Magyar Tudomány 184(2023)7, 851–863
DOI: 10.1556/2065.184.2023.7.4

MI AZ ÁBRA? SZOCIOLÓGIAI FELMÉRÉS A KUTATÁSI ADATOK KEZELÉSÉRŐL, ÉS AZ EREDMÉNYEK ÁTFORDÍTÁSA AZ INFORMATIKA NYELVÉRE

WHAT'S THE SITUATION? A SOCIOLOGICAL SURVEY ON RESEARCH DATA MANAGEMENT AND THE TRANSLATION OF ITS RESULTS INTO IT LANGUAGE

Egyed-Gergely Júlia[1], Gárdos Judit[2], Horváth Anna[3], Meiszterics Enikő[4], Tóth Zoltán[5], Vajda Róza[6]

[1]PhD, gyűjteménygondozási munkatárs,Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központ, Budapest
egyed-gergely.julia@tk.hu
[2]PhD, a Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központ vezetője
gardos.judit@tk.hu
[3]gyűjteménygondozási munkatárs, Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központ, Budapest
horvath.anna@tk.hu
[4]gyűjteménygondozási munkatárs, Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központ, Budapest
meiszterics.eniko@tk.hu
[5]mérnökinformatikus, fejlesztőmérnök
ELKH Számítástechnikai és Automatizálási Kutatóintézet Elosztott Rendszerek Osztály, Budapest
toth.zoltan@sztaki.hu
[6]gyűjteménygondozási munkatárs, Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központ, Budapest
vajda.roza@tk.hu

ÖSSZEFOGLALÁS

Hogyan építsünk adatrepozitóriumot? Mit kell tudnia egy adatrepozitóriumnak, és honnan szerezzünk erről információkat? Milyen nehézségekkel kell számolni, és mennyire felkészültek a potenciális felhasználók a szolgáltatás igénybevételére?

Az Eötvös Loránd Kutatási Hálózat (ELKH) országos szolgáltatást nyújtó központi adatrepozitóriumot fejleszt, amely a kutatási adatok hosszú távú tárolására, archiválására ad majd lehetőséget az ELKH kutatói közössége számára. Annak érdekében, hogy a készülő infrastruktúra minél jobban kiszolgálja a hálózat kutatóit, kutatócsoportjait, komplex felmérést készítettünk a kutatók archiválással kapcsolatos igényeinek megismerése céljából. Tanulmányunkban az ELKH Adatrepozitórium Platform (ARP) megalapozását biztosító felmérés eredményeit és ezeknek a repozitóriumfejlesztéshez történő felhasználását mutatjuk be.

A vizsgálat során az ELKH-intézetek kutatóinak kutatásiadat-kezelési, -tárolási és -megosztási gyakorlatait, igényeit, repozitóriumhasználattal összefüggő tapasztalatait tártuk fel. Kérdőíves megkereséssel és interjús beszélgetésekkel közel kétszáz kutatót – és rajtuk keresztül számos kutatócsoportot – értünk el, az adatfelvétel a kutatási hálózat minden intézményét és diszcip-

línáját lefedte. Ezzel a vizsgálat az eddigi talán legszélesebb körű ilyen típusú magyarországi felmérés lett.

Tanulmányunkban a nyílt tudomány elvével kapcsolatos nemzetközi trendekből kiindulva mutatjuk be az adatgazdálkodás hazai helyzetét. A kutatói beszámolókból, véleményekből kiindulva a legnagyobb magyar kutatási hálózatban uralkodó viszonyok bemutatásakor kitérünk a speciális kihívásokra, hiányosságokra és a megismert jógyakorlatokra. Végül bemutatjuk, hogy a felmérés eredményeiből hogyan készült szoftverspecifikáció, azaz hogyan támogatta az igényfelmérés a számítástechnikai követelményrendszer előállítását a fejlesztés alatt álló adatrepozitóriumhoz.

ABSTRACT

How to build a data repository? What does a data repository need to know and how should we find out about it? What are the difficulties and how prepared are the potential users?

The Eötvös Loránd Research Network (ELKH) is developing a central data repository to provide for long-term storage and archiving of research data produced by ELKH researchers. In order to best serve the research community, a complex survey was carried out to identify the needs of researchers related to archiving. This paper presents the results of a survey realized in the framework of the ELKH Data Repository Platform (ARP) project, and shows how they were used in the process of developing the repository.

The survey explored the practices and needs of researchers within the ELKH related to the management, storing and sharing of research data as well as their experiences concerning the use of repositories. Via questionnaires and interviews we reached nearly two hundred researchers—individuals and teams—covering all the institutions and disciplines represented in the research network. The survey is thus the most comprehensive data collection ever undertaken on research data management in Hungary.

This paper presents the current situation of data management in the largest Hungarian research network in the context of international trends with respect to open science. In presenting the state of affairs in the largest nation-wide research infrastructure, we discuss the challenges and gaps as well as some good practices drawing on the accounts and opinions of researchers. Finally, we show how the results of the survey were translated in terms of software specifications in order to support the IT requirements framework of the project.

