Magyar Tudomány 179(2018)5, 653–667
DOI: 10.1556/2065.179.2018.5.9

AZ ÚJ ADATKEZELÉS LEHETŐSÉGEI ÉS KOCKÁZATAI A TÁRSADALOMKUTATÁSBAN

POSSIBILITIES AND RISKS OF NEW DATA MANAGEMENT IN SOCIAL RESEARCH

Nagy Péter Tibor[1], Veroszta Zsuzsanna[2]

[1]egyetemi tanár, Eötvös Loránd Tudományegyetem Társadalomtudományi Kar, Wesley János Lelkészképző Főiskola
nagypetertibor@gmail.com

[2]tudományos főmunkatárs, Központi Statisztikai Hivatal Népességtudományi Kutatóintézet
veroszta@demografia.hu

ÖSSZEFOGLALÓ

A legtágabb perspektívában minden olyan „új adatról" gondolkodunk, amely a jelenben folyamatosan termelődik, és kezelhetővé tételük strukturálásra vagy újrastrukturálásra készteti a kutatót, átalakítva ezzel a kutatási eljárásmódokat is. Szűkülő fókuszunk ezután a nem kutatási céllal létrejött, ám strukturált mikroadatokra helyeződik át, melyeknek kutatási célú felhasználása egyfelől megköveteli a maga tudományos validációs eljárásait, másfelől azonban az adatfelhasználás új lehetőségei felé mozdítja el az empirikus vizsgálatokat. Legszűkebb perspektívában ezen lehetőségek megmutatása írásunk célja.

ABSTRACT

In the broadest perspective we are thinking of any kind of "new data" that is constantly produced in the present forcing the researcher to structure or re-structure them, to modify the research methods too. Our shrinking focus concentrates for micro data, which are produced definitely for non-research purposes. Using them for research purposes on the one hand, require their own scientific validation procedures, on the other hand open new perspectives. In the narrowest perspective the goal of our paper is to show these new opportunities.

Kulcsszavak: társadalomkutatás, Big Data, mikroadat, adatkapcsolás, longitudinális

Keywords: social science, Big Data, micro data, data-merging, longitudinal

Évszázadok óta termelődnek olyan adatok, melyeket a kutató kigyűjtött, de a számítástechnika előtti korban csak egyediségüktől megfosztva és csak aggregált formában kerülhettek tudományos felhasználásra. Az aggregáció a „változók"

kombinálhatóságának természetes határt szabott. Más esetekben az adatok „a változók kombinálhatóságát" megőrző helytörténeti elemzések formájában dolgoztathattak fel. Ez utóbbi viszont kizárta az általánosíthatóság reális igényét. Voltak ugyan olyan történészi programok, melyek szisztematikus helytörténeti elemzések utólagos összekapcsolhatóságát vették célba – az Ila Bálint kezdeményezte megyei monográfiasorozat, vagy az 1929-es Zsidó Lexikon hitközség címszósorozata példa erre –, de részben e szakirodalom korabeli szisztematikus feldolgozásának gyakori elmaradása, részben az utólagos – napjainkban folyó – adatbázisba szervezés nyomán egyértelművé váló adathiányok jelzik e törekvések tényleges korlátait. Ha pedig az adattömeg felhasználásának célja nem egyes jelenségek gyakoriságának megragadása volt – például jelentős életpályákra vonatkozó adatgyűjtések esetében –, a kutatóknak eleve le kellett mondaniuk arról, hogy konkrét tárgyukra vonatkozóan „minden" adatot összeszednek. Hiszen a „jelentős" életpályák, „jelentős" egyszeri történések épp azért minősíttettek önálló feldolgozásra méltónak, mert mind kortársaikra, mind az utókorra olyan hatást gyakoroltak, hogy távoli levéltárakban, félreeső sajtótermékekben is maradtak velük kapcsolatos adatok. Az újkori levéltári és könyvészeti termelés nagyságrendje pedig gyakorlatilag kizárta, hogy a kutató – egy konkrét személy vagy esemény nyomai után kutatva – *minden* olyan adatforrást átnézzen, ahol semmilyen logikus előfeltevés nem valószínűsítette vonatkozó adat felbukkanását. A személyekkel vagy egyedi történésekkel kapcsolatos különböző forrásokban megőrzött adatok feltárása és összekapcsolása csak a források tömeges digitalizálása nyomán vált lehetségessé, amit az olvasó akkor tapasztal meg leginkább, ha sok millió oldalas szövegtárakban – mint az Adtplus vagy az országos levéltár honlapja – indít el egy-egy névre vagy évszámra vonatkozó keresést. A kutatási célra felvett modern személysoros szociológiai adatbázisokhoz szokott, korrelációs vagy oksági hipotéziseket feltevő kutató számára a régi szövegek vagy táblázatok egyszerű szkennelésen alapuló digitalizált változata nem elégséges: ezek csak alapanyagul szolgálhatnak a kutatásra alkalmas adatbázisok felépítéséhez. E régi források vagy régi feldolgozások adatbázisba szervezése után azonban már az egyedi adatok tömeges összekapcsolása – avagy kis aggregátumok valószínűségi összekapcsolása – is lehetségessé válik. A régi adatok egy része a kortársak számára is adatként funkcionált (például adóösszeírások, anyakönyvek). Szociológiai vagy tudományos adattá válásuk az összekapcsolás, a tömegesedés, illetve – kortársak számára elvileg talán ismert, de gép nélkül elképesztő időigényű – matematikai elemzések révén történik meg. Vannak azonban olyan történeti források, melyeket a kortársak semmilyen értelemben *nem adatként* éltek meg: ilyenek a történeti térképek, ahol a települések, az úthálózat, a csatornahálózat adatait önmagában is elemzésnek vetheti alá a számítógéppel felszerelt kutató, de összekapcsolhatja adataikat szöveges forrásokkal, illetve napjaink térképadataival is. Ilyen a templomképek, családi fotók millióinak elemezhetősége is. De

nem élték meg „adatteremtésként" a szavak és kifejezések megválasztását az írók és újságírók, a magánleveleket írók, és az államigazgatási szövegeket, vállalati ügyiratokat termelők sem. Mindezeknek a forrásoknak a használata, „adattá alakítása", adatként elemzése a következő évtizedek történészei számára olyan lehetőségeket nyit a számítástechnika előtti évszázadok elemzéséhez is, amelyek folyamatosan újraalkotják a múltról szóló képünket.

