ISMERTETŐK

PÁRI ANDRÁS

A mikroadatok hozzáférésével és az adatok felfedés elleni védelmével kapcsolatos kérdéseiről

„A mikroadatok hozzáférésével és az adatok felfedés elleni védelmével kapcsolatos kérdéseiről" szóló műhelykonferenciát 2012. november 6-án rendezte meg a Statisztikai Felhasználói Tanács (SFT), a Magyar Szociológiai Társaság (MSZT), a Magyar Statisztikai Társaság (MST) Társadalomstatisztikai szakosztálya, valamint az MST Területi szakosztálya szervezésében a Központi Statisztika Hivatal Keleti Károly termében. A témában kezdeményezett párbeszéd elsődleges célja a konszenzusos megoldások kialakítása volt a mikroadatokhoz való kutatói hozzáférés és a felfedés elleni védelmével kapcsolatos kérdésekben.

A konferencia vagy inkább műhelybeszélgetés két fontosabb téma köré összpontosult, nevezetesen a kutatók mikroadatokhoz való hozzáféréséhez és a 2011. évi népszámlálás adatainak eléréséhez kapcsolódóan. Az előbbi téma az egész napos konferenciabeszélgetés délelőtti (négy előadás), míg az utóbbi a délutáni (egy előadás) részében zajlott le. A rendezvény jellege nehezen meghatározható, hiszen a hagyományos konferencia és a műhelybeszélgetés elemei vegyesen ötvöződtek. Az előadók alapos szakmai prezentációi és az előadások formai keretei a konferencia jegyeit hordozták, azonban az előadások közben, illetve az azokat követő kérdések, javaslatok, észrevételek egyértelműen inkább a műhelybeszélgetés és a közvetlen szakmai párbeszédről szóltak. Ezért a szervezők által meghirdetett műhelykonferencia meghívója teljes mértékben tükrözte a rendezvény jellegét.

A vitavezető *Dr. Kovách Imre*, az MTA Társadalomtudományi Kutatóközpont Szociológiai Intézet igazgatójának bevezetője után a nyitó előadást *Vereczkei Zoltán*, a KSH Módszertani osztályának vezetője tartotta az adathozzáféréssel kapcsolatos nemzetközi előírások és azok gyakorlati alkalmazásáról. Kiemelte, hogy a magyar adatvédelmi szabályok egy magasabb nemzetközi és EU-s szabályozási környezetnek megfelelően kell, hogy érvényesüljenek. Ilyen jelentős jogszabály a 223/2009/EK rendelet. A kötelező szabályokon kívül – több esetben párhuzamosan – alkalmazandó nem kötelező érvényű előírások, az iránymutatások vagy irányelvek (*guideline*), valamint a kézikönyvek (*handbook*) is hasonló jelentőséggel bírnak gyakorlati oldalról. A hallgatóság részéről megemlítették, hogy az Európai Unió statisztikai törvénye viszont nem határozza meg pontosan ki a kutató, vagyis ennek meghatározása nem kógens[1] jogszabályi keretek között történik és így a tagállami elbírálás vagy meghatározás diszpozitív jelleget kap.

A mikroadatokat illetően az Európai Unió egy újabb rendeletben szabályozza az ezekhez való hozzáférést (μ-Argus rendszer). Az EU gyakorlati kódexe került még szóba, ami a téma szempontjából fontos kutatói hozzáférést és a bizalmas adatok kezelését tartalmazza. A kutatószobák létrehozása mint a biztonságos környezetben való hozzáfé-

1 Kógens: a meghatározottól eltérést nem engedélyező.

rés helyisége az adatvédelem és az adathozzáférés szempontjából kiváló megoldásnak tűnik. Addig, amíg a kutatók – titoktartási nyilatkozat aláírását követően – hozzáférhetnek a számukra fontos és elemzésükhöz, pályázataikhoz nélkülözhetetlen mikroadatokhoz, a külső felhasználók számára ez elérhetetlen. Ennek kapcsán egyrészt az előadás közben is több kérdés, megjegyzés és észrevétel is érkezett az előadóhoz, hogy a kutatószoba és a hozzáférés részleteit tisztázni és pontosítani szükséges, másrészt pedig a korreferensek *Dr. Halpern László*, az MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaság-tudományi Intézet igazgatóhelyettese és *Dr. Benczúr Péter*, a Magyar Nemzeti Bank kutatásvezetője is kitértek erre a kérdésre a prezentációt követően.

