
VALÓSÁG

A rovatunkban megszólaltatott szakértők az adminisztratív adatok felhasználásának két területét mutatják be. Kádárné Fülöp Judittal, az EMMI korábbi OECD szakértőjével az oktatás területén rendelkezésre álló adminisztratív adatállomány kezeléséről és államigazgatási felhasználási lehetőségeiről beszélgettünk. Rudas Tamás egyetemi tanárt (ELTE), a TÁRKI tudományos igazgatóját az adminisztratív adatok társadalomkutatási kezelésének gyakorlatáról kérdeztük.

„Informatikai fejlesztés és információ fejlesztés AVAGY eljutunk-e oda, hogy az adatbázisok mesélni tudnak az oktatási rendszerről?”

Interjú Kádárné Fülöp Judittal

Educatio: Miért fordult a figyelem az adminisztratív oktatási adatbázisok felé nemzetközi és hazai szinten?

Kádárné Fülöp Judit: Nyilvántartás értelemben nagyon régóta van oktatási adatbázis, hiszen a beiratkozási napló és a kartoték is az. A fejlett oktatási informatikát működtető rendszerekben az informatika fejlődésének megfelelő fázisában áttértek a kartoték-rendszerrel az elektronikus adminisztrációra. Amikor az adattovábbítási lehetőségek ezt engedték, értelemszerűen létrejöttek központi elektronikus nyilvántartások, amelyekből azután statisztikai célokra használható adatbázisokat is lehetett építeni – az eredeti nyilvántartási cél fenntartása mellett. Ma az Európai Unió tagállamainak kb. a fele ilyen nyilvántartásokból állítja elő az oktatástatisztikai adatok jelentős részét.

Az elektronikus adatgyűjtő rendszer nyilvánvaló előnyei azt a látszatot keltik, hogy ezekből automatikusan elő lehet állítani statisztikai célra is használható adatforrást, adatbázist. Pedig ez tévedés. Az igazgatási célú adatok egy része szöveges adat (pl. név, lakcím), amire igazgatási okokból szükség van, ám egy statisztikai elemzésre szánt adatbázisban kényelmetlenek, mert ezekben a táblázatba rendezhető adatokat lehet jól kezelni.

Az igazgatási adatok elsődleges gyűjtői maguk az intézmények és az intézményfenntartók. (Igazgatási ok például a beiratkozások, bizonyítványok nyilvántartása, az órafelosztás elkészítése, munkaszerződések készítése, a bérelszámolás, stb.). Ezzel szemben az

oktatáspolitikának többnyire összesített statisztikai adatokra és elemzésekre van szüksége. Amiért az oktatáspolitikai is igényli az egyedi adatokat (pl. hogy a tanuló melyik iskolába jár, mikor született, hányadikos, ismételt-e osztályt, stb.) az éppen az elemzési szükségletből fakad. Ugyanis elemi adatokból sokkal többféle összesítést és elemzést lehet készíteni, mint a már eleve összesített intézményi statisztikai adatokból, éppen ezért sokkal kevesebb adatból sokkal több információt lehet nyerni.

Magyarországon a 2000-es években kezdődött igazán nagyvonalú informatikai fejlesztés az oktatásügyben. A 2000-es évek elején az iskolaadminisztrációs programok elterjesztése is megindult, de azután ez hosszú időre leállt. Viszont modernizálták az intézményi alapú statisztikai adatszolgáltatást (az ún. „októberi statisztikát”), ami korábban – egészen 2000-ig – kizárólag papíron történt. Ez megteremtette az igényt az elektronikus adatszolgáltatásra. Az iskolaadminisztrációs programfejlesztések leállása miatt a statisztikai adatszolgáltatás fejlesztése egy másik irányt kezdett követni, t.i. a web-es statisztikai adatszolgáltatás vált uralkodóvá. Ennek két oka volt. Egyrészt igény volt arra, hogy a központi statisztikai adatbázis létrehozásához ne kelljen adatrögzítéshez folyamodni (ami komoly hibalehetőség), ezért mindenképpen arra törekedtek, hogy az intézmények számítógépen keresztül szolgáltatassanak adatot. Másrészt az akkori számítógép ellátottság mellett könnyebb volt megoldani egy web-es felületen történő adatszolgáltatást, mint elérni azt, hogy minden intézmény iskolaadminisztrációs programot használjon.