Kulcsszavak: adatrepozitórium, igényfelmérés, kutatási adat, adattárolás, szoftverspecifikáció

Keywords: data repository, needs assessment, research data, data archiving, software specifications

BEVEZETŐ

Egy kutatási adatrepozitórium hatékony működtetéséhez elengedhetetlen a felhasználók, a különböző tudományterületeken tevékenykedő kutatók adatkezelési gyakorlatainak és igényeinek megismerése. Írásunkban az Eötvös Loránd Kutatási Hálózat (ELKH) adatrepozitóriumának tervezéséről, ezen belül is egy szociológusok és informatikusok szoros együttműködésében megvalósult kutatásról

számolunk be. Bemutatjuk felmérésünk néhány fontos eredményét, és azt is, hogy ezek hogyan használhatók fel, fordíthatók át digitális megoldásokká elektronikus kutatási szolgáltatások tervezésekor.

Nem csak a jelen írásban bemutatott felmérés foglalkozik a kutatási adatok tárolásának és megosztásának kérdéseivel. A Springer Nature kiadó és a Figshare repozitórium közös, *A nyílt adatok helyzete* című kutatásában 2016 óta gyűjt évente adatokat a világ minden tájáról a kutatók adatokkal, adatkezeléssel és -megosztással kapcsolatos véleményeiről, gyakorlatairól. Az éves jelentések mellett a kutatás kérdőívét és az adatbázisokat is nyíltan hozzáférhetővé teszik a Figshare repozitóriumából (URL1).

Az Európai Bizottság Kutatási és Innovációs Főigazgatósága 2021 júniusában kezdeményezte az *Európai kutatási adatvilág* című felmérés elindítását, melynek keretében kutatók és adattárak gyakorlatairól, tapasztalatairól gyűjtöttek információkat az EU tagállamaiban, a Horizon 2020 társult országaiban és az Egyesült Királyságban. A tanulmány (URL2) kulcsfontosságú lesz az Európai Bizottság számára a nyílt adatokkal kapcsolatos gyakorlatok általánosabbá tételét célzó szakpolitikák és támogatási intézkedések kidolgozásához (European Commission, 2021).

Magyarországon jelenleg több kutatóhely, tudományos intézmény szeretne adatrepozitóriumot létrehozni, vagy már dolgozik is annak felállításán. Ezekben az intézményekben ugyancsak felmérésekkel alapozzák meg a fejlesztést: az ELKH Adatrepozitórium Platform projektje mellett többek között a Debreceni Egyetem Egyetemi és Nemzeti Könyvtára (Száldobágyi, 2022) és az *Eötvös Loránd Tudományegyetem* Egyetemi Könyvtár és Levéltár Oktatás- és Kutatástámogatási Osztálya (Móring, 2022) is végzett ilyen kutatást.

A FELMÉRÉSRŐL

A kérdőíves felmérésre 2022. február eleje és március vége, az interjús beszélgetésekre 2022. január vége és május eleje között került sor, a Covid19-világjárvány miatt kialakult, a személyes találkozások minimalizálását megkövetelő helyzet miatt többnyire *online* formában. A Társadalomtudományi Kutatóközpont (TK) és a Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI) munkatársai közösen koordinálták és végezték a munkát valamennyi ELKH intézmény megkeresésével és bevonásával. Kérdéseinket az összes intézménynek kiküldtük, közel 130 kitöltött kérdőív érkezett vissza. A kérdőíves felmérés során a nagyobb ívű trendekre, jellemző számadatokra, könnyen jellemezhető gyakorlatokra voltunk kíváncsiak, míg az interjúk lehetőséget adtak arra, hogy kitérjünk a kutatói munkafolyamatok egyes mozzanataira, és ezáltal teljesebb képet kapjunk a kutatók elgondolásairól, igényeiről, és ezen keresztül az intézményi és strukturális

lehetőségekről, akadályokról. Az összesen 52 interjú során ELKH-kutatókkal, intézményvezetőkkel, néhány esetben kutatóval szorosan együttműködő egyéb szakértővel beszélgettünk.

MIT ÉS MENNYIT ŐRIZZÜNK MEG?

Az igényfelmérés tanúsága szerint az ELKH intézményeiben végzett kutatások során jellemzően új kutatási adatok keletkeznek, bár vannak olyan kutatók és területek is, akik és ahol csak feldolgozott, más kutatók által keletkeztetett adatokkal dolgoznak. Tudományterületenként és azon belül az egyes intézmények szintjén, néhány esetben pedig akár a kutatócsoportok különféle projektjei szerint is eltérő a keletkező nyers és/vagy feldolgozott adatok mennyisége, formátuma, illetve az, hogy ezeket az adatokat milyen időtávra visszamenőleg, milyen mennyiségben tárolják, illetve archiválják – azaz mennyi és milyen formátumú adatot helyeznének el az ELKH készülő adatrepozitóriumában.[1]

