kedem, mert látok az oktatáspolitika 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 oktatáspolitika 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 interjút 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 at 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 táppénztől a közúti szabálysértésig tartalmaznak egymással összekapcsolható egyéni adatokat. Erre a lehetőségre is utaltam akkor, amikor a nagy adatállományok kezeléséhez, tárolásához és közzétételéhez 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 kivesz 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 rendszeréből hallom 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. Ezalatt – 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 pozíciómból 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 találkoztam 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 egybevágóságot 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. Sejtem 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 félelmet 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 ördögit 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 humánerőforrás 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 hátteréből 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 z mögött 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