2008-ra teljeskörűvé vált az intézmények web alapú statisztikai adatszolgáltatása, és ezzel *párhuzamosan* elkezdődött a központi elektronikus nyilvántartások fejlesztése is bizonyos területeken. Az utóbbiak közé tartoznak a személyi nyilvántartások (az intézmény nyilvántartás, a tanulói és a pedagógus nyilvántartás, a középfokú beiskolázást szolgáló adatbázis, az érettségi adatbázis, a felvételi információs rendszer és a felsőoktatási információs rendszer is). Ezek a rendszerek – néhány kivétellel – a mai napig egymással párhuzamosan és egymásról kevéssé tudva működnek.

A megnövekedett figyelem az elektronikus adatbázisok iránt annak is köszönhető, hogy 1990 után alaposan átalakult az oktatási rendszer, ami a szabad iskolaválasztás és a felsőoktatási képzési programok sokasodása révén jóval komplikáltabb lett. 1990-től kezdve az intézmények a fogyó gyermeklétszám miatt kénytelenek voltak különböző speciális, egyedi ajánlattal, tantervvel előállni annak érdekében, hogy minél több diákot vonzzanak. Ugyanez valósult meg a felsőoktatásban is, ahol viszont a csökkenő demográfiai tendencia a bejutási küszöböt csökkentette, ezzel egyidejűleg viszont olyan programok indítására ösztönözte a felsőoktatási intézményeket, amelyek ezeket az újabb hallgatói rétegeket is vonzotta. Egy ilyen sokszínű rendszert persze statisztikailag sokkal nehezebb követni, mint egy olyan rendszert, amely egységes központi tanterveken, szakokon és követelményeken alapszik. Eltekintve attól, hogy sokkal aprólékosabb bontásokra van szükség, ha a változásokat követni akarjuk, egyéb problémák is adódnak, ami a statisztikusoknak fejtörést okoz. Például többek között az, hogy ugyanaz a diák (pl. szakképzésben, felsőoktatásban) párhuzamosan több programban is részt vehet, akár több intézményben is. Így aztán sokkal kifinomultabb, személyes adatokat tartalmazó ún. mikroadatbázisokra van szükség az olyan egyszerű kérdések megválaszolásához is, hogy „hány szereplő van a rendszerben”.

E: Mindez milyen fejlesztéseket és új irányokat hozott előtérbe?

KFJ: A 2000-es évek elején megérett az idő az oktatási azonosító bevezetésére. Ez egy olyan egyedi azonosító szám, amit minden gyermek (most már az óvodába lépéskor) megkap, és ez az azonosító elkíséri egész tanulmányi pályafutása során. Bevezetése persze nem ment zökkenőmentesen. Eleinte előfordult olyan eset, hogy egy tanuló, ha két intézményben tartották számon (pl. egy iskolában és egy szakképző intézményben) két azonosítót is kapott. Problémát jelentett továbbá, hogy majdnem egyszerre kezdték el bevezetni az oktatási azonosítót a közoktatásban és a felsőoktatásban, így jó ideig átfedések is voltak. Körülbelül 2010/2011-re jutott el a rendszer odáig, hogy az oktatási azonosító nagyjából be tudja tölteni a funkcióját, t.i. azt, hogy a tanuló egész iskolai pályafutása követhető legyen. Ehhez az is kellett, hogy az intézményeket rá lehessen bírni, hogy 5 napon belül jelentsék diákjaik belépését, illetve kilépését nemcsak az intézmény, hanem a program vonatkozásában is.

Sajnos az iskolák informatikai ellátottsága mindmáig nagyon különböző. Van olyan iskola, ahol nincs szélessávú internet vagy lassú és elavult számítógépek vannak. Súlyosabb probléma azonban, hogy az intézmények számára a megnövekedett adatszolgáltatási igény egyre nagyobb terhet jelent, ráadásul minden haszon nélkül, hiszen az összegyűjtött adatokból a központ nem szolgáltat vissza olyan összesítéseket és elemzéseket, amiknek az iskolák hasznát vennék. Ebben egyetlen kivétel az Országos Kompetencia Mérés, ami többek között azzal a céllal készült, hogy az iskolák össze tudják hasonlítani saját diákjaik teljesítményét más hasonló iskolákéival.

E: Milyen a főbb oktatásstatisztikai adatforrások mai állapota?

