezett és digitalizált kulturális örökség tartalmainak hozzáférését oktatási, kutatási és megőrzési célokra. A kutatási adatok előállítása, dokumentálása, közzététele erőforrás-igényes, az e téren való előrelépés csak akkor lehetséges, ha a kutatókat egyszerre támogatják a megfelelő infrastruktúra és szakemberek (data stewardok) biztosításával, valamint a kutatásértékelési rendszerek revideálásával.

Irodalomjegyzék

Harrower, Natalie, Maciej Maryl, Timea Biro, és Beat Immenhauser. 2020. *Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities*. Berlin: ALLEA – All European Academies. https://doi.org/10.7486/DRI.tq582c863.

Leskinen, Petri, Eero Hyvönen, és Jouni Tuominen. 2018. „Analyzing and Visualizing Prosopographical Linked Data Based on Biographies". https://aaltodoc.aalto.fi:443/handle/123456789/35320.

Maróthy Szilvia. 2020. „A nyílt és a zárt tudományról". In *Kulturális iparágak, kánonok és filterbuborékok*, szerkesztette Bárány Tibor, Hermann Veronika, és Hamp Gábor, 25–38. Budapest: Typotex. https://edit.elte.hu/xmlui/handle/10831/46729.

Micsik András, és Gárdos Judit. 2014. „Tudományos repozitóriumok az MTA-ban: a KDK és a SZTAKI tanulságai". In *Informatika a felsőoktatásban*. Debreceni Egyetem Informatikai Kar. http://openarchive.tk.mta.hu/340/.

Tóth-Czifra, Erzsébet. 2020. „10. The Risk of Losing the Thick Description: Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem". In *Digital Technology and the Practices of Humanities Research*, szerkesztette Jennifer Edmond, 235–66. Open Book Publishers. https://doi.org/10.11647/obp.0192.10.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, és mtsai. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Kutatásiadat-kezelés a csillagászat területén

Holl András

ORCID: 0000-0002-6873-3425
MTA Könyvtár és Információs Központ

A cikk bemutatja egy tudományterület kutatásiadat-kezelési gyakorlatát. A csillagászat élen jár a kutatási adatok megosztásában – érdemes tanulmányozni az évtizedek alatt elért eredményeket.

1.) Fejlett adatkezelés – okok, történet

A csillagászat elöl jár a kutatási adatok kezelésének területén.[1] Ennek a ténynek az okát részben a tudomány erősen haszon-talan (non-profit) jellegében kell keresnünk. Részint abban, hogy az égi jelenségeket nemhogy kisajátítani nem lehet, de gyakorta szükség van egymástól távoli földrajzi helyű megfigyelők együttműködésére, vagy különböző megfigyelési eszközök adatainak egyesítésére.

Az obszervatóriumok közötti adatcsere igénye hozta létre a Flexible Image Transport System (FITS) nevű adatszabványt 1981-ben (Greisen, 2003). A szabványos adatformátum léte, de még inkább azon tulajdonsága, hogy a rugalmasan bővíthető, de mégis szabványos metaadatok magában az adatfájlban kaptak helyet, kínálta a lehetőséget arra, hogy az (amúgy nyílt forráskódú, szabad) feldolgozó szoftverek is támogassák. Mindennek az lett az eredménye, hogy a teljes adatfeldolgozási folyamatba beépült a részletes dokumentáció (Holl, 2020).

Az adatok megosztása, újrafelhasználása egyre nagyobb jelentőségre tett szert a megaprojektek korában. Fél évszázaddal ezelőtt a kutatási adatokból kevés volt, az adatok összegyűjtése (megfigyelés) sok időt és energiát emésztett fel. A megosztás praktikus lehetősége az adatpublikálás volt.