A kutatások során keletkező adatmennyiség nagy szórást mutat: van, ahol egyetlen mérés alatt keletkezik annyi adat, amennyi máshol évek, akár évtizedek alatt gyűlik össze. Általánosságban megállapítható, hogy a keletkező legkisebb adatmennyiség néhány tíz kilobájtnyi, míg a legnagyobbak között mérésenként több terabájtos adatcsomag is szerepel. Nemcsak a hálózaton, hanem az intézményeken és az egyes kutatócsoportokon belüli eltérés is jelentős. Van több olyan kutatóközpont, ahol az egyik kutatócsoport vagy projekt 2–100 megabájt közötti mérési adatokkal dolgozik, míg a másik több terabájtosokkal. Mindeközben a kutatók kevesebb mint fele (44%) ismer olyan kutatási adatrepozitóriumot, amely az ő vagy kutatócsoportja számára szolgáltathatna, ráadásul az említettek közül több nem is kifejezetten adat-, inkább publikációs repozitórium, esetleg online adatbázis vagy egyéb, adatok megosztására szolgáló platform.

MIÉRT ÉRDEMES ARCHIVÁLNI?

A beszélgetésekben az adatarchiválás legfőbb motivációjaként az adatbiztonság és az adatok visszakereshetősége fogalmazódott meg. Több adatintenzív kutatási területen nem létezik hazai ágazati repozitórium, ugyanakkor egyes technológiailag igényes és fejlett területeken azt láttuk, hogy az adatmegosztás

[1] Az interjúk tanúsága alapján az ELKH kutatói által jelenleg tárolt és/vagy archivált adattípusok és -mennyiségek ugyanakkor nem feltétlenül egyeznek meg teljes mértékben azzal az adatmennyiséggel és adatstruktúrával, melyet akkor archiválnának, ha lenne erre egy központi megoldás, mint amilyen a készülő adatrepozitórium.

korszerű lehetőségeivel élve egy-egy kutatóintézet már bekapcsolódott a nemzetközi vérkeringésbe. Inkább természettudományos területekre jellemző, hogy a repozitóriumhasználat az adattárolási és -megosztási gyakorlatok szerves részét képezi. Alapvetően máshol is nyitottságot tapasztaltunk, de fenntartásokkal: az adatféltés, az adminisztratív többletfeladatok vagy az anonimizálás humán és pénzügyi erőforrásigénye merült fel mint korlátozó tényező.

Van, amikor a kutatómunka során használt adatok jellege, illetve az erről való vélekedés fogja vissza az adatrepozitóriumok iránti érdeklődést. Nem tartják szükségesnek a használatát az olyan területeken (matematika, kvantumkémia) dolgozó kutatók, ahol nincs nyers adat, vagy ahol a végeredmény számít, nem pedig a munkafolyamat során keletkező nagy mennyiségű adat (számítások). Van, hogy azért nem tűnik érdemesnek megőrizni az adatokat, mert gyorsan elavulnak, érvényüket vesztik, így csak feleslegesen foglalnák a helyet. Más eset, amikor nem világos, egyáltalán mi tekinthető kutatási adatnak, mivel a gyűjtésen kívül nem történt beavatkozás. A terepmunka során keletkezett adatokat többen rendezetlennek vagy éppen túl személyesnek (például: jegyzetfüzet, terepnapló) tartják ahhoz, hogy közhasználatú adatbázisba kerülhessenek. Megjegyzendő azonban, hogy egy megfelelő paraméterekkel rendelkező repozitórium az ilyen, jellegükben strukturálatlan adatok gyűjtésére, rendszerezésére is lehetőséget biztosít.

KÖTELEZETTSÉGEK ÉS LEHETŐSÉGEK

Az ELKH kutatóinak többsége viszonylag szabad kezet kap intézményétől – annak minden előnyével és hátrányával együtt – adatainak kezelését, kutatás alatti és utáni tárolását illetően. Néhány, szigorú szabályokat előíró intézmény és néhány, erősen szabályozott rendszerben előálló adattal dolgozó terület (például mérőműszerekkel dolgozó, azokon adatot keletkeztető kutatások) kivételével a legtöbb tudományág művelői saját adattárolási és -kezelési gyakorlatokat alakítottak ki a napi kutatómunka során.

A kutatói adattárolás a szabályozáson kívül természetesen nagyban függ az intézetek által biztosított lehetőségektől is. Kutatásunkból kiderült, hogy a válaszadók alig több mint fele érzi úgy, hogy intézményében teljes mértékben megoldott a munkavégzéshez, kutatáshoz szükséges adattárolás. Az interjúkban nyilatkozó kutatóktól pedig tudjuk, hogy az intézményi és az ELKH-szintű lehetőségekhez képest a saját megoldások (saját gépen, saját külső háttértárolón, kutatócsoport által kialakított és működtetett online felületen való tárolás) az elterjedtebbek.