KFJ: Az „októberi statisztika” máig a legmegbízhatóbb adatforrás az oktatási rendszer állapotáról. Ez részben annak köszönhető, hogy ezeket az adatokat rendszeresen ellenőrzik, mégpedig több szervezet is, mert ebből szolgáltatunk adatokat a nemzetközi szervezeteknek (Eurostat, OECD). A köznevelési intézmények nyilvántartása ma már közhiteles nyilvántartásnak tekinthető, ami nagy előrelépés az 5-10 évvel ezelőtti állapothoz képest, és fontos a statisztikai elemzés szempontjából. A felsőoktatási információs rendszer mint igazgatási adatforrás kitűnő, statisztikai elemzésre kevésbé könnyen kezelhető. Meglehetősen sok fejlesztést igényel még annak megoldása, hogy a személyi nyilvántartásokból lehessen tanulói és pedagógus adatokat szolgáltatni, de nem kétséges, hogy ez a fejlesztés egyedüli útja. Ehhez elsősorban a személyi nyilvántartások statisztikai adatfeldolgozásához szükséges eszközrendszer kiépítésére van szükség. Általában igaz az, hogy az igazgatási nyilvántartások tervezői nemigen vették figyelembe azt az elvárást, amit a Központi Statisztikai Hivatal régóta hangoztat: t.i. hogy mind a pontosság és megbízhatóság érdekében, mind pedig költséghatékonyság okán arra kell törekedni, hogy a nyilvántartásokat tegyék statisztikai elemzésre alkalmassá, ahelyett, hogy újabb és újabb adatgyűjtéseket szorgalmaznának.

E: Miért volt szükség pedagógus-nyilvántartásra is?

KFJ: A finanszírozási rendszer változása és az új pedagógus előmeneteli rendszer bevezetése motiválta a pedagógus nyilvántartás létrehozását. 2013-ig az iskolák önkormányzati fenntartásban működtek, amihez az állam normatív hozzájárulást – fejkvótát – adott a tanulólétszám alapján. Ebben az ún. normatív finanszírozási rendszerben az önkormányzatok maguk döntötték el, hogy hány pedagógust vesznek fel, természetesen a törvényi

keretek között. Amikor a 2011-es köznevelési törvény alapján az állam részben átvette az iskolák közvetlen finanszírozását, a normatív támogatásról áttértek a pedagógusbérek és a közvetlen oktatási költségek finanszírozására. Emiatt a pedagógusbérek költségvetési tervezéséhez a személyes adatok központi nyilvántartása is szükségessé vált (pl. végzettség, életkor, illetve szakmai gyakorlat). A pedagógus életpálya modell bevezetése is megerősítette ezt az igényt, ami korábban fel sem merült.

E: Kialakultak-e az adatok előállításának és kezelésének egységes keretei?

KFJ: A 2011-es köznevelési törvény végrehajtási rendelete (229/2012. kormányrendelet) meghatározza, hogy az intézményeknek milyen adatokat kell nyilvántartaniuk, és az Oktatási Hivatalnak milyen nyilvántartásokat kell vezetniük. A technikai háttér a győri Educatio szervezetében adott volt, ezért az Oktatási Hivatal összes adatgyűjtését erre az elektronikus útra terelték. Az ún. Köznevelési Információs Rendszer keretében működő adatgyűjtések egyúttal kommunikációs felületek is, amelyeken keresztül az iskola bejelenti az újonnan érkező diákokat, megadja a felvételi jelentkezéshez szükséges adatait, illetve elküldi az érettségi jelentkezéseket, ami szerint az OH megszervezi a vizsgákat. Az Országos Kompetenciamérések szervezésének nagy része is ezen a kommunikációs felületen történik.

E: Hogyan lehet ezeket az adatbázisokat a jogosultságok kezelésén és a szervezésén túl elemzési célra is használni?

KFJ: A közoktatási információs rendszerben rengeteg adat van. Ezek egy része igazgatási vagy szervezési célra készült és ennek általában meg is felel. Ugyanezeket az adatbázisokat azonban többnyire komoly fejlesztés után lehetne rendszeresen elemzési célokra használni. Például a tanulói nyilvántartásban van kb. tíz millió rekord: adatok, amiket különböző időpontokban és intézményekben vettek fel ugyanarról a tanulóról. Az oktatási azonosító alapján a tanuló persze azonosítható, és az is megtudható, hogy az adott időpontban miért jelentették be. De ez más kérdés, mint az, hogy egy adott időpontban hány tanuló van a rendszerben. Vagy, hogy az összes közül mennyi a lány, hányan vannak a nyolcadik osztályosok, hányan ismételnék osztályt, stb. Ahhoz, hogy az efféle lekérdéseket rendszeressé lehessen tenni, az adatbázist kifejezetten erre a célra kivonatolni kellene, és évente egyszer egy adott ún. referencia időpontban „kimerevítve”, tisztítva és megfelelő dokumentációval ellátva szerkesztett állapotban kellene a statisztikusok rendelkezésére bocsátani. És ez nagyjából így van a többi nyilvántartással is. Ezzel a fejlesztéssel még adós az Oktatási Hivatal.