A kutatás számára rendelkezésre álló „régi adatok" mellett napjaink jelenségeiről a számítástechnika tömeges alkalmazása következtében korábban elképzelhetetlen mennyiségű új adat keletkezik, melyek egyrészt befolyásolják a régebbi dolgok, párhuzamos trendek „kifutásáról" való tudásunkat – miáltal a napjainkról szóló adatok robbanásszerű bővülése a közelmúlt történetének folyamatos átírására késztet bennünket –, másrészt a „gyorsuló időben" maguk is hamar történeti forrássá válnak. Nyugodtan feltételezhetjük, hogy amiképpen az első adatrobbanás korának történészei és levéltárosai nem látták át mindazokat a módszereket, ahogyan koruk adattömegét – például a frissen megszületett napisajtót, az alfabetizmus általánossá válása következtében nagyságrendekkel megnövekedett magánlevelezést, a városias ügyintézés, üzemszerű termelés exponenciálisan megnőtt írásbeliségét – a jövő tudósai használni fogják (vagy használhatnák, ha megőrizték volna…), úgy valószínűleg ma sem látjuk pontosan, hogy a spontán keletkező digitális adattömeg hogyan hasznosul majd a jövőben. Mint ahogy azt sem, hogy az *evidence based* döntéshozás korában az egymással konkuráló álláspontok alátámasztására felépített strukturált, „tudományos" adattömegrész mennyire éli túl az alátámasztandó álláspont kisebbségben maradását.

E gondolatmeneten haladva tovább elméleti fejtegetésünk során a „jövő történészeire gondolva" az új adatok természetéről és kezeléséről két fő momentumot emelnénk ki. Egyfelől az új dolgokról szóló új, de strukturálatlan adatok keletkezésének társadalmi természetét, másfelől a strukturált, illetve „tudományos" adat születését és visszatöltődését az adatok körforgásába. Ezt követően a szakadatlanul keletkező „új adatok" újrastrukturálásának gyakorlati kutatási szempontjait, lehetőségeit és elemi szabályait gondoljuk át, a folyamatosan frissülő nyilvántartási – regiszter – adatok tudományos célú felhasználásának konkrét példáin.

ÚJ, STRUKTURÁLATLAN ADATOK ÚJ DOLGOKRÓL

A régi adatok digitalizálásának, történeti adatbázisba szervezésének valamilyen tudományos motivációja van. Az adatbővülési folyamatok teljesen más típusát jelenti az a napról napra az interneten megjelenő adattömeg, melyet termelője közigazgatási, politikai, üzleti, társadalmi vagy „társasági", esetleg magánéleti célból tesz nyilvánossá. Ez az adattömeg a közigazgatási, politikai, üzleti stb. valóságot soha nem látott mértékben teszi kutathatóvá. Az egyedi konkrét ese-

ményekre irányuló tudományos kutatás szempontjából – a gyakorlatilag korlátlan ingyenes tárhelykapacitások megnyílása óta – az adattermelődés és hozzáférés legfontosabb korábbi korlátja tovatűnt. Még két évtizeddel ezelőtt is biztosak lehettünk abban, hogy – az anyagi erő különbözőségénél fogva – nem mindenki publikálhatja adatait vagy mondanivalóját, akinek szándéka van azokat publikálni. A nem szöveges, hanem képi, különösen mozgóképi anyagok vonatkozásában ez még tíz éve is igaz volt.

Természetesen az egyedi történésekre vonatkozó kutatásnak továbbra is van négy alapvető korlátja.

Az egyik, hogy semmilyen mód nincs annak ellenőrzésére, hogy egy interneten publikált adat hogyan keletkezik „valójában", ki vagy kik készítették ténylegesen, az adatot készítők láthatták-e a kontextust, amelybe az általuk szolgáltatott adat vagy szövegrészlet illeszkedik, hogyan és miért módosult a szöveg, illetve az adat hogyan viszonyul a „valósághoz" (például dolgok ténylegesen a leírttól eltérő darabszámához vagy a tényleg elhangzott, a megjelentnél sokkal keményebb hozzászóláshoz, egy békésen üldögélő közönséget mutató kamera látószögéből gondosan kihagyott – bár az összes jelenlévő számára jól látható – kompromittáló szimbólumhoz stb.). Természetesen az internet előtti adatokkal kapcsolatban is igaz ez, csak az adatpublikáló felületek akkor még korlátozott száma valószínűsíthetővé tette, hogy a „valóságtól" történő eltérésre valaki azonnal felfigyel, és az eltérés lelepleződik.

A második, hogy az internetes adat használója minden korábbinál nagyobb mértékben ki van téve a hamisításnak. A hagyományos adatokat is lehetett hamisítani – azt a látszatot kelteni, hogy mástól származnak, mint akitől származnak –, de a csalás a Photoshop előtti korszakban jelentős munkabefektetést igényelt, s szakértők általában ki tudták zárni a hamis iratokat, hamis fotókat.

A harmadik, hogy semmilyen mód nincs annak ellenőrzésére, hogy maga az adattermelő nem termelt e belső használatra a nyilvánosan megjelent adattal teljes mértékben szembenálló adatot, akár tömegesen is. Természetesen az internet előtti korszakban is tömeges volt a kettős valóság – azaz voltak „belső jelentések" –, de sosem volt olyan könnyű és olcsó hamis vagy érdektelen adattömegbe „fullasztani" az érdeklődőt.