A kutatás közbeni és utáni adattárolásnál egyaránt jellemző tehát az általános szabályozás hiánya; az interjúk alapján kisebbségben vannak azon intézmények, amelyekben intézményi előírással vagy ajánlással látnák el a kutatókat. A kérdő-

íves felmérés eredménye szerint a kutatási adatok archiválásának lehetőségét a válaszadók csupán 40%-a tartja intézményén belül teljesen megoldottnak. Ennek megfelelően a legkülönbözőbb megoldások születnek, kezdve attól, hogy a kutató a saját gépén, mappákban tárolja a munkája során keletkező adatokat, a mérőműszer automatikus mentésén át egészen a többszörös gépi, felhőalapú vagy külső háttértárolós megoldásokig. Az interjúalanyok alig fele említette, hogy valamilyen nyilvános adatbázisba vagy adatrepozitóriumba tölti fel adatait, legalább eseti szinten. A kérdőíves felmérésből is az derül ki, hogy a vezető megoldások a külső adattárolón és a saját gépen való mentés, miközben viszonylag ritka a repozitóriumi elhelyezés *(1. ábra).* Utóbbi jellemzően nem automatikusan, minden kutatás esetében történik, legfeljebb alkalomszerűen, konkrét helyzetekben élnek a lehetőséggel.

1. ábra. Archiválás a kutatás lezárulta után az ELKH intézményeiben (%-os megoszlás a kérdőíves felmérés eredményei alapján) (saját szerkesztés)

Felmérésünkből kiderül továbbá, hogy többnyire nincsen kötelező érvényű, egységes, következetes metaadatolás a kutatásokhoz rendelve. A metaadatok rögzítése területenként, intézményenként nagyon egyenetlen, általában nem szabványszerű és nem is teljes körű: inkább úttörő gyakorlatokkal, semmint általános, rögzített sztenderdek szerint építkező, bejáratott adatbázisokkal találkoztunk. Ugyanakkor, az intézményes elvárások hiánya ellenére előfordul alapos és módszeres adatnyilvántartás. Vannak területek, ahol (előírás szerint vagy nem kötelező jelleggel) a publikációk magukban foglalják az adatokat, azok jegyzékét, illetve a keletkezéstörténetükre vonatkozó információkat. Ám egy-egy kutatás

nyomát olykor pusztán a szakavatottak számára elérhető kutatási jelentések őrzik, amelyek nem tartalmaznak rendszerezett és kellő mélységű leírást a kutatáshoz kapcsolódó adatokról. Minimum szintet jelentene a projektnyilvántartás, csakhogy többen ennek módszereit is csak hírből, illetve külföldi példákból ismerik.

MIÉRT KELL ADATOT TÁROLNI, ÉS MIÉRT JÓ MEGOSZTANI?

Amilyen rengetegféle archiválási gyakorlat létezik, olyan sokféle oka lehet annak, hogy mit és miért archiválnak az egyes kutatók. Gyakori, hogy egy-egy kutatás lezárultával egyszerűen „megmaradnak" az adatok a számítógépen. A tudatosabb kutatók rendszerezik is anyagaikat a későbbi visszakereshetőség végett, a még tudatosabbak biztonsági mentést is készítenek, minimalizálandó az adatvesztés esélyét. Viszonylag kevesen vannak, akik vagy publikációs elvárásnak engedelmeskedve (több területen jellemző, hogy épp most, a legutóbbi időkben jelent meg ez a fajta igény), vagy általában a nyílt tudomány elvének jegyében (tudományos hozzájárulásként) töltik fel az adataikat nyilvános repozitóriumba – utóbbiak elsősorban azok közül kerülnek ki, akik munkájuk során maguk is adatletöltők.

Ahogyan a tárolásra, úgy az adatok megosztásának módjára sincs szabályozás számos ELKH-intézményben. A kérdőívre adott válaszokból kiderül, hogy bizonyos gyakorisággal, valamilyen (intézményen kívüli) körrel a kutatók túlnyomó többsége (kilenctizede) megosztja az adatait, vagy a kutatás közben, vagy annak lezárultával. A megosztás módja azonban – mind a kérdőíves, mind az interjús felmérés eredményei alapján – igen változatos képet mutat. Általában, bár korántsem mindenhol, létezik valamilyen közös tárhely, jellemzően intézeti felhő, ugyanakkor a kutatók leggyakrabban informálisan, főként e-mailben, Google Drive-on vagy Dropboxban osztják meg a kutatási adatokat egymással és azokkal, akikkel a közös munka révén kerülnek kapcsolatba, esetleg szívességből másokkal is. Néhány területen bizonyos érzékeny, nagyon védett vagy féltett adatokat nem enged kijutni az intézetvezetés, vagy a kutató nem szeretné azokat online módon továbbítani. Ezekben az esetekben kutatószobás elérés vagy személyes pendrive-os vagy külső merevlemezes megosztás a bevett forma.