Jó jel, hogy az elemzés igényeit felismerve az Oktatási Hivatal megvizsgáltatta a KIR adatbázisokat a statisztikai elemzés lehetősége szempontjából, és javaslatokat kért egyes adatbázisok továbbfejlesztésére.

Az oktatási adatbázisokból származó adatok elemzése statisztikai és oktatáspolitikai szakértelmet igénylő munka. Jól érzékelteti ezt, hogy arra a gyakran feltett egyszerű kérdésre, hogy „Hány pedagógus van?”, az elemző általában zavarba ejtő viszontkérdéssel válaszol, t.i. hogy „Milyen célra kell az adat?”. Pedig nem bosszantani akarja a kérdés feltevőjét, hanem csak azt szeretné tudni, hogy hogyan számoljon. Csak a főfoglalkozású pedagógusokat kérdezik? Vagy a főfoglalkozású pedagógusok statisztikai létszámát (amiben a részfoglalkozásúak az óraszámoknak megfelelő súlyukkal szerepelnek)? Esetleg az összes pedagógus számát, aki megfordul az osztályokban? Vagy éppen azok létszámát,

akiket pedagógus státuszban foglalkoztatnak, noha beosztásuk szerint nem végeznek tanítási munkát? Mindegyik esetben más lesz a számítás eredménye. Mást is kell kérdezni, ha az egy pedagógusra eső gyermeklétszámot kérdezzük, illetve ha a pedagógusbérekre szánt összeget kell tervezni.

E: Ki fér hozzá az adatbázisokhoz és mire használják az adatokat?

KFJ: Sajnos jelenleg ez az egyik legégetőbb probléma. Az Oktatási Hivatal egyre kevesebb oktatásstatisztikai adatot tesz közzé, aminek többféle oka is lehet. Az egyik ok nyilván az a félelem, hogy az adatok minősége nem elég jó. Ami részben igaz is, csak éppen a megoldás nem az információ visszatartása, hanem ellenkezőleg: a nagyobb nyilvánosság és az adatok rendszeres használata, publikálása, ami kikényszeríti az adatok minőségének javítását. Ugyanis az adatszolgáltatók aszerint veszik komolyan az adatszolgáltatást, hogy mennyire van tétje a „lazaságnak” az adatszolgáltatásban. Két példát mondok erre. Az iskolai könyvtárak adatait a statisztika minden évben elkéri. Tudjuk, hogy nem egy iskola ugyanazt az adatsort adja minden évben. De ebből még sohasem lett baj és haszon sem származott belőle. Semmit sem publikáltak az adatokból, és sohasem kaptak az iskolai könyvtárak fejlesztési forrásokat könyvekre, attól, hogy nem gyarapodott a könyvtármányuk egyik évről a másikra. Az ellenkezőjére is van példa. Abban az időben, amikor az önkormányzatok normatív támogatást kaptak a tanulók után, szeptemberben az önkormányzatoknak le kellett adniuk a tanulólétszámot, mert ennek alapján kapták a támogatást. Ha ez a létszám több mint 5%-kal eltért az októberi statisztikától, a többlet támogatást vissza kellett fizetniük. Nos, nem mondanám, hogy nem igyekeztek pontosak lenni.

E: Van-e szervezett, egységes stratégia az oktatási adatvagyon felhasználására, menedzselésére, fejlesztésére, esetleg adat-összekapcsolások végzésére?

KFJ: Szerintem nincs. Erre abból következtettek, hogy az elemző kapacitás fejlesztése rendre elmarad, sőt inkább leépül az is, ami van. A háttérintézmények összevonása és a létszámcsoökkentések azt a benyomást keltik, hogy éppen az elemzés az, amiből az oktatáspolitikai nem kér. Pedig a rendszer statisztikai elemzése égetően fontos. Az oktatáspolitikai intézkedések hatását és mellékhatását csak ezen az úton lehet vizsgálni. Ha egy diák nem tanul meg olvasni nyolc év alatt, az elég nagy baj a diáknak és esetleg a pedagógus is kudarcként éli meg. De ha az analfabetizmus az országban minden tízedik gyereket érint, az már nagyon nagy baj. Arra való a statisztika, hogy megmutassa a probléma nagyságrendjét, hiszen ettől függ, hogy mikor, milyen módon kell beavatkozni, hogy a kedvezőtlen tendenciákat megállítsuk. Ugyanakkor az intézkedések hatását is figyelni kell. És ez is csak statisztikai elemzés segítségével lehetséges – még pedig hosszabb időszakon át összehasonlítható adatokkal.

E: Összességében milyennek látja az adatalapú oktatáspolitikai döntéshozás hazai helyzetét?