A negyedik ellentmondás pedig, hogy semmilyen mód nincs annak ellenőrzésére, hogy valamely kérdésben az adott ügyben érdekelt és aktív, de a nyilvánosságban nem érdekelt szereplő milyen adatokat termelt. Természetesen ez is jelen volt korábban, de minthogy az aktív irattárból a levéltárba kerülésre csak sok év után lehetett számítani, a kinyomtatásnak pedig költségei voltak, mindenki számíthatott rá, hogy egy döntésmechanizmusnak vannak a nyilvánosság számára láthatatlan anyagai is. Az e-kormányzás korában bármely döntéshez szükséges adat „láthatatlansága" csak abból eredhet, hogy az adat tulajdonosa *nem akarja* azt megosztani a nyilvánossággal.

Egészen más viszonyok jellemzik a kutatott tömeges jelenségeket. A találat-százezrek összehasonlítása olyan problémákat vet fel, amelyek a hagyományos adatok esetében nem merülnek fel. Ilyen például az adattöbbszöröződés, illetve a keresőmotorok sajátosságai, melyek – egyértelműen üzleti érdekek mentén – „fel-" és „lesúlyoznak" jelenségeket. A hagyományos adattömeg esetében az adatokat létrehozó, tároló, rendelkezésre bocsátó rendszer (például az államigazgatás vagy egy konkrét sajtóvállalat) érdekei áttekinthetők. Az internetes adattömeg esetében a keresőmotorokat vezérlő üzleti és reklámérdekek áttekinthetetlenek.

Az internetes adattömeg „nagy testvére" a „Big Data", melynek nagy része közvetlenül nem érhető el az interneten, amely a cégek, az intelligens hálózatok, a magánszektor és az egyéni felhasználók által világszerte és napi szinten előállított óriási adatmennyiséget jelenti. Ez közismerten folyamatosan növekszik, nagyságrendjét ma már csillagászati számokkal szokták meghatározni. Ezzel az adattömeggel persze inkább az informatikusok, mint a társadalomtudósok foglalkoznak. Az adattermelődés korábbi állapotával szemben az alapvető különbség, hogy az egyének mindennapi cselekvéseit korábban részben lehetetlen volt tömegesen adattá szervezve regisztrálni, részben csak tudatos előzetes döntések alapján lehetett – jelentős anyagi ráfordításokkal – regisztrálni és akár üzleti, hatósági, titkosszolgálati vagy kutatási célból felhasználni. A Big Data az az adattömeg, amelyet a minket körülvevő digitális eszközök – elvileg előzetes döntés nélkül szinte mindenről szinte mindent – regisztrálnak. Azaz az adattermelésünk legnagyobb része ma már nem tudatos jellegű.

Mindenképpen igaz azonban, hogy azt, hogy az egy adott másodpercben technikai értelemben mindenképpen létező adattömegből a következő másodpercre vagy a következő napra mi őrződik meg, az adatok birtokosainak érdekei határozzák meg. De a kérdés nem csak a megőrződés. Üzleti vagy kormányzati érdekek és igen komoly erőforrások kellenek ahhoz, hogy a valamiképpen megőrződő strukturálatlan adattömegből társadalomtudósok által már használható adatok váljanak.[1]

„TUDOMÁNYOS", STRUKTURÁLT ADATOK

Akár az internet, akár a Big Data adattömegéhez képest elhanyagolható nagyságrendű, de a korábbi korok hasonló adattermelésének többszörösét jelenti az a tudatosan strukturált, interneten vagy kiválasztott kör számára zártkörűen hoz-

[1] Úgy tűnik, érdekek a valóság egy lehatárolt darabjának megismerésére vannak csupán: tudomásom szerint még nem történt olyan valódi társadalomtudományi elemzés, mely – valamiféle Új-Ulyssesként – akár valamely társadalmi csoport által (illetve ról) meghatározott időszakban termelt valamennyi adatot feldolgozta volna. Elképzelhető, hogy a New York-i 9/11 után nagyobb titkosszolgálatoknál történtek ilyen feldolgozások, de ezeket a hétköznapi társadalomtudományi gyakorlat még nem ismeri.

záférhetővé tett táblázat-, számítás- és grafikontömeg, amely valamiféle szakértői munka, háttéranyag, tudományos munka.

Az *evidence based decision* gyakorlatilag az egész fejlett világban elterjedt. A már strukturált és emberi fogyasztásra emberi beavatkozással (ha mással nem, egy táblázattermelő *syntax* megírásával és lefuttatásával) előkészített adattáblák és grafikonok „csapnak össze" az ellenérdekelt felek vitáiban, mely ellenérdekelt felek természetesen saját, általuk vagy általuk is ellenőrizhető adatgyűjtő, -feldolgozó és adatértékelő szervezetek fenntartásában érdekeltek. Mi több, „az ellenérdekelt felek" nélkül működő politikai rendszerek is – tervezésre, de ha arra nem, akkor propagandára – tömegesen érdekeltek strukturált adatok létrehozásában.

Az adatgyűjtő, adatfeldolgozó és adatértékelő szervezetek pedig saját erő- és érdekviszonyaikkal alapvetően meghatározzák a tudományos célokból működő társadalomtudós helyzetét is. A strukturált adat, amellyel a társadalomtudós dolgozni kénytelen, magán viseli születésének körülményeit és az előállítók (esetleg szintén valamiféle társadalomtudósok) lét- és tudati viszonyait is. A tanulmányokba „beledolgozott" adatra – tekintettel arra, hogy válogatás és interpretáció eredménye – fokozottan igaz ez.

Nyilvánvaló, hogy számos társadalomtudományi mű esetében elmosódik a határ a politikai, kormányzati, üzleti erőviszonyokat közvetlenül figyelembe vevő alkalmazott kutatási publikációk, jelentések és a „tiszta tudomány" céljait szolgáló publikációk között. A publikációk termelői – finanszírozási vagy más okokból – gyakran érdekeltek abban, hogy konkrét érdekek szolgálatában álló adatközléseiket és adatelemzéseiket „tisztán tudományosként" tüntessék fel, vagy hogy valójában „tisztán tudományos" – azaz kizárólag szerzőjének a tudományos közösségben való előrejutását célzó vagy önkifejezési vágyát kielégítő, illetve a tudományos igazságkeresés transzcendentális igényét kielégítő – munkát alkalmazójuk vagy megrendelőjük érdekeit szolgáló praktikus műnek tüntessenek fel.