1 „Astronomy has been a pioneer of Open data sharing, and remains at the forefront." Turning FAIR into reality. European Union, 2018. doi: 10.2777/1524

Publikálás előtt a megfigyelési adatokat gondosan őrizték, és amikor publikálták (intézeti kiadványsorozatban, katalógusban), akkor az hivatkozhatóvá is vált. Az űrobszervatóriumok (csillagászati megfigyelést végző szatelliták, mint a *Hubble Űrteleszkóp*, *Kepler*, *Gaia*) és a földfelszíni nagyprojektek (pl. *Sloane Digital Sky Survey*) térnyerése adatbőséget teremtett a csillagászatban. Immár több az adat, mint amennyit a pályán lévő csillagászok belátható idő alatt fel tudnak dolgozni. Az adatbőség teremtette meg a szükségességét annak, hogy az adatok dokumentáltan, mind könnyebben hozzáférhetően – természetesen digitális formában – adatbázisokba kerüljenek. A Hubble Űrteleszkóp adatain alapuló publikációk többsége immár az archívum felhasználásával készül (Novacescu et al., 2018, 1. ábra).

2.) Adatmegosztás: a Virtuális Obszervatórium

Virtuális obszervatóriumként működhetnek a megfelelően dokumentált és hozzáférhetővé tett kutatási adatok: esetenként nem kell új (drága) megfigyeléseket végezni, mivel a meglévő, kutatható adatok is képezhetik a kutatás alapját. Az Egyesült Államok *National Virtual Observatory* kezdeményezésének alapdokumentumát 2002-ben adták ki.[2] Ugyanebben az évben hozták létre a Nemzetközi Virtuális Obszervatórium Szövetséget (*International Virtual Observatory Alliance*).

Nem kellett sokat várni az NVO első eredményeire. Egy 2003-as demonstráció során, amikor két nagy katalógus (SDSS és 2MASS) adatait párosították egy szoftverprototípus segítségével, felfedeztek egy új barna törpecsillagot. Az új technikával a két hatalmas adathalmazban való keresés (és találat) mindössze két percet vett igénybe – ez korábban több hétnyi vagy hónapnyi emberi munkát igényelt volna.[3]

A VO húsz éve alatt elsősorban szabványokat teremtett, de létrejöttek szoftverek és adatközpontok által nyújtott VO szolgáltatások is. Ezek közül mutatunk be néhányat.

2 Towards the National Virtual Observatory. http://www.virtualobservatory.org/documents/sdt-final.pdf

3 Johns Hopkins University. „Virtual Observatory Prototype Produces Surprise Discovery; Early Demo Project Identifies New Brown Dwarf." ScienceDaily. 12 March 2003. http://www.sciencedaily.com/releases/2003/03/030312071232.htm

A strasbourgi CDS-ben készült az *Aladin* égbolt-térkép szoftver, melynek van webes szolgáltatásként működő, személyi számítógépre telepíthető és böngészőben futó javascript változata is. A *Google Maps*-hez hasonlítható, csak nem a Földet, hanem az eget ábrázolja. Számos nagy, teljes égboltot lefedő felmérés képeit vagy katalógusait meg lehet vele jeleníteni, de ki-ki saját képeit is megnézheti a „beépített" térképekre vetítve. A hazai *Information Bulletin on Variable Stars* (IBVS) folyóirat is használta megjelenítő eszközként: mind egy égboltterület több színben fotometrált csillagainak adatait tartalmazó táblázatot, mind egy kis égterület fényképét meg lehetett jeleníteni a segítségével. Az olvasó csak kattintott egyet, és a folyóiratcikkhez tartozó adatok megjelentek a felbukkanó térképi alkalmazásban (Holl, 2022).

Ahhoz, hogy az Aladin segítségével „privát" adatokat fel lehessen tenni a térképre, szabványokra van szükség. Az egyik ilyen a táblázatos adatokban található mennyiségek szabványos leírására szolgáló Universal Column Descriptor, a másik maga a szabványos táblázat formátum: a VOTable.

A gyors lefutású égi események felfedezése esetén a közösség riasztására szolgáló protokoll a VOEvent. Például a gravitációshullám-obszervatóriumok ilyen híradásokat adnak ki azért, hogy az általuk megfigyelt események elektromágneses sugárzását a készenlétben álló teleszkópok percek múlva megkísérelhessék detektálni (Williams és Seaman, 2006).

A csillagászatban korán létrejött tudományterületi szabványok egyike volt a közlemények (de akár publikált adatok) azonosítására szolgáló BIBCODE. Viszont ennek az azonosítónak a használata esetenként késleltette a DOI bevezetését.