Mivel a kutatóközösségek általában maguk kénytelenek meghatározni a módszereiket, átlátható szabályok helyett sokszor egyéni szokások szabják meg az adatgazdálkodást. Ez alkalmanként adatvesztéshez, még több esetben adatláthatatlansághoz vezet. A megkérdezett kutatók többsége szerint a kutatói szabadság megőrzése mellett jó volna kiszélesíteni az ELKH biztosította lehetőségeket, mindenekelőtt az adatok tárolásának, visszakereshetőségének és megoszthatóságának biztosítása érdekében, hogy megismerhessék más (akár saját intézményen belüli) kutatók adatait, valamint, hogy eleget tudjanak tenni a publikációs elvárásoknak és a nyílt tudomány kívánalmainak.

ADATARCHIVÁLÁSI ISMERETEK ÉS ATTITŰDÖK

A saját kutatási anyagok rendbe- és közhasznúvá tétele több helyen forrás-, infrastruktúra- vagy ismerethiányba ütközik. Technikailag összetettebb adatmegőrzési módszerek esetén egy korszerű adatbázis fenntartásához és működtetéséhez szakszemélyzet (digitális levéltáros, adatgazdász) igénye is felmerül. Egyértelmű, hogy nem fogható minden probléma egyszerűen a pénzhiányra vagy a külső körülményekre. Az intézmények működése, vezetése mind pozitív, mind negatív irányban erősen befolyásolja a kutatás folyamatának átláthatóságát és az eredmények közhasznúvá tételét. Interjúalanyaink meglepően gyakran számoltak be arról, hogy intézményük semmiféle adatkezelési szabályzattal nem rendelkezik. A korszerű adatmenedzsmentet illetően mindenekelőtt tehát az intézményi elvárások lazasága, nemléte és ebből következően a repozitóriumok, adatbázisok használatában való jártasság hiánya látszik a legfőbb akadálynak.

Az adatbiztonság ügye többféle jogi kérdést is felvet. Ide tartoznak a szerzői jogokkal, illetve az adatok tulajdonlásával kapcsolatos aggodalmak, kétségek, illetve a vonatkozó ismeretek vagy a gyakorlati megvalósítások hiánya.

Több beszélgetőpartnerünk kárhoztatta a kutatókat, amiért nem szívesen osztják meg a nyers adataikat. Van, aki életkori megosztottságot tapasztal, miszerint az idősebb, „technikailag értetlen" generáció kevésbé nyitott az átállásra. Nehezen vitatható ugyanakkor, hogy a „kutatói szabadságra" történő gyakori hivatkozás elméletileg, módszertanilag és/vagy etikailag megalapozott állásponton alapulhat. Innen nézve értelmezési kételyeket vetnek fel az eredeti adatfelvétel szempontjaitól eltávolodó másodfeldolgozások (lásd erről bővebben: Gárdos, 2011). Ezen felül humánkutatásoknál sokszor nem egyértelmű, hogy tisztességes-e egyáltalán a kutatásban részt vevőkkel szemben a tőlük származó információ továbbadása. Olyan meggyőződéssel is találkoztunk, miszerint a nem átlátható, közös használatú adatbázisok általában veszélyt jelentenek az intézményes irányítás és a felelősségmegoszlás szempontjából.

AZ IGÉNYFELMÉRÉS LEFORDÍTÁSA KÖVETELMÉNYRENDSZERRE

Az igényfelmérésnek a tudományos érdeklődésen túl alapvető célja volt az új repozitóriumi rendszer fejlesztőinek támogatása a szoftverrendszer követelményspecifikációjának meghatározásában. Egy általános szoftverspecifikáció erősen formalizált dokumentum, előállítása közvetlenül az interjúk szövegéből nem egyszerű feladat. Az interjúalanyok nem a programozás által megkívánt formalizmusnak megfelelően fogalmazták meg az igényeiket, amelyek sok esetben nem igazán konkrétak, mivel a kutatók hajlamosak nagyon magas absztrakciós szintről tekinteni egy-egy témakörre. További problémát jelentett, hogy ugyanazt a követelményt

több interjú során is megemlíthették a kutatók, csak más megfogalmazásban, így ezeknek az összesítését is meg kellett oldani. A szoftverfejlesztés szempontjait szem előtt tartva szükségessé vált tehát egy köztes nyelv kialakítása és valamiféle problémakör-klasszifikáció használata. Ilyen egyéni tapasztalatok problémaorientált lefordítását lehetővé tevő köztes nyelv, specifikációs eszköz gyanánt elsősorban az ún. felhasználói történeteket (user story) (URL3), pontosabban az ezekből kialakított használati eseteket (use case) alkalmaztuk. Ez egy hibrid módszertani eszköz interakciók leírására, melynek használata nem korlátozódik szigorúan a számítástechnikai környezetre. Mint az alábbiakból kiderül, a használati esetek definiálása alkalmas a feladatkörök dekompozíciójára és ezen keresztül bizonyos értelemben vett klasszifikációjára, ami átláthatóbbá teszi a fejlesztés céljait.