Természetesen a „tiszta tudomány" kritériumainak megfelelő és önmagát oda is soroló munkák is objektíve illeszkednek egyfelől a témájával kapcsolatos vagy akár asszociatív kapcsolatba hozható ideológiai és társadalmi viszonyokba, másrészről az adott tudományág mikrotársadalmi viszonyaiba, harmadrészt abba az általános versengésbe, ami a dolgok megközelítésének általános legitimációja körül folyik a vallás, a tudomány, a művészet, azaz a világ megismerésének fő történelmi formái, s az ezzel foglalkozó értelmiségi csoportok között.

Az „adatokra épülő" tudományos mű maga is adattá, adathalmazzá válik háromféle értelemben is: egyrészt a benne lévő adatok súlya, hatása, fennmaradási esélye a tanulmányba kerülés révén nagyságrendileg megnő a többi – tanulmányba nem kerülő, ott nem idézett, nem interpretált – azonos forrású adathoz képest. Másrészt az adatok új értelmet nyernek más adatforrásokból kiválogatott adatokkal való összekapcsolás révén. Harmadrészt pedig, az is adat, hogy kik,

miről, milyen terjedelemben, hol publikálnak, kikre hivatkoznak, milyen érvelési algoritmusokat használnak.

A tanulmány – immár megrendelőjétől, szerzőjétől, kiadójától függetlenül – azután újabb adatképződés forrásául szolgál, hiszen sokféle érdek által meghatározottan részévé válik irodalomjegyzékeknek, hivatkozásoknak, vagy fejt ki – hivatkozatlan formában is – alapvető hatást más tudományos tanulmányokra. Részévé válik a tudományos nyilvánosság ma még kétarcú intézményrendszerének: a papíralapú könyvkiadás világának, s a részben papír, részben azonban már egyre inkább internetes módon is hozzáférhető folyóiratok világának. A könyvkiadás és folyóirat-kiadás világa egyaránt különböző rangú elemekre oszlik. A „jó helyen megjelent" nyomtatott társadalomtudományi könyv persze magasabb presztízst jelent, mint a csak interneten elérhető szövegek. Könyvkritikák születnek róla, bevételszerzővé válhat, tankönyvvé válhat. A mű elterjedése szempontjából egyértelműen az interneten is elérhető folyóiratokban megjelenő művek élveznek előnyt. Pontosabban: élveznének, ha nem alakult volna ki a – filmek világához hasonlóan – a szűkebb közösségekben megosztott, illetve torrentoldalakon működő másodlagos piac, melyek a könyvek elektronikus elterjedésének is kedveznek, a kiadók kétségbeesett ellenállása ellenére. Az új adatok keletkezése, régi adatok újrahasznosulása még tovább bonyolítják azokat az erőviszonyokat, amelyek az egyes témákat kutató történészek, a lezárult dolgokat kutató nem történész identitású társadalom- és bölcsészettudományi kutatók, a napjainkban is élő dolgok előzményeit és gyökereit bemutatni kívánó kutatók, a múlt társadalmi jelenségeit analógiaként és példatárként használó közvéleményformáló értelmiségiek között fennállnak. Ezek az erőviszonyok ugyanis természetesen kihatnak a forrásfeltárási, forráspublikálási folyamatokra, a tudományos művek megszületésére és elterjedésére – végső soron a tudomány egészére.

NEM TUDOMÁNYOS, STRUKTURÁLT ADATOK

A tudományos céllal rendszerezett adattömeg kezelési sajátosságai mellett a nem tudományos – például adminisztrációs, regisztrációs – céllal gyűjtött, ám strukturált adatok kutatási szempontú kezelése is új lehetőségek előtt áll. Első ránézésre a regiszteradatok rendelkezésre állása nem tűnik nagy újdonságnak a kutatók számára, hiszen a statisztikai rendszerek hosszú történetük során mindig is produkálták ezt az információforrást, és a tudomány élt is a lehetőséggel. (A magyar statisztikát és államigazgatást e szempontból közepes mértékű előrelátás jellemzi. 1880 óta megőrizték ugyan a népszámlálási kiadványok alapjául szolgáló elsődleges aggregációkat – ez a számítástechnika korában kis aggregációkat esetként megjelenítő, s összekapcsolást lehetővé tevő adatbázisok felépítését teszi lehetővé –, de a népszámlálás személysoros adatait még az 1980-as és 1990-es

népszámlálás után sem őrizték meg, noha akkor már a számítástechnika lehetőségei mindenki számára ismertek voltak. Az intézmények, illetve a levéltárak az 1850–1950 közötti évszázadban megőrizték a középiskolai és egyetemi anyakönyveket, de azt a vélhetőleg szintén eredetileg személysoros adatlapot, melyek alapján országos összesítések készülhettek, például a tanulók anyanyelvéről, már nem lehet fellelni.) A lényeges fejleményt ezen adatbázisok egyénsoros összekapcsolási lehetőségeinek felfutása jelenti, amit mind az informatikai lehetőségek, mind a jogi szabályozás alakulása nagyban elő tud segíteni. A technológiai környezet esetében ez olyannyira érvényes, hogy egyenesen a Big Data kutatási felhasználásának alapproblémájaként szembesülünk azzal, miszerint a tudományos célú felhasználás gyakorlati és szemléleti szempontból egyaránt nehezen tud lépést tartani az informatikai lehetőségek gyors fejlődésével.

Mielőtt azonban a kutatói szemléletmód átalakulásán és állandóságán, majd a megújuló kutatási eljárásokon gondolkodnánk az új közegben, röviden a regiszteradatok természetével kell foglalkoznunk.