Fontos itt megjegyezni, hogy a KSH-ban már korábban is működött kutatószoba az Információszolgálaton. Az utóbbi időben megnőtt az igény a kutatószoba szolgáltatásai iránt, köszönhetően a népszámlálási adatok megjelenésének, valamint a kutatók és az elemzők mikroadatok iránti érdeklődésének, ezért a KSH-nak az ilyen jellegű adatkérési megkeresésekre tekintettel kell lennie, a kutatószobában szükséges kiépíteni és üzemeltetni az igényeknek megfelelő számú számítógép-hálózatot, másrészt kutatószoba-hálózatot lenne érdemes létrehozni az igazgatóságokon is, ahogy ez a délutáni műhelybeszélgetés során említésre is került. Az ESS DARA-program került elő példaként, ami az Eurostat kutatószobájából távoli hozzáféréssel is elérhető. A távoli végrehajtás során a kutató beküldi az igényét, majd a statisztikai hivatalban lefuttatott program kiszűri és leválogatja az elemzendő adatokat.

A korreferensi vélemények a KSH-val kapcsolatos tapasztalatait megfogalmazva a végrehajtáshoz szükséges időintervallum jelentőségét emelték ki „hiszen bármennyire is hihetetlen a kutatók is határidőre dolgoznak" – jegyezte meg az egyik korreferens. Válaszként erre a kapacitás és a munkaerőforrás korlátaival érvelt Vereczkei Zoltán, amellett, hogy egyetértett és szükségesnek tartja, hogy a kutatók ésszerű időkorlátot várhatnak el a hivataltól. A minnesotai modellt említették még a hallgatóságból, amelyhez hasonló letölthető nyers adatbázis közzétételén a KSH-ban is gondolkodnak.

Az adatkérők (például ki lehet adatkérő), az egyedi adatkérések és az adatvédelmi politika kialakítása, valamint az adatkiadás gyakorlata került említésre az előadás végén. Az egységesítés szükségességét emelte ki ezzel összefüggésben az előadó, különösen a legelső és legutóbbi esetében – válaszolva ezzel az előadás során felmerült kérdésre, amely a szakfőosztályi adatkiadási döntési szempontokra vonatkozóan fogalmazódott meg.

Dr. Halpern László megemlítette, hogy az eredmények, adatok kiadásakor (*output checking*) a Módszertani főosztály értelmezést ad az adatokhoz annak ellenére, hogy a kutató felelőssége az adatfelhasználás és az adatok elemzése. Külön kiemelte a költségkialakítást is, miszerint a költségekről szóló kiadási listát osszák meg a kutatóval. Így ennek fényében a kutató dönthetne arról, hogy kéri-e az adatokat vagy sem. Dr. Benczúr Péter az adminisztratív és survey típusú adatok közötti különbségről beszélt, amerikai példákkal részletesen illusztrálva, kiemelve az adminisztratív adatok felértékelődését, és példaként a hosszú távú időszakra történő adatsorok előnyeit méltatta. Az IRS-adatbank példáján keresztül az adatbázisok összekapcsolásával az Egyesült Államokban olyan adathalmazok válnak elérhetővé és nyilvánossá, amelyeket hazánkban el sem tudunk képzelni. Ilyen adatok például a munkajövedelem, az egyetemi tanulmányi tandíj, a jelzáloghitelek kamatai, a családi állapot, a lakóhely, az adózási adatok stb. egy adatbázisban tárolása. (A tengerentúlon hatályos adatvédelmi törvény és a hozzá kapcsolódó elvek,

valamint a szankciórendszer és a közgondolkodás a magyar adatvédelmi szabályozástól eltérően liberálisabb, azonban a társadalmi környezet, amiben ezt szabályozni hivatott, teljes mértékben különbözik az említettől. A szankciórendszer a két ország esetében nehezen összehasonlítható, hiszen az Egyesült Államokban korlátlanul perelhető az, aki más személyes adataival visszaél.)

A korreferensek egyetértettek abban, hogy a kutatószoba megléte és bővítése jó irány a mikroadatokhoz való hozzáférés tekintetében, és a távoli adatfelhasználás is pozitív kezdeményezés, ha a következő években ezt majd folyamatosan fejlesztik. Dr. Benczúr Péter javasolta, hogy a KSH lehetne egy ”gyűjtőhely” – az amerikai példánál maradva –, így a több intézménytől bekért adatokat egy helyen lehetne elérni és lekérdezni. Ennek első lépéseként természetesen a jogi szabályozást kellene módosítani. Kérdésként merült fel, hogy van-e, vagy várható-e kezdeményezés ebbe az irányba.