INTERJÚKBÓL FELHASZNÁLÓI TÖRTÉNETEK

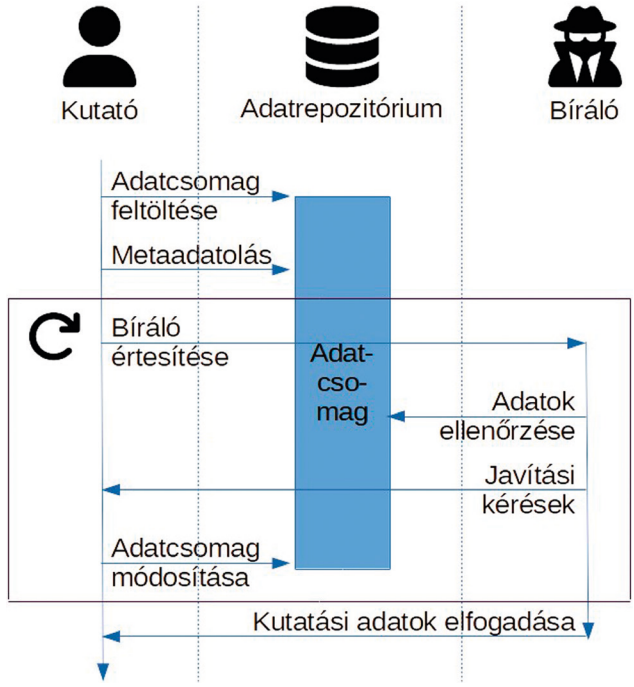
A módszer használatának előkészítéséhez a vezetett beszélgetések megpróbálták feltérképezni, hogy milyen felhasználói feladatok és interakciók merülnek vagy merülhetnek fel egy létező/elképzelt repozitóriummal kapcsolatban. A beszélgetések alapján meghatároztuk a tényleges kutatói feladatokat, a szerepköröket és a feladatok megoldásához szükséges lépéseket. Megállapítottuk, hogy hol szükséges adatrepozitóriumok használata, az adatrepozitóriumokkal kapcsolatban a kutató kivel kommunikál, hova, milyen adatokat tölt fel, milyen interakciókat végez, vagy amennyiben erre a jelenleg általa használt eszközkészlet nem alkalmas, milyen interakciókat szeretne elvégezni. A felhasználói történetek egyszerűen megfogalmazott, nem formális tevékenységleírások, vagyis esetünkben a kutató megfogalmazta, hogy mit szokott csinálni, vagy mik lennének az igényei.

FELHASZNÁLÓI TÖRTÉNETEKBŐL HASZNÁLATI ESETLEÍRÁSOK

Ezeknek a felhasználói történeteknek a formalizáltabb lebontásai a használati esetleírások. Ezek lényegüket tekintve olyan forgatókönyvek, amelyekben adott szereplők/szerepkörök egymás utáni interakciói vannak rögzítve. A forgatókönyv felbontása a módszertan szempontjából tetszőleges, amitől alkalmassá válik nagyon bonyolult kommunikációs kapcsolatok elnagyolt vagy akár részletekbe menő leírására is.

Idevágó példa egy publikációhoz tartozó adatok feltöltésének és elbírálásának folyamatleírása *(2. ábra).* A felhasználói történet kimerülhet abban, hogy a kutató megállapítja: „a publikáció elfogadásához szükséges adataimat feltöltöm a repozitóriumba, ahol a bíráló megnézi, leellenőrzi azok elérhetőségét, a formai kritériumoknak való megfelelését, és engedélyezi, hogy a publikáció megjelen-

jen”. Ez a tevékenység a használati esetleírásban két szereplőt azonosít (kutató – K, bíráló – B), illetve a szoftverkörnyezetet, ahol a forgatókönyv lezajlik. Vannak előfeltételei (publikáció zajlik, ahol a publikálás folyamatához szervesen kötődik a kutatási adatok közreadása, vagyis K rendelkezik kutatási adatokkal), és van eredménye is a folyamatnak (a publikációs folyamat továbbmegy, vagy visszakerül egy olyan állapotba, ahol ismételten szükséges az adatok ellenőrzése annak megfelelően, hogy B hogyan ítélte meg azok megfelelését az elvárásokhoz képest). Vannak továbbá lépései is: K feltölti az adatokat a repozitóriumba, jelzi valamilyen módon, hogy az adatok elérhetőek, B ellenőrzi az adatokat, majd dönt a publikációs folyamat további menetéről. Meg kell jegyezni, hogy bár a példánkban az interakciók sorozata lineáris, a lépések végrehajtása nem feltétlenül ilyen. A döntési helyzeteknek megfelelően a felhasználási esetek szokásos ábrázolása a felhasználási esetdiagram, amely egy irányított gráf megfeleltetése a forgatókönyv lépéseinek, ahol a döntésnek megfelelő interakciók vezethetnek korábban már végrehajtott lépésekre is. Ha a példánkat úgy módosítjuk, hogy a kutató addig változtatja a kutatási adatok leírását, amíg a bíráló el nem fogadja az aktuális állapotot, rögtön bevezettünk a leírásban egy visszacsatolást, egy ismétlődő mintát, ahol a szereplők nem, csak a rendszer állapota (az adatok leírása) változik a lépések során.