Kutatóként az adatok sokféle típusával lehet dolgunk. Az adatok sokfélesége azt is jelenti: kezelésükkel, értelmezésükkel kapcsolatban sokkal többféle kérdés merül fel, mint amikor a kutató egyféle (például egy konkrét népszámlálásból, egy konkrét adóbevallásból vagy egy konkrét közvéleménykutatásból származó) adatot próbál meg értelmezni. Az adat magában „az információ formalizált módon való megjelenítése, amely alkalmas feldolgozásra, továbbításra, közlésre, értelmezésre" (KSH, 2014). Az egyedi adatok közvetlenül hozhatók kapcsolatba egy egyénnel vagy szervezettel.[2] Ezek (például a név, a lakcím, a születési dátum, de mesterséges azonosítóként akár az adószám, TAJ-szám is) személyes adatnak vagy üzleti titoknak tekinthetők mindaddig, amíg az adatkezelés során kapcsolatuk az egyénnel vagy egyedi szervezettel fennmarad vagy helyreállítható. Még mindig egyéni szintű, ám anonimizált az a mikroadat, amely egy adott alanyra vonatkozik ugyan, de a közvetlen és közvetett azonosíthatóság lehetősége nélkül, azaz személyes jellegétől megfosztva áll rendelkezésre. A döntő különbséget ez az egyéni szint jelenti a statisztikai adatokhoz képest, amelyek az egyedek megfigyeléséből, statisztikai műveletek eredményeként jönnek létre. A teljes körű adatbázisok azután aggregáltsági szintjük mellett céljaikat tekintve is különböznek. Adminisztratív adatforrások esetében a szervezet nyilvántartási, engedélyezési, jogosultsági (köz)feladatainak ellátása az adatgyűjtés célja, ami egyben a célcsoport teljes lefedettségének és egyedi azonosíthatóságának követelményét is magába foglalja. Abból eredően, hogy az állami rendszerek (adózás, társadalombiztosítás, oktatás) az adminisztratív adatokat igazgatási, szabályozási, regisztrációs, szolgáltatási stb. – de nem kutatási – céllal gyűjtik, a kutatásban történő hasznosításuk csakis másodlagos felhasználásként értelmezhető. (A kutatónak gyakran

[2] A statisztikáról szóló 1993. évi XLVI. törvény alapján.

vállalnia kell ennek azt a következményét is, hogy az eredeti adatfelvétel irányítói szempontjai szerint a leghatalmasabb adatbázisokból éppen a legfontosabb háttéradatok (például szülők társadalmi státusza, felekezet vagy politikai orientáció) hiányoznak. Ehhez a hasznosításhoz egyébként, épp a célzott adatfelvételekhez képest remélhető gazdaságossága miatt állami és nemzetközi szakpolitikai szinteken nagy várakozások és sikeres gyakorlatok kötődnek.[3]

Látszólag valóban nem jelent komoly pluszterhet az állami adatvagyon részét képező, már meglévő, akár összekapcsolt személysoros adatállományok kutatási felhasználása. Valójában azonban (a jogi és etikai megfontolásokat most nem említve) önálló kutatási tevékenységről van szó: az adatok olyan újrastrukturálásáról, melynek során az adminisztráció célhoz kötöttségét kell a kutatási célhoz transzformálnunk. Az adminisztratív adatok struktúrájukban – és nem csak aggregátumként – magukon hordozzák, leképezik ugyanis az aktuális hatalmi, döntéshozási struktúrákat. A kutatás során csak az adminisztratívból kutatásivá transzformált elemi adatokon válik majd lehetővé új struktúrák keresése. Ezen kezelési és újraértelmezési igény a kutatói munkát állítja új feladatok elé mind a tudás és szemléletmód, mind a készségek, mind pedig az eszközök szintjén, bázist adva ezzel olyan új szakmák, mint az adattudomány/adattudós megerősödéséhez.

Az adattudomány részeként a nem tudományos céllal gyűjtött és strukturált regiszteradatok kutatási becsatornázásához – a törvényi háttér megismerése, az adatszolgáltatással kapcsolatos érdekek felmérése és a rendelkezésre álló adatok körének feltérképezése mellett – a tudományos kutatás lépéseit kell hozzárendelnünk. Ehhez az adminisztratív adattípus megkülönböztető jegyeinek megismerése szükségeltetik. A témában született gazdag irodalom (lásd például: Dixon, 2000; Elias, 2015; Hotz et al., 2008; McNabb et al., 2009; Roos et al., 2008; Smith et al., 2004) jól mutatja, hogy a kutatási céllal gyűjtött adatokhoz képest a regiszteradatok nagy erőssége az adatok mennyisége és a célcsoport teljes lefedettsége, ami a válaszadási torzítás gondját csökkenti. (Persze a regiszteradatok szisztematikus hiányokat hoznak: mondjuk a hajléktalanokkal vagy éppen a nem legális jövedelmekkel kapcsolatban.) Az adminisztratív adatbázisok kiterjedése időben is nagyobb mozgásteret nyújt, mint a célzott adatfelvételek, ahol a múltbéli információk egy részéhez csak retrospektíve, a válaszadó memóriájára, a válaszadó pszichológiai törvényszerűségek szerint torzuló narratívájára bízva magunkat férhetünk hozzá, a jövőt pedig vagy tervekből, vagy későbbi adatfelvételek lebonyolítása árán követhetjük. A regiszterek számára – állandó vagy konvertálható adatstruktúrát feltételezve – nem kihívás a longitudinalitás biztosítása különösebb anyagi és időráfordítás nélkül. (Természetesen a szavak – például foglalkozásnevek – jelentésének változása mint probléma, így is fennmarad.) Igaz, e kétségtelen rugalmasság és bőség ellenére nagyon is erős rigiditásba ütközünk

[3] Lásd például a 2007-es OECD *Istanbul Declaration*t.

a kutatási felhasználás során. Abba az áthidalhatatlan távolságba például, ami a kutatni kívánt jelenség változószintű megragadásának adminisztratív és primer kutatási lehetőségei között feszül. A regiszteralapú kutatásnak olyan változókészlettel és kategóriarendszerrel kell boldogulnia, amelyet nem kutatási céllal hoztak létre. Ahhoz, hogy ezekből kutatási adatok és változók váljanak, végül is más sorrendben, de a klasszikus kutatási eljárás- és szemléletmód szerint kell eljárnunk, vagyis definiálnunk kell a változókat és meghatározni mérési módjukat. A kutatási célú újradefiniálás és validálás során az adatok létrejöttének kontextusát és vonatkozásának korlátait épp azzal a fegyelemmel építjük újjá, mint ahogyan egy primer kutatás tervezése során eljárunk – jóllehet az eltérő sorrendhez hozzá kell szoktatni gondolkodásunkat, és át kell hangolni eljárási szabályainkat.