KFJ: Ma egy kicsit úgy néz ki, mintha a politika árnyékában az oktatásstatisztika és a statisztikai elemzés csak pár ember hobbjá lenne. Emellett van egy erős informatikai lobby, melynek elsődleges érdeke az informatikai termékek és fejlesztések eladása. Ezért a fejlesztési források rendre újabb és újabb „fejlesztések” felé terelődnek, miközben kidobjuk azt is, ami jó volt, de legalábbis nincs pénz annak a fenntartására. De azért remény-

kedem, mert látok az oktatáspolitikai területén néhány olyan embert (nem sokat), aki szeretne és tud is tenni azért, hogy az oktatási információs rendszer jelentősen javuljon. Emellett hiszek abban, hogy az oktatáspolitikai igazi támasza a nyilvánosság. Az oktatás végül is közügy, nemcsak a politikusok és az újságírók – vagy éppen a pedagógusok ügye. És a hiteles oktatási adatok olyan semleges tények, amelyeket – harag és részrehajlás nélkül (sine ira et studio) – megvitathatunk. No nem egymás legyőzése, hanem a legjobb közös megoldás keresése érdekében. Valahogy úgy, ahogy a finnek teszik.

Az interjú Hideg Gabriella készítette

„Az adatok nem önmagukért beszélnek, mi nézünk valahogy rájuk” Interjú Rudas Tamással

Educatio: Elsőként a TÁRKI-ban zajlott adminisztratív adat alapú kutatások áttekintésére kérném.

Rudas Tamás: Tudtommal a TÁRKI-ban az első ilyen jellegű nagy kutatás a '90-es évek közepén volt, és azt hiszem mind szakpolitikai döntés-előkészítő jellegét, mind tudományos tartalmát tekintve messze megelőzte a korát. E kutatás révén a kormányzat az adórendszer különféle elemein bevezetett változtatások hatásait kívánta modellezni. Ebbe a munkába részben adminisztratív adatállományokat is bevontunk – ezek legfőképpen maguk az adóbevallások voltak. Emellett több survey adatállománnyal dolgoztunk: részben a KSH háztartási költségvetés-felvételét, részben TÁRKI-s monitorfelvételek állományait kapcsoltuk össze többszörös imputációs eljárással. E köré a virtuális adatállomány köré építettünk egy valamelyest interaktív lekérdező felületet. Mikroszimulációnak neveztük ezt az elemzési módszert, melynek lényege, hogy az adott, meglévő adatokat tekintjük a valóság leírásának, és azt vizsgáljuk, hogy különböző policy-változatok milyen hatással vannak ezek alakulására. Ezt a modellt a kormányzat frissített adattartalommal még hosszú ideig – talán a 2000-es évek közepéig – használta. Maga az adatösszekapcsolási gyakorlat ekkor a világon már bevett eljárás volt ugyan, ám ez elsősorban record linkage-t jelentett. Ennek során az egyazon személyekről rendelkezésre álló rekordokat, feljegyzéseket azonosítjuk és kapcsoljuk össze. A mi eljárásunk virtuálisabb volt, hiszen az egyes adatbázisokban beazonosíthatóan nem ugyanazok a személyek szerepeltek, hasonlóság alapján kapcsoltunk tehát. Lett is valami sikere ennek a módszernek, de amennyire én tudom, a magyar közigazgatási tevékenységből kikopott. Fennállásáig, a Költségvetési Tanács volt az utolsó megrendelője a munka frissítésének.

E: Hogyan fejlődött ezek után a TÁRKI adatbankja?

R.T.: Az előbbi volt az első olyan munka a TÁRKI-ban, amely már ekkor tartalmazta ennek a most zajló nagy adatrobbanásnak minden elemét: nagy adatállományon zajlott, adminisztratív rekordokat használt, összekapcsoláson alapult, és szakpolitikai döntéselőkészítést szolgált. Azóta - és főleg jelenleg - a TÁRKI jóval kisebb adatgazda, mint azt kívülről gondolni lehetne. A TÁRKI adatbankja viszonylag hosszú ideig az összes, vagy majdnem az összes lényeges társadalomtudományi kutatás adatait megkapta különböző konstrukcióban. Ezek bárki számára ingyenesen elérhetőek és továbbra is rendelkezésre állnak, de az újabb adatállományokat már a TÁRKI saját adatfelvételei je-

lentik. Az adatszolgáltatási megállapodások megszűntek, a partnerek túlnyomó többsége nem kívánta újrakötni ezeket. Emögött két fő tendenciát látok magyarázatként. Az egyik az, hogy technikailag sokkal egyszerűbbé vált az adatok tárolása és közzététele. A másik pedig, hogy az adatok politikai és gazdasági értéke vált sokkal világosabbá. Azt gondolom, e két tényező vezetett oda, hogy pillanatnyilag Magyarországon – tudomásom szerint – nem működik olyan központi társadalomtudományi adatbank, mint amilyen hosszú ideig a TÁRKI adatbank volt.