Az International Virtual Observatory Alliance weboldala jelenleg 58 szabványjellegű dokumentumot tartalmaz.[4] A szervezet tevékenységét Berriman et al. (2020) foglalja össze.

4 https://www.ivoa.net/documents/

3.) IBVS

A már említett IBVS adatközlő folyóiratként is működött. Közölt adat-cikkeket, melyekben az adatok a cikkben foglalt táblázatokban a nyomtatott/PDF cikk részét képezték, más esetekben pedig a cikk digitális mellékleteként voltak az adatok a webről letölthetőek (1995 után). Ez utóbbi adatfájlok egyszerű szerkezetűek voltak, gépi feldolgozhatóságra alkalmas formában. A fentebb említett, CDS Aladinban való megjelenítésre szánt adatfájloknak elérhető volt egy VOTable változata is.

Az IBVS már nem aktív folyóirat – az utolsó cikkek 2019-ben jelentek meg. A cikkek linkekkel gazdagított HTML változatai sem érhetőek már el, csak a PDF-ek. A többnyire egyszerű, szöveges állományok formájában elhelyezett adatfájlok viszont elérhetőek továbbra is a folyóirat honlapján, és külső archívum(ok)ban is.

4.) Az élenjáró és a kevéssé fejlett területek

A kutatási adatok FAIR követelmények szerinti kezelése munkaigényes. A megfelelő szintű dokumentálás legkönnyebben a nagy projektek (például űrobszervatóriumok vagy nagy költségvetésű földfelszíni projektek) esetében oldható meg. Ezekben az esetekben az adatfeldolgozás automatikus folyamatban történik, így a megfelelő dokumentálás, gazdag metaadatkészlet biztosítható.

Megoldható a FAIR adatkezelés a nagy adatközpontokban is – ezeknél rendelkezésre áll a költségvetés, és az adatokkal képzett adatgazdászok (data steward) foglalkoznak. A csillagászatban ilyen adatközpont például a *Centre Données de Strasbourg*, az *ESAC Science Data Centre*, a *Mikulski Archive for Space Telescopes* és az *ESO Science Archive Facility*. Külön említendő a NASA támogatásával működő *Astrophysics Data System*, amely ugyan bibliográfiai adatbázis, de nyilvántartja a cikkekhez kapcsolódó adatjellegű forrásokat is.

Ha a kis költségvetésű projektek adatait vizsgáljuk, az adatok nyilvánosságának, dokumentáltságának helyzete már korántsem ilyen jó. Bár a kutatóknak adódnak lehetőségei az adatok látható közzétételére – ritkábban a folyóiratoknál digitális mellékletekként, gyakrabban a generalista repozitóriumokban mint a *Figshare* vagy a *Zenodo*, alkalmanként

intézményi adatrepozitóriumokban, de ezekkel a lehetőségekkel ritkán élnek. Talán azért, mert munkaigényes az adatokat közzétehető formába hozni, talán mert hiányzik a motiváció, nincs meg a megfelelő ösztönző rendszer. A szakterület legfontosabb folyóiratainak egyike, az európai *Astronomy & Astrophysics* szerzőknek szóló útmutatójában leírja, hogy a cikkekhez kapcsolódó elsődleges adatok elhelyezését a CDS végzi, és az archiválható állományok létrehozásában szükség esetén segítenek is. Mindazonáltal adatelhelyezési kötelezettséget az útmutató nem említ.

Egy hazai kutatóhely éves publikációit végignézve az ADS-ben arra a következtetésre juthatunk, hogy bár a cikkek többségénél az adatbázis jelöl valamiféle adatkapcsolatot, ezek többségét a CDS által a cikkekből kigyűjtött objektumnevek, esetleg táblázatok adják. A hazai csillagászati kutatóhelyek esetében egyre nő a nagy projektek adatait használó cikkek száma – ezeknél a megfigyelési adatok elérhetősége és újrafelhasználhatósága megoldott, így a cikkhez használt adatok is elérhetőek. A hazai teleszkópokkal végzett észlelések adatai viszont kevéssé felelnek meg a FAIR kritériumoknak, de ez gyaníthatóan így van világszerte más kisebb obszervatóriumok esetében is.