2. ábra. Az adatfeltöltés és a *peer review* folyamatábrája (saját szerkesztés)

A példán megmutatkozik a módszertan előnye, azaz, hogy nem szükséges minden aktor részletes ismerete a folyamatleírásban, és az is, hogy a leírás részletessége igazodhat annak céljához. Ebben az esetben elnagyolt például a környezet leírása. A készülő rendszer számára, amennyiben az nem szándékozik támogatni a kutató-bíráló közvetlen kommunikációt (mert például erre hivatalos csatorna van egy publikálás során), teljesen mellékes, hogy a bíráló hogyan jelzi a kutató számára, hogy korrekcióra van szükség a feltöltött adatokkal kapcsolatban. Elnagyolt továbbá az esetleírásban szereplő lépések kifejtése is. Adott környezetben mást-mást jelenthet például a „kutatási adatok feltöltése" lépés, amely több, akár meglehetősen komplex módon megadható, egyedi használati esetre bontható. Például egy következő granulációs szint lehet a feltöltés lépéseinek részletezése: K az adatokat adott formátumban webes interfészen keresztül feltölti a repozitóriumba, majd kitölti a szükséges metaadatokat, és nyilvános állapotúra állítja a leírást, hogy tetszőleges személy férhessen hozzá. Egészen részletes leírásig is le lehet akár menni, ahol például a webes feltöltés van tovább bontva, miszerint milyen gombokra kell kattintani a feltöltés során, milyen nyilatkozatokat kell elfogadni stb.

HASZNÁLATI ESETLEÍRÁSOKBÓL SZOFTVERKÖVETELMÉNYEK

Az interjús felmérés során meghatároztuk a felhasználók által elvárt interakciókat. A rendszer alapját egy nyíltan hozzáférhető platform, a Dataverse (URL4) adja, amely a jelenleg is működő hazai nyílt, multidiszciplináris kutatási repozitórium, a CONCORDA (URL5) platformja is egyben. Ebből kinyertünk egy meglehetősen részletes szoftverspecifikációt, amely alapját a felhasználói és adminisztrátori kézikönyv (dokumentációanalízis), a működő szoftver elemzése (funkcionális dekompozíció, interfészanalízis, mérnöki visszafejtés) adta. Az így kapott specifikációt, funkciólistát kellett összevetni azzal a valósággal és azokkal az igényekkel, amelyekkel a megkérdezett kutatók találkoznak a munkájuk során, azaz fel kellett mérni, hogy az általuk végrehajtani kívánt tevékenységek kivitelezhetőek-e vagy sem az adott szoftverkörnyezetben.

Az Adatrepozitórium Platform projekt interjúbeszélgetései alapján közel negyven egyedi felhasználói esetleírást határoztunk meg. Az interjúk természetesen nem alkalmasak az esetleírások teljes mértékű feltérképezésére (erre a felhasználók munka közbeni megfigyelése lehetne jó eszköz, de erre ennek a projektnek a keretein belül nem volt mód). A felhasználói történetek begyűjtése után az esetleírások meghatározásához az aktuálisan használt adatrepozitóriumi rendszerekben szokásos lépéssorokat vettük alapul. Az általunk így meghatározott esetleírások komplexitásukat, felbontásukat tekintve sem egységesek, de technikai értelemben is különféle kategóriákba esnek: már támogatott; fejlesztést igényel; a projekt keretein belül tudjuk implementálni.

A felhasználók aktuális repozitóriumhasználatán alapuló igényeken felül megfogalmaztunk még további tíz, jövőbe mutató esetleírást is. Ezekről, mivel a megkeresésekre és interjúkra a projekt kezdeti stádiumában került sor, az interjúk során nem tudtunk érdemben információt kérni a kutatóktól, de a FAIR-adatkezelés megvalósításához alapvető fontosságúnak tartjuk őket. Összességében az esetleírások elemi lebontásai már egy számítástechnikai értelemben vett követelményrendszert határoznak meg, és ez, kiegészítve a Dataverse szoftverből kinyert funkciólistával, megadja a fejlesztők számára szükséges specifikációt.

ÖSSZEFOGLALÁS

Összességében elmondható, hogy az adatmegosztás kultúrája igen kezdetleges Magyarországon. „Mindenki ül a dolgain", vagyis az adatain, ahogy az egyik válaszadónk fogalmazott. Vajon hol fogható meg a probléma a lehetőségek és kényszerek sűrűjében? Eleve gondok vannak az adatmegőrzéssel, ami technikai feltételek és követendő szabályok hiányában sok esetben nem csak a kutatók felelőssége. Erre jönnének még rá az adatmegosztással járó terhek. „[S]zép dolog az *open access,* de a valóság az, hogy nagyon sok korlátozást be kell tartani" – jelezte az egyik beszélgetőpartnerünk. És természetesen lényeges az is, hogy nincs meg a felülről érkező inspiráció, hiszen az állami hivatalok sem iparkodnak közzétenni a kutatók számára is értékes alapadatokat.