Kodaj Katalinnak, a Nemzeti Fejlesztési Minisztérium főreferensének előadása az adathozzáférés jogszabályi környezetének változásáról szólt. Az adatszolgáltatók bonyolult és szövevényes kötelezettségéről és az adatbázisok összekapcsolásának lehetőségéről tett említést, mint elérendő célról. A prezentáció során az anonimizálási törvény, a közadatok újrahasznosításáról (*Public Sector Information*) szóló törvény, az infotörvény és a statisztikai törvény került elő. Az első törvény esetében módosítási javaslatot kívánnak kidolgozni, amely lazítaná a jelenlegi szigorú adatvédelmi elveket, így a kutatók adatkérővé minősítésével könnyebben elérhetővé válhatnának egyes adatbázisok. A kutató fogalmának kérdése itt is előkerült a hallgatóság részéről és válaszul az OTKA kutatási listáját említette Kodaj Katalin. Szerinte arra a kérdésre, hogy ki a kutató, sokkal inkább szakmai választ kell adni, mint bármilyen minisztériumi meghatározást.

A második – a közadatok újrahasznosításáról szóló – törvény említése nem volt igazán mérvadó, mert a gyakorlati jogalkalmazás szempontjából még értelmezhetetlen, hiszen 2013. január 1-jén lép majd hatályba. A törvény az OSAP-adatokra épít, amelyekkel majd egyes közadatok elérhetővé válnak. A 2011. évi CXII. törvény (az információs önrendelkezési jogról és információszabadságról) kapcsán a szankcionálás hiányát, míg a statisztikai törvény említésekor azt emelte ki, hogy a törvény értelmében a KSH nem anonim módon kapja az adatokat, viszont az adatfeldolgozás során anonimizálja azokat, így az elemi adatokból anonim módon hozzáférhetők lehetnének az adatok. Az adatok használhatósága szempontjából véleménye szerint hátrány, hogy nincs egységes adatszolgáltatás és így az adatok érvényessége is kérdéses. Az előadó felvetéseit nem korreferálták.

A délelőtti szekció második előadása után érkeztek hozzászólások, kérdések, amelyek közül *Dr. Németh Zsolt*, a KSH társadalomstatisztikai elnökhelyettese megerősítette az októberi Statisztikai Felhasználói Tanács találkozóján elhangzott azon álláspontját, miszerint a statisztika bizalmi üzlet, vagyis amíg kap adatot a KSH, addig ad is. Hozzáfűzte, hogy az adminisztratív források elsősorban nem statisztikai célra készülnek, ezért ezek feldolgozása sok idő- és erőforrás-ráfordítást igényelnek. Éppen ezért az adminisztratív adatok ”lassabbak”, mint a survey adatok, de mindkettőnek van előnye és hátránya. Nem szabad figyelmen kívül hagyni azt sem, hogy a KSH nem nyilvántartó hivatal, vagyis amikor a rendőrség vagy a bíróság keresi meg a hivatalt nem statisztikai célú adatkéréssel, akkor azt el kell utasítani. Végül az ár-költség témájához reagálva elmondta, hogy nem a profit a cél, hanem a szolgáltatás – utalva a statisztika bizalmi jellegére.

A harmadik előadást megosztva tartotta *Dr. Cseres-Gergely Zsombor*, az MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaság-tudományi Intézet tudományos munkatársa és *Dr. Fábián Zoltán*, a TÁRKI (Társadalomkutatási Intézet Zrt.) vezető kutatója. Az előadás a kutatók oldaláról megfogalmazott álláspontokat és kérdéseket helyezte a középpontba és több ponton érintette a legelső előadást. Dr. Cseres-Gergely Zsombor szerint figyelembe kell venni a munka célját és a környezetet, amibe illeszkedik a munka, így ennek megfelelően lehetne mérlegelni az adatkiadást, ezzel rugalmassá téve az adat- és információáramlást. Az adatok hozzáférése kapcsán a projektgondolkodás a versenyképességet fokozná, vagyis az időben megkapott adat a verseny szempontjából elengedhetetlen. Az információvesztésre is felhívta a figyelmet, amely az adatvédelem következtében keletkezik, ez pedig – a versenyképesség gondolatát tovább fűzve – hátrányt eredményez, méghozzá versenyhátrányt. Az információvesztés kiküszöbölésére azonban szerinte a kutatószoba nem a legjobb megoldás. A harmonizált és származtatott adatok vonatkozásában a kutatószoba nem alkalmas.

Dr. Fábián Zoltán egy norvég társadalomtudományi adatbank, a Norwegian Social Science Data Services (NSD)[2] példáját, a *'public use file'* és a *'scientific use file'* mint az elérhető biztonságos környezetben való hozzáférés és az adathordozón kiadott adatállományok közötti különbséget említette meg.

A korreferens, *Gárdos Éva*, a KSH, statisztikai főtanácsadója a kutatók támogatásáról, az adatvédelmi szempontok betartásáról és az adatszolgáltatói terhek csökkentéséről beszélt. Javasolta, hogy a kutatók adatokhoz való hozzáférésének feltételeit tisztázni kellene, legyen meg a kapcsolat az adatgazdákkal (együttműködési megállapodások jelentősége), továbbá távoli kutatószoba megvalósítását szorgalmazta.