E: Milyen trendeket lát jelenleg az adatkapcsolásban?

R. T.: Az adminisztratív adatállományok összekapcsolására vonatkozó legnagyobb kísérlet pillanatnyilag az egész világon a register based census. Ez azt jelenti, hogy a népszámlálást már nem azzal a hagyományos módszerrel végzik, amely igyekezett kapcsolatba lépni az ország minden egyes lakójával, akár személyesen, akár elektronikusan. Ehelyett a már létező regiszterek alapján próbálják az ország lakóit összeszámolni és bizonyos tulajdonságaikat megállapítani. Magyarországon ettől végtelenül messze vagyunk, de Ausztriában az utolsó népszámlálás már regiszter alapon zajlott. Ezek az adminisztratív adatállományok a táppénytől a közúti szabálysértésig tartalmaznak egymással összekapcsolható egyéni adatokat. Erre a lehetőségre is utaltam akkor, amikor a nagy adatállományok kezeléséhez, tárolásához és közzétételeéhez kapcsolódó politikai és gazdasági érdekek láthatóvá válásáról beszéltem. Például egy regiszter alapú népszámlálás nagyon sok pénz megtakarítását tenné lehetővé egy ország számára. A módszernek azonban nem ez a legnagyobb előnye a mostani hagyományos népszámlálási adatokkal szemben. A lényeges elem, hogy kivész bizonyos politikai döntési szabadságot a népszámlálásból. Ha ugyanis a népszámlálást regiszter alapon végezzük, akkor a regiszterek kötött adattartalma használható fel, míg ha megkereséses alapon, akkor a gyűjtött információk köre politikai döntéstől is függ. Ebben az értelemben a big data nem csak nagy lehetőség, hanem bizonyos értelemben véve korlát is az államigazgatás működésében. Persze ez az eljárás olyan regiszterek működését kívánná meg, melyekben érvényes és megbízható, friss adatok vannak.

E: Az adminisztratív adatok eltérő kezelése eltérő felhasználásukat is jelenti?

R.T.: Az adminisztratív adatok összekapcsolása a központi döntések előkészítése érdekében nem kizárólag technikai feladat. Az összekapcsolt adatállományokból kinyerhető információk nyilvánvalóan valamilyen politikai interpretáción mennek keresztül, mielőtt egy politikai döntés alapjait képeznék. Példa erre a felsőoktatásban az alapszakok számának adminisztratív adatösszekapcsolás eredményeire hivatkozó, közelmúltban zajlott módosítása. Tulajdonképpen ha egy más ország, más közigazgatási rendszerből hallok azt az érvelést, hogy az alapszakok listáján tervezett változtatások, a fenntartásukra avagy megszüntetésükre vonatkozó döntés adminisztratív adatok alapján történt, akkor mindez nagyon megnyugtatónak hallatszott volna. Viszont ebben az esetben úgy vélem, hogy nagyjából tudom, milyen adatok állnak rendelkezésre, és azok hogyan keletkeztek, és arról is van legalábbis véleményem, hogy ezeknek mennyire van közük az egyes szakok értelmes vagy értelmetlen voltához. Mindezen ismereteim alapján valahogy nehéz elképzelnem, hogy az adatokból erős politikai interpretáció nélkül is következne a meghozott döntés. Egy olyan közegben, amikor nemegyszer úgy vélhetjük, hogy az adat inkább csak

hivatkozási alapként szolgál, nagyon fontos erősíteni a hivatalos adminisztratív állományok megfelelő minőségű összekapcsolását és felhasználását.

E: A szabályozási keretek ehhez adottak?