Mindennek tükrében érdemes újraértelmezni a többek által legfőbb problémának tekintett adatféltést. Nemegyszer azt tapasztaltuk, hogy a bizalmatlanság mögött komoly intézményi hiányosságok húzódnak meg: a kutatók magukra vannak utalva, és kénytelenek „hagyományos", azaz nem korszerű módon kezelni a saját adataikat, ami számos kényelmetlenséggel jár.

Emellett természetesen szükség van hozzáállásbeli változásra is, azonban nem egyéneket vagy akár egyes munkahelyeket nézve, hanem a tudományszervezés szintjén. Ennek lényege az volna, hogy valamilyen módon publikációkként ismerjék el, és értékeljék a kutatási adatok hozzáférhetővé tételét (Kratz–Strasser, 2014). Ez a motivációs bázis elengedhetetlen a nagy adatbázisok, repozitóriumok felépüléséhez. A tömeges adattárolásra alkalmas struktúrák elfogadtatásában fontos továbbá annak tudomásul vétele, hogy az ilyen rendszerek a megszokottnál nagyobb hibaaránnyal tudnak csak működni, ám ebbe bele kell törődni, ha az adatok hosszú távú megőrzése és elérhetővé tétele a cél.

Az is egyértelmű, hogy az egyes területek, intézmények különböző módokon és mértékben tudnak részt venni adatrepozitórium-építésben, illetve profitálni abból. A fejlesztéskor tehát különösen figyelembe kell venni az eltérő adottságokat és kompenzálni a hátrányokat. A lehetőségek egyenetlen megoszlása hosszú távon egy-egy tudományág háttérbe szorulását is eredményezheti, ami a teljes tu-

dományos élet kárára válhat. Ahogyan az egyik interjúalanyunk baljós előérzete diktálja: „[A]z összes tudományterület szegregálódni fog olyanokra, akik nagyon jól elő tudnak állítani adatokat, mert nagyon-nagyon jó kísérleteik vannak, nagyon-nagyon jó számítógépeik vannak, és olyanokra, akik pedig csak a mások által előállított adatokból fognak tudni dolgozni."

Mindenki számára nyilvánvaló, hogy az adat maga érték, és ezért is fontos a tulajdon- és értékesítési viszonyok, a hozzáférhetőség és a hivatkozás módjának szabályozottsága, valamint az adatgondozás humán kapacitásának biztosítása. Cikkünkben megmutattuk azt a kontextust, amelyben ez Magyarországon megvalósulhat, és rámutattunk arra, hogy a kutatásunk során kirajzolódó kutatói igények és felkészültségek ismerete hogyan használható fel egy korszerű kutatási adatrepozitórium megtervezése során.

IRODALOM

European Commission Directorate-General for Research and Innovation (2021): *The EC Kicked off a Study That Will Provide a Full Characterisation of the "European Research Data Landscape".* https://tinyurl.com/38vmsads

Gárdos J. (2011): Interjús szociológiai források újrafelhasználása. *Szociológiai Szemle,* 21, 3, 125–145. http://www.szociologia.hu/dynamic/06_gardos.pdf

Kratz, J. – Strasser, C. (2014): Data Publication Consensus and Controversies. *F1000Research,* 23 April 2014. 3:94. DOI: 10.12688/f1000research.3979.3, https://pdfs.semanticscholar.org/e74b/40a08ca3b83bce726a7573601186c8f064f6.pdf?_ga=2.134950001.1423098832.1664977457-30365543.1664977457

Móring T. (2022): *Oszd meg és uralkodj – Felmérés a kutatási adatkezelésről az ELTE kutatói körében*. Konferencia-előadás, Networkshop, 2022. 04. 21. Budapest: ELTE Egyetemi Könyvtár és Levéltár Oktatás- és Kutatástámogatási Osztály, https://networkshop.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=2b6e59e0-a81c-45a0-9d90-ae8c009755fc

Száldobágyi Á. (2022): *Esettanulmány egy intézményi adatrepozitóriumi rendszer és a hozzá tartozó szolgáltatás kialakításához*. Konferencia-előadás, Networkshop, 2022. 04. 21. Debrecen: Debreceni Egyetem Egyetemi és Nemzeti Könyvtár, https://networkshop.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=9172b05a-2c0d-42a5-ae93-ae8c0097352d

URL1: Springer Nature Kiadó – Figshare repozitórium: *A nyílt adatok helyzete* című kutatás adatai a Figshare repozitóriumban. https://doi.org/10.6084/m9.figshare.17081231

URL2: European Commission, Directorate-General for Research and Innovation: *European Research Data Landscape : Final Report*. Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2777/3648

URL3: *User Story.* Wikipedia, https://en.wikipedia.org/wiki/User_story

URL4: *The Dataverse Project.* https://dataverse.org

URL5: *CONCORDA*. https://science-data.hu/