Persze, még ha el is végezzük az adatok átstrukturálását – újrakonceptualizálunk és -operacionalizálunk változókat kutatási céljainknak megfelelően, adminisztratív adatból kutatási adatot hozva ezzel létre –, akkor is pótolhatatlan hiányokba fogunk ütközni a lágy társadalomtudományi változókat és általában minden, adminisztrációs, ellenőrzési célt nem szolgáló adatot illetően. Szembe kell néznünk azzal is, hogy a hamis adatszolgáltatáshoz sokkal többféle és sokkal szisztematikusabb egyéni és intézményi érdekek kötődnek, mint a kutatási adatfelvételek esetén. Magának az adatfelvevőnek is sokkal inkább vannak – rejtett – igazgatási vagy politikai érdekei, amelyek éppúgy befolyásolhatják a kérdések megfogalmazását, mint az adatfelvétel módját. Ezeket az ellentmondásokat hidalhatja át – némiképp – a kutatási és adminisztratív adatbázisok összekapcsolása, amellyel a két típusú adatforrásban rejlő pozitívumokat ötvözhetjük.

ADATOK ÚJRASTRUKTURÁLÁSA MINT KUTATÁS

Az egyéni szintű adatokat tartalmazó – akár kutatási, akár adminisztratív – adatbázisok összekapcsolásával az adatok kutatási célú újrastrukturálásának új lehetőségeihez érkeztünk el. (A különféle helyeken felvett adatok összekapcsolása kutatási szempontból végzett adatfelvételeknél sem túl gyakori. Nagyobb, tehát államigazgatásival összevethető nagyságrendű esetszámú adat-összekapcsolásra két példát tudnánk említeni: az egyik a Csákó Mihály vezetésével folyt adatfelvétel, amely 1998 tavaszán valamennyi 12. évfolyamos középiskolással kérdőívet vett fel, majd – később eltüntetett – név és születési dátum alapján összekapcsolta ezt az adott év őszén a felsőoktatásba beiratkozott diákok adataival.[4] A másik a Karády Viktor és Nagy Péter Tibor vezetésével felvett adatbázis, amely az 1870

[4] *A felsőfokú továbbtanulás tényezői* című kutatás. Szervezeti háttér: Eötvös Loránd Tudományegyetem Szociológiai és Szociálpolitikai Intézet; Finanszírozás: Soros Alapítvány; A vizsgálat éve: 1998; Kutatásvezető: Csákó Mihály.

és 1918 között érettségizett diákok középiskolai értesítőkből, illetve anyakönyvekből felvett adatsorát kapcsolta össze – név és születési év alapján – az egyetemi anyakönyvekkel és diplomakönyvekkel.[5]) Az állami adminisztrációban több olyan egyénsoros adatbázis áll rendelkezésünkre, amely egyedi azonosítói vagy egyéni adatkombinációi révén lehetővé teszi az információtartalom integrálását. Ezen lehetőségek teljes köre aligha térképezhető fel, és talán hamar aktualitását is vesztené, éppen ezért a következőkben már megvalósított vagy tervezett konkrét adatkapcsolási eljárások bemutatásához fordulunk. Két hazai példát veszünk alapul: az egyik az Oktatási Hivatal által a Diplomás Pályakövetési Rendszer keretein belül végrehajtott államigazgatási adatbázisok pályakövetési célú integrációja. Ebben az esetben a felsőoktatásban abszolutóriumot szerzettek évfolyamának alapsokaságához kötötték hozzá az adózási, társadalombiztosítási, foglalkoztatási rendszerekben található adatokat a végzést követő időszakra nézve, egyénsoros anonimizált módon.[6] Az adatok átstrukturálása itt az oktatási adminisztrációban összegyűjtött információhalmaz munkaerőpiaci szempontú rendezését jelenti annak érdekében, hogy a friss diplomások kilépési sikerességét az elérhető paraméterek mentén – foglalkoztatottság, jövedelem, foglalkozás – becsüljék. Ebben a példában adminisztratív adatbázisok közötti összekapcsolásról van szó, amely egy előzetesen definiált alapsokaság teljes körére kiterjedően integrálja a lehetséges információkat. A regiszteradatok kutatási célú transzformálása ez esetben is szükségeltetik, a validálási eljárást minden bevont változóra szükséges kiterjeszteni a kutatási nézőpontnak megfelelően. Ezt követően az egyénsoros összekapcsolás abban az esetben, ha minden adatbázis tartalmazza ugyanazt az egyedi azonosítót (ez esetben a TAJ-szám tölti be ezt a szerepet), *match-merge* eljárással mechanikusan végrehajtható. Ennek hiányában az összekapcsolás – *deterministic linkage* – a személyes adatok egyedi megkülönböztetést lehetővé tévő kombinációjára támaszkodhat (például a név, születési hely és idő együttesen már biztosíthatja az azonosítást). Az adatkapcsoláson alapuló kutatási adatbázisból aztán vizsgálhatóvá válnak például az egyes szakokon végzettek NAV-nál rögzített havi bruttó keresetei néhány évvel a végzettség megszerzése után.

Adminisztratív mikroadatok kutatási átstrukturálásának longitudinalitáson alapuló példáit a Központi Statisztikai Hivatal Népességtudományi Kutatóinté-

[5] URL1, Finanszírozás: FP7-es projekt.