R.T.: Magyarország a szabályozást tekintve egy viszonylag extrém helyzetben van. Ez alatt – nem minősítő, hanem leíró jelleggel - az adatvédelmi törvények igen erős voltát értem. Az erős szabályozó környezethez társul (személyes tapasztalataim alapján) a lakosság viszonylag alacsony szintű érzékenysége vagy érdeklődése az információ felhasználása iránt. Mondhatni, hogy bizonyos értelemben egy, a mostaninál liberálisabb politikai rendszer a lakosságot próbálta védeni a szigorú adatvédelmi törvények a bevezetésével. Ezeket a kutatók jelentős része támadja, hiszen lehetetlenné teszik a különböző adatállományok összekapcsolását. Nagyon nehéz az én pozícióból megítélni, hogy a szigorú adatvédelmi törvények inkább hasznosak vagy inkább korlátoznak egyébként hasznos tevékenységeket. Gyengének látom azt az intézményrendszert, amely ennek a betartásán őrködik. A magyar viszonyok kapcsán azt kell mondanom, hogy az egy dolog, amit a törvény leír és egy másik az, hogy ezt milyen mértékben és milyen pontossággal, precizitással tartják be. Zárójelben jegyezném meg ezzel kapcsolatban, hogy éppen TÁRKI-s tevékenységek kapcsán találok olyan helyzettel, amelyben a megrendelő adminisztratív adatállományok összekapcsolását kívánta volna tőlünk, amelyhez mi a nemzetközi szakirodalomban ismert algoritmusokat kívántuk alkalmazni. A meglévő szabályozási keretek azonban a megvalósítást jogi kérdéssé minősítették és a jogász kolléga, akinek a törvénnyel való egybevághóságot kellett megítélnie, nem tudott egyértelműen állást foglalni a kérdésben, az adott keretek között. A projekt így nem tudott megvalósulni. Azóta az adatvédelmi törvényeket lazították, és létezik egyfajta speciális mód az adatösszekapcsolásra, de a magyar törvények szigorúak e tekintetben.

E: Eszerint a jogi környezet nem kedvez a technikailag egyébként megvalósítható felhasználásnak?

R.T.: Mikor egy ilyen eljárás jogi kérdéssé transzformálódik, onnantól nem lesz sikeres. Korábban is volt egy államigazgatásból érkezett gazdasági vonatkozású adatkérésünk, amely során a minisztériumi támogatás ellenére döntöttek úgy az adatgazda jogászai, hogy nem látják biztosítva a megfelelő szintű adatvédelmet. Érthető részükről, hiszen az eljáráshoz nem jogi védelem, hanem nemzetközi szokásos algoritmikus védelem állt rendelkezésre. Hasonlóan például a bankautomatába beütött PIN kódhoz, az algoritmikus védelem technikailag akadályozza meg (vagy teszi hatékonytalanná) a kódok visszafejtését. Mindez bizonyos matematikai és statisztikai eljárásokkal biztosítható, mint ahogy az egész világon komoly pénzügyi tranzakciók zajlanak ezen algoritmikus védelmek ernyője alatt. De ez nem jogi garancia.

E: Az adatok felhasználása a technikai és jogi keretek mellett az államigazgatás részéről bizalmat is feltételez. Megvan ennek a bázisa?

R.T.: Az adatfelhasználást valóban nagyon nagy mértékben beárnyékolja az a bizalomhiány, ami az egész ország, az egész társadalom működésének sok területén látszik. Az alap az, hogy sokan bizonyos naivitással adatokat szolgáltatnak. Azok a hozzáértő szervek és emberek pedig, akiknek az ezeket védő törvény betartása felett kellene őrködniük, inkább bizalmatlansággal viseltetnek a lehetséges felhasználók iránt, legyen az ku-

tató vagy gazdasági vállalkozás. Tudomásom szerint nagyon kis mértékben valósul meg az, hogy gazdasági vállalkozások által gyűjtött óriási és szerintem fantasztikus adattartalmú állományokhoz pusztán kutatási célra hozzá lehessen férni, leszámítva persze az adott vállalkozás érdekében történő hozzáférést, hiszen ez az adatgyűjtésnek célja. Tényleg azt hiszem, hogy a bizalmatlanság írja le leginkább a szereplők hozzáállását. Ezt a bizalmatlanságot azonban nem érzem indokolatlannak. Nyilván adott egy államigazgatás, amely nem feltétlenül készült fel szakmailag az adatrobbanás megértésére. Az adatokat termelő procedúrák elmúlt évtizedekben bekövetkezett gyors változására és ennek társadalmi hatásaira senki sem volt – én sem – felkészülve. Érthető tehát az államigazgatás bizonyos fajta óvatossága.

E: Ugyanezt a tendenciát érzékeli a lakosság körében?