5.) A könyvtárak szerepe

Kis mennyiségű adat a közleményekben megjelentethető. Konkoly Thege Miklós rendszeresen közölte az ógyallai csillagvizsgálójában végzett megfigyeléseinek adatait az Akadémia kiadásában, például az „Értekezések a mathematikai tudományok köréből" sorozatban. Az – akkori mércével – sok adatközlés úgy tűnik, terheket jelentett az Akadémiának: Konkoly levélben panaszolta el Eötvös Lorándnak az őt ért kritikákat (Vargha 2001).

Konkoly obszervatóriuma a későbbiekben is adott ki kötetekben megfigyelési anyagot (Tass 1925). A publikált megfigyelési anyagok papíron megtalálhatóak a könyvtárakban, digitalizálva pedig a könyvtárak által üzemeltetett repozitóriumokban.

A XX. század második felére az obszervatóriumi kiadványsorozatok szerepe csökkent (bár a Konkolyról elnevezett hazai obszervatórium kiadványsorozatában még ez időszakban is sok adatot közöltek). Az adatok jobbára az ekkorra uralkodóvá vált folyóiratok mellékleteibe

(*Supplement Series, Ergänzungshefte*) szorultak. 1993 és 1998 között az *American Astronomical Society* folyóirataihoz CD-ROM mellékletet adtak ki, ezután már az adatok a hálózatra kerültek. CD-ROM-os formában indult a *Journal of Astronomical Data*, majd ez a kiadvány is a webre került.[5]

A Gutenberg-korszakban keletkezett, ránk maradt megfigyelési eredmények jelentős része a könyvtárak polcain, és papírról digitalizálva a repozitóriumokban találhatóak, ám ezek az adatok gépi feldolgozásra nem alkalmasak (átalakítás nélkül). Mint fentebb láthattuk, a kisebb volumenű, modern adatok is legtöbbször a folyóiratok közvetítésével kerülnek a nagy adatközpontok adatbázisaiba.

Az adatok „könyvtári útjának" fontosságát mutatja, hogy az adatokat (szakirodalmi kapcsolataikkal együtt) kereshetővé tévő adatbázisok is szoros kapcsolatban állnak a könyvtárakkal. A NASA/SAO ADS átfogó tudományterületi bibliográfiaként teszi láthatóvá a cikkekhez kapcsolódó adatokat, míg a CDS erős könyvtáros csoportot fenntartva, az adatok oldaláról mutatja meg az adatokat említő tudományos cikkeket.

– *Library and Information Services in Astronomy (LISA)*
A csillagász könyvtárosok konferenciasorozata a LISA 1988-ban indult, a legutóbbi, kilencedik összejövetel 2021-ben volt (a járványhelyzet miatt online). A konferenciák visszatérő témája a kutatási adatok kezelése, a nyílt tudomány. A konferenciák kiadványköteteiben a kutatási adatok kezelésével foglalkozó cikkek száma az elmúlt két évtizedben a 2002-es hétről tizenötre nőtt.

A legtöbb esetben külön szekció foglalkozott a kutatásiadat-kezeléssel, 2017-ben annyira hangsúlyos volt a téma, hogy három szekciót töltöttek meg az adatkezeléssel kapcsolatos előadások.

– *Tudománymetria – adathasználat követése az ESO-ban*
Az Európai Déli Obszervatórium könyvtára vállalta fel az intézményben végzett megfigyelések publikálásának szakirodalmi követését, mérését. A *telbib* projekt szakirodalmi könyvtárosi ellenőrzéssel megerősített

5 JAD: http://journalofastronomicaldata.be/

adatbányászattal követi a távcsőidő-pályázati azonosítók szereplését a publikációkban (Grothkopf, 2018).

6.) Közösségi tudomány

Ez a terület az, ahol a kívülálló leginkább felmérheti a csillagászati kutatásiadat-kezelés helyzetét. A csillagászati közösségi tudomány hazai előképe Konkoly Thege Miklós meteorészlelő hálózata 1875-ből (Bartha, 1988). 1911-ben alakult az *American Association of Variable Star Observers*, többségében amatőr csillagászok megfigyeléseinek szervezésére, az adatok összegyűjtésére.

