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 repozitóriumi 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.