[6] Az adatkapcsolásra több alkalommal is sor került (2010, 2012, 2013, 2016). A 2013-as vizsgálatba bevont szervezetek például: Felsőoktatási Információs Rendszer (FIR), Diákhitel Központ Zrt., Magyar Államkincstár (MÁK), Nemzeti Adó- és Vámhivatal (NAV), Országos Egészségbiztosítási Pénztár (OEP), Országos Nyugdíjbiztosító Főigazgatóság (ONYF), Nemzeti Munkaügyi Hivatal (NMH). A vizsgálat alapsokasága a 2009/2010-ben, illetve 2011/12-ben a felsőoktatásban végzettek teljes köre, az adatgyűjtés a 2013-as státuszra, illetve visszamenőlegesen három évre vonatkozott. Az adatkapcsolás-sorozat napjainkban is zajlik. Ismertetését lásd például: Nyüsti–Veroszta, 2014.

zetben most induló Magyar Születési Kohorszvizsgálat – Kohorsz ’18 – tervezett adatkapcsolásainak bemutatásával érzékeltetjük.[7] A vizsgálat a 2018-ban születendő gyermekek tízszázalékos mintáján méri fel a gyermekek magyarországi felnövekedésének számos aspektusát. A longitudinális *survey* adatgyűjtés a várandósság időszakában indulva fél-, egy-, négyéves korig, sőt a tervek szerint még tovább követi a gyermekek kognitív, érzelmi fejlődését, családi és környezeti hátterét, egészségét, társadalmi helyzetét. Az azonos alapsokaság időben folyamatos követése a survey módszertanon belül is folyamatos egyéni szintű adatkapcsolást feltételez. Ötletszinten emellett egy születési kohorszvizsgálatot számos adminisztratív adatkapcsolási lehetőség segíthet. Az elvi lehetőség adott például az ugyanazon adminisztratív adatbázison történő adatkapcsolásra. Ennek során egyazon regiszter különböző időpontokra vagy tartalmakra vonatkozó adatait kapcsoljuk össze a kutatás alappopulációjának egyedi adatsoraihoz. Ez esetben az adatok egyéni szintű azonosítása a rendszer adottsága. Ilyen kutatási elem lehet például a 2018-ban gyermeket vállaló nők szülés előtti és utáni munkaerőpiaci életútjának vizsgálata a változó nevű és jogállású egészségbiztosítás adatain. De ilyen longitudinális adatkapcsolásra kerülhet sor abban az esetben is, ha a védőnői rendszerből vett szülői adatokhoz ciklikusan kapcsoljuk a gyermek ugyanabban a rendszerben rögzített fejlődési adatait.

Amennyiben a kutatás mintájához, illetve az ez alapján rögzített survey adatokhoz kapcsolunk regiszteradatokat, azonosítási szempontból két úton haladhatunk. Az adatbázistól elkülönülten kezelve (a válaszadó beleegyezésével) rögzíthetjük azt a kapcsolati kódot (például TAJ-szám), amellyel a válaszadóra vagy a mintába került gyermekre vonatkozó adattartalmak az adminisztratív adatbázisokban azonosíthatók. Ez esetben akár regiszteralapú surveyre is lehetőség nyílik, amennyiben a teljes körű adminisztratív adatbázis megfigyelési egységre vonatkozó egyedi adatsorai egészülnek ki survey kutatásból származó, a mintára vonatkozó szintén egyedi adatsorokkal. A két típusú (hivatalos és személyes) információforrás összekapcsolása nemcsak a két adattípus erényeit kombinálja, hanem jelentősen csökkenti a válaszadói terheket (ezáltal a költségeket) is. A kohorszkutatás esetében regiszteralapú kutatási designt jelentene, ha a mintába került várandósok terhesgondozási nyilvántartásból származó regiszteradatai már az első személyes megkereséskor egy adatsorrá kapcsolódnának össze a felvett adatokkal. Egyedi azonosító esetében persze minderre utólag is sor kerülhet. Ha azonban ennek alkalmazására nem nyílik lehetőség, akkor még mindig lehetséges a már meglévő, különböző forrású, ám azonos alappopulációt lefedő adminisztratív, illetve survey adatbázisok egyedi szintű összekötése. Ennek során – a deter-

[7] A kutatás az EFOP 1.9.4. – VEKOP-16 EMMI-felhívás: *A szociális ágazat módszertani és informatikai megújítása* keretében valósul meg. A kutatási program az induló szakaszban jár, ismertetését lásd az URL2 honlapon.

ministic linkage eljáráshoz hasonlóan – a surveyben szereplő egyes válaszadók adatkombinációiból állítunk össze egyedi azonosításra alkalmas adatcsomagokat, és keressük ezek megfelelését az adminisztratív adatbázisban szereplő, elvileg teljes alappopuláció egy tagjával. Az ehhez alkalmazott módszer a valószínűségi adatkapcsolás (probabilistic record linkage)[8] amely statisztikai eljárással azonosítja a két, azonos alappopulációt lefedő adatbázis tagjai közti kapcsolat statisztikai valószínűségét.

A születési kohorszvizsgálatban emellett az időbeliség, életút kezelésén leginkább alapuló adatkapcsolási kutatási elem lehetne egy olyan „adminisztratív jelzőrendszer" kidolgozása is, amely az időben rögzített kutatási szakaszok mellett a megfigyeltek (gyermek, illetve anya) megjelenését figyeli a szociális és oktatási rendszer, valamint a munkaerőpiaci adminisztráció adatbázisaiban, és erre reagálva eseti adatfelvételt tesz lehetővé. Ezáltal a fejlődési szakaszokhoz igazított kutatási szakaszok mellett az adatgyűjtés egyedi életeseményekhez is igazodhat. Amennyiben például az adminisztratív rendszerek a kutatás számára visszacsatolást küldenének, ha a mintában szereplő anya foglalkoztatottként jelenik meg, vagy a gyermek közoktatásba kerül, lehetővé válna a gyermekek vizsgálata egységesen a bölcsődei ellátás kezdete utáni hónapokban vagy az óvoda megkezdésekor, vagy az anyák felkeresése munkába állásuk után néhány hónappal, vagy újabb gyermekvállalásuk, sőt esetleges munkanélkülivé válásuk esetén.