R.T.: Maga a bizalmatlanság társadalmi szinten is érzékelhető. A big data valamilyen szinten az információs monopólium csökkentésére is képes, ami még a hozzáértő felhasználókban is nagyon erős és jogos bizalmatlanságot szülhet. Egy pár perccel ezelőtti jó példát is mondhatok erre. Két számítógépet lát az asztalon. Az egyik gépen a booking.com-on néztem egy szálloda adatait. Nem volt nyitva semmilyen ablak, amiben e-mail lett volna és nem jelentkeztem be a booking.com rendszerébe. Ennek ellenére, amint ott bezártam az ablakot, egy perc múlva érkezett egy e-mail a másik gépemen nyitva levő, egy másik e-mailemhez csatolt ablakban, amelyben az állt, hogy a booking.com látta, hogy abbahagytam a keresést, és ha gondolom, ide kattintva tudom folytatni. Sejttem persze, hogy valahol az apró betűs részben jóváhagytam az eljárást. A mobiltelefonra letöltött applikációk esetében a helyzet hasonló, amennyiben a működésükhöz engedélyezni kell, hogy bizonyos információkhoz hozzáférjenek. Egy ilyen közegben tényleg azt hiszem, hogy jó lenne, ha a lakosság tudatosabb lenne a saját adatainak megadásával kapcsolatban. Amíg tehát a politikai döntéshozók részéről az információs monopólium elvesztésétől való féltelmet és az ebből fakadó bizalmatlanságát látom, addig a lakosságnál néha meglepő az óvatosság hiánya. Számos olyan törzsvásárlói program létezik például, amely 50 levásárolható pontért elkéri a telefonszámot, e-mail címet, születési időt és egyéb személyes adatokat. Semmi ördögít nem látok ezekben a tevékenységekben, amennyiben deklarált, világos módon és feltételekkel zajlanak. De az azért jó volna, ha mindenki tudatosabban döntene az adatközlésről.

E: Hogyan reagál minderre a társadalomkutatás?

R.T.: Ahogy előbb is említettem, a társadalom egészen máshogy működik ma, mint 10 vagy 20 éve. Ilyen mérvű fejlődést még a társadalomkutatók sem láttak előre. Az elképesztő mennyiségű rendelkezésre álló adat érdemi felhasználása és elemzése azonban – részben az adatgazdák érdekei miatt – csak bizonyos területekre korlátozódva történik meg. Ha például egy szupermarket gyűjti a vásárlások adatait, akkor azt kizárólag saját fejlesztési céljaira használja fel és csak nagyon ritkán válik hozzáférhetővé más típusú elemzések számára. Ez nem egy örömteli dolog. Egy másik, amit tendenciaként látni vélek a világban, az az új eljárások – mint az adatbányászat vagy a machine learning – speciális humán erőforrás igénye. Ezeket a tevékenységeket elsősorban nem társadalomtudósok, hanem mérnökök és fizikusok végzik. Nyilván ezek az emberek nagyon ügyesek az algoritmusok gyártásában, szoftverek írásában, tehát az adatbázis tevékenység, gépi programozás nagyon magas színvonalon tud megvalósulni. Ami viszont ezeknek a szak-

embereknek az inkább természettudományi, mint társadalomtudományi háttéréből adódóan nehezebben valósul meg, az az eredményekhez tartozó helyes interpretáció. Érdekes kérdés persze, hogy mi is a helyes interpretáció. Azt gondolom, ennek legfőbb jegye mindenképpen a fenntartás. Vagyis nem gondolni azt, hogy a társadalmi vagy gazdasági jelenség mögött, amelyet valamiféle big data forráson éppen vizsgálunk, egyetlen összefüggés, egyetlen igazság van. Vagy akár azt, hogy az általunk azonosított összefüggés bizonyosan helyes. Minden adatállománynak lehetnek alternatív interpretációi, elemzései, még akkor is, ha sok, a survey-t jellemző bizonytalanság az adminisztratív adatállományokat nem jellemzi. Úgy látom, a szokásos értelmezésből ma hiányzik az a nézet, hogy az adatok nem önmagukért beszélnek, hanem mi nézünk valahogy rájuk. Ez a szemléletmód az adminisztratív adatok kutatási felhasználásában változatlanul releváns. Ennek kezelésében maga a világ van lemaradva abból adódóan, hogy a hatalmas és egyre növekvő hozzáférhető adminisztratív adatállományokat jóval kevésbé társadalomtudósok, sokkal inkább mérnökök és fizikusok elemzik. Ezt a szomorú tendenciát látom legerősebbnek.

E: Mindez a társadalomkutatók iránti elvárásokra is kihat?

R.T.: Szerintem nyilvánvaló, hogy itt is fejlődésre, új kompetenciákra van szükség. Már 30-35 évvel ezelőtt elindult egy változás, melynek során a társadalomkutatók akkori új generációja elkezdett a survey követelményeihez igazodó skilleket megszerezni. Ez a tudáskészlet most, az adatbányászati lehetőségek és feladatok következtében frissítésre, bővítésre szorul. A társadalomtudományi felhasználásban a szükséges változást nehezíti, hogy a technikai belépési küszöb nagyon magas, hiszen mind az alkalmazott algoritmusok, mind az ezeket megvalósító IT eszközök rendkívül szofisztikáltak.

Az interjút Galántai Júlia készítette