A modern közösségi tudomány egyik legfontosabb platformja a *Zooniverse* portál. A névadó első alkalmazás, a *Galaxy Zoo* önkénteseinek száma jelenleg közel 87 ezer. Nem valósulhatott volna meg a projekt az SDSS és a többi időközben bevont égboltfelmérés adatainak nyilvánosságra hozatala nélkül.

Még korábbi az emberi agy jelfeldolgozó képességei helyett az otthoni számítógépek szabad processzorkapacitását felhasználó *SETI@home*.

7.) Összefoglalás

Negyven évnyi kooperatív kutatásiadat-kezelési gyakorlat birtokában a csillagászat ma olyan tudomány, ahol saját megfigyelések nélkül, kizárólag publikusan elérhető adatokra alapozva lehet tudományos kutatást folytatni. A csillagászatban (és egyre több más tudományterületen is) sok a fóka [az adat], és kevés az eszkimó [a kutató]. Chris Lintott nyilatkozta a Time magazinnak: „In many parts of science, we're not constrained by what data we can get. We're constrained by what we can do with the data we have."[6] Ahhoz, hogy a digitális korszakban keletkező adatáradatot a feladatra vállalkozók feldolgozhassák, megfelelő dokumentáltságra, és nyilvánosságra van szükség.

6 https://web.archive.org/web/20100331061938/http://www.time.com/time/health/article/0,8599,1975296,00.html

Irodalom

Bartha L., 1988. Az első magyarországi észlelőhálózat, Meteor, 7/8. 7–11.

Berriman G.B. et al., 2020. The International Virtual Observatory Alliance (IVOA) in 2020. arXiv:2012.05988 [astro-ph.IM]

Greisen, E.W., 2003. FITS: A Remarkable Achievement in Information Exchange. In: Heck, A. (eds) Information Handling in Astronomy - Historical Vistas. Astrophysics and Space Science Library, vol. 285. Springer, Dordrecht. https://doi.org/10.1007/0-306-48080-8_5

Grothkopf, U., Meakins, A., Bordelon, D., 2018. ESO telbib: learning from experience, preparing for the future. ArXiv:1806.08746 [astro-ph.IM] https://doi.org/10.1117/12.2311667

Holl, András, 2020. A kutatási adatok dokumentálását elősegítő szoftverek. In: Networkshop 2020. Országos Online Konferencia. 2020. szeptember 2–4. HUNGARNET. pp. 7–12. http://real.mtak.hu/119187/ https://doi.org/10.31915/NWS.2020.1

Holl, András, 2022. IBVS Data Files – Case Study of a Small Data Journal. Bulletin of the AAS, 54 (2). https://doi.org/10.3847/25c2cfeb.f14d187b

Novacescu, J. et al., 2018. Elevating MAST-Data Publications with Digital Object Identifiers (DOIs). Jenny EPJ Web Conf., 186, 10003. https://doi.org/10.1051/epjconf/201818610003

Tass, Antal, ed., 1925. Photometric Observations of Variable Stars = Photometrische Beobachtungen Veranderlicher Sterne = Változó csillagok photometrikus megfigyelései. Publications of the Royal Hungarian Astrophysical Observatory, Foundation of Konkoly in Budapest, 2. Konkoly-Alapítványú Budapesti Magyar Királyi Csillagvizsgáló Intézet, Ógyalla – Budapest. http://real-eod.mtak.hu/9417/

Vargha Domokosné: Konkoly Thege Miklós magyar nyelvű írásai. Magyar Tudomány, 2001/7. 867. o. http://www.matud.iif.hu/01jul/vargha.html http://real-j.mtak.hu/158/1/MATUD_2001.pdf#Page=895

Williams, R. D., Seaman, R., 2006. VOEvent: Information Infrastructure for Real-Time Astronomy. ADASS XV. https://ui.adsabs.harvard.edu/abs/2006ASPC..351..637W/abstract