Előbbi példáink annak érzékeltetésére szolgálnak, ahogyan a nem kutatási céllal létrejött, hatalmas és folyamatosan bővülő adattömegeket – „új adatokat" – a kutatás a maga számára hasznosíthatja. Az, hogy példáink egy része jelenleg zajló eljárás, másik része még csak tervezési szinten érvényes, jól érzékelteti ezen új adatok keletkezésének gyorsaságát és a kutatási reagálás előtt álló kihívásokat.

Elméleti szempontból a legnagyobb kihívás persze annak feldolgozása, hogy az *adatfeltárás, adatrendezés és adatelemzés* jelenségvilága egy komplex tudományszociológiai összefüggésrendszer része.

Funkcionalista megközelítésből ez azt jelenti, hogy minden társadalomtudományi elemzés legnagyobb kihívása, hogy *egyszerre* kellene…

…elemeznünk – önmagukban – a rendelkezésre álló adatokat;

…megállapítanunk, hogy a rendelkezésre álló adatok mennyire tükrözik annak a társadalomrésznek a viszonyait, amelyről szólni akarunk (van-e szisztematikus torzítás);

…megállapítanunk (például analógiák vagy hasonló kutatások alapján), hogy a rendelkezésre álló adatok, változók közül nem hiányzik e valamilyen alapvető adat, melynek hiánya az összes magyarázatot irreálissá teszi;

[8] Az eljárás matematikai alapjainak kidolgozásának forrása: Fellegi–Sunter, 1969.

…megállapítanunk, hogy az adattermelődés közegének nyelvi valósága az adatok mértékének mekkora homogenitását biztosítja, azaz, hogy minden adatszolgáltató számára többé-kevésbé ugyanazt jelentik-e az adatszolgáltatás során használt szavak;
…megállapítanunk, hogy mennyiben módosítják adatainkat a másoktól átvett vagy magunktól kitalált elemzési kategóriák.

Konfliktuselméleti paradigmában az önreflexió és álcázás sajátos dinamikáját fenntartva ugyanakkor arról beszélünk, hogy…
…témaválasztásunk mennyire szabad és mennyiben megrendelt;
…adataink mennyire valósak és mennyire konstruáltak;
…kategóriáink mennyire rugalmasak és mennyire előre programozottak;
…hipotézisünk, illetve tanulságaink mennyire voltak meg előzetesen, vagy mennyire keletkeztek az elemzés eredményeképp;
…doxikusan használt okságfogalmunk valójában mely tudományfilozófiai rendszer részeként legitim, illetve illegitim.

Az államigazgatási alapú adattömeg használatával egyszerre növeljük – finanszírozási esetlegességektől függetlenítve – a társadalomtudományi kutatások lehetőségeit, s egyben nézünk szembe azzal a kísértéssel, hogy ne arra akarjunk válaszolni, ami tiszta tudományos kérdésként megfogalmazódik, hanem arra, amire a más célú és más motivációjú adatfelvételek választ kínálnak. Egyénenként és kutatóközösségként is vigyáznunk kell arra, hogy a lényegtelen kérdésekkel kapcsolatos lehetőségek ne szorítsák háttérbe a lényeges, de nehezebben „adatolható" kérdések megfogalmazását, felelős megvitatását.

IRODALOM

Dixon, S. (2000): Using Administrative Data Sources in Labour Market Research. *Labour Market Bulletin*, 2, Special Issue, 26–30.

Elias, P. (2015): *New Forms of Data – New Opportunities for Research.* Trans-Atlantic Platform Social Sciences and Humanities, 11th February 2015.

Fellegi, I. A. – Sunter, A. B. (1969): A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 328, 1183–1210. https://www.researchgate.net/publication/228057942_A_Theory_for_Record_Linkage

Hotz, V. J. – Goerge, J. – Balzekas, J. – Margolin, F. (eds.) (2008): *Administrative Data for Policy-relevant Research: Assessment of Current Utility and Recommendations for Development. A Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research.* http://public.econ.duke.edu/~vjh3/working_papers/adm_data.pdf

KSH – Központi Statisztikai Hivatal (2014): *Módszertani dokumentáció/Fogalmak, definíciók.*

McNabb, J. – Timmons, D. – Song, J. – Puckett, C. (2009): Uses of Administrative Data at the Social Security Administration. *Social Security Bulletin,* 69, 1, https://www.ssa.gov/policy/docs/ssb/v69n1/v69n1p75.html

Nyüsti Sz. – Veroszta Zs. (2014): *Diplomás pályakövetési adatok 2013 – Adminisztratív adatbázisok integrációja*. Budapest: Educatio Társadalmi Szolgáltató Nonprofit Kft. https://www.felvi.hu/felsooktatasimuhely/dpr/kiadvanyok/adminisztrativ_adatbazisok_integracioja2013

Roos, L. L. – Brownell, M. – Lix, L. et al. (2008): From Health Research to Social Research: Privacy, Methods, Approaches. *Social Science & Medicine*, 66, 1, 117–129. DOI: 10.1016/j.socscimed.2007.08.017

Smith, G. – Noble, M. – Anttilla, C. et al. (2004): *The Value of Linked Administrative Records for Longitudinal Analysis. Report to the ESRC National Longitudinal Strategy Committee.* http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=21358D6605BCE27CF3130F4C-539B121A?doi=10.1.1.630.1056&rep=rep1&type=pdf

URL1: Culturally Composite Elites, Regime Changes and Social Crises in Multi-Ethnic and Multi-Confessional Eastern Europe. (The Carpathian Basin and the Baltics in Comparison - cc. 1900–1950). http://elites08.uni.hu

URL2: www.kohorsz18.hu