Dr. Lakatos Miklós, a KSH adatvédelmi felelőse, statisztikai főtanácsadója a kutatói felvetésekre, előadásokra reflektált. Az Adatvédelmi Bizottság munkájáról beszélt és arról, hogy idáig összesen 12–13 KSH-s állásfoglalás született adatvédelmi kérdésekben. Felhívta a figyelmet arra, hogy az anonimizált adatok hozzáférhetők, itt az adatvédelemnek nem kell szükségszerűen megjelennie, viszont a nem anonimizált adatok esetében elengedhetetlen a megfelelő módon történő kezelésük. Kiemelte, hogy nem a személyes adatokkal van a probléma, hanem a gazdasági szervezetek adataival.

További hozzászólásokat *Dr. Tardos Róbert*, az MTA–ELTE Kommunikációelméleti Kutatócsoportjának tudományos főmunkatársa és *Dr. Róbert Péter*, a TÁRKI vezető kutatója részéről hallhattunk. Róbert Péter szerint az adatvédelem említésekor két párhuzamos világ tűnik fel egymás mellett, ebből az egyik a statisztikai hivataloké – nem csak a KSH, hanem általánosan a statisztikai hivataloké – és a kutatói világ. A kettő közötti összhang és együttműködés kellene, hogy elérendő cél legyen.

Dr. Bartus Tamás, a Budapesti Corvinus Egyetem docense a felfedés elleni védelem statisztikai következményeit illusztrálta *Az anonimizálás esetei és szabályai* címmel. Egy többváltozós elemzéssel az anonimizálásra fókuszálva változtatta meg az iskolai végzettség, a településtípus és más változók értékét. Kimutatta, hogy a súlyozott átlagban és szórásbecslésben kismértékű torzulás, míg a korrelációs együtthatókban komolyabb torzulás következett be. Minél nagyobb részt védünk az adatbázisból, annál nagyobb lesz a torzulás. Kérdésként vetette fel, hogy az anonimizált adatbázis használható-e, majd a

2 http://www.nsd.uib.no/nsd/english/index.html

kutatási eredményei alapján rámutatott arra, hogy ez nagymértékben függ a védett adatok méretétől és a kifinomult technikák az adatokat torzíthatják.

Dr. Szép Katalin, a KSH statisztikai főtanácsadójának korreferensi válaszát *Dobány Máté*, a KSH statisztikusa tolmácsolta, aki elismerését fejezte ki a felfedés elleni védelemről szóló elemzésről. Javaslata szerint a fogalmak és a módszertani megfogalmazások pontosítása szükséges, hiszen olyan téma került górcső alá, amelynek fogalomtára még nem tisztázott.

A délelőtti előadások lezárása előtt *Dr. Harcsa István*, a KSH statisztikai főtanácsadója emelte ki, hogy a legnagyobb probléma az, hogy a két oldal nem érti egymást, ezen kell elsősorban változtatni, míg *Szabó István*, a KSH Tájékoztatási főosztályának vezetője hozzátette, hogy az „adatért elemzés" elvét a KSH is alkalmazza, vagyis KSH-s adatokat külsős kutatók is elemeznek. Kihangsúlyozta az együttműködési keretmegállapodások fontosságát, amelyek keretében mindkét oldal jól jár.

A délutáni részben Dr. Harcsa István váltotta Dr. Kovách Imrét a vitavezetői székben. A népszámlálási adatokhoz való hozzáférés és ezen belül a területi részletezettségű adatokról szóló összefoglalót Szabó István tartotta. A népszámlálási adatok adatszolgáltatási körének (kinek, mit, hogyan és mikor) felvetésével kezdődött az előadás. Szóba kerültek a népszámlálási adatok publikálási, tájékoztatási időpontjai. A következő évi tájékoztatási tervben szerepel, hogy 2013–2014-ben összesen 14 kötetet jelentet majd meg a KSH (a 2001-es népszámlálás során 30 kötetet adott ki a hivatal). A 2011-es népszámlálás előzetes adatairól már két kiadvány elkészült, a részletes adatok közlése 2013 márciusában kezdődik. Az elektronikus tájékoztatás során a KSH honlap népszámlálási aloldaláról Excel-formátumban elérhetők lesznek a népszámlálási adatok, 2014-től pedig dinamikus táblázatokat is le lehet majd kérni. A térképek települési szinten (legkisebb egység a tömb lesz) lesznek megjelenítve. A kutatószoba működtetéséről Szabó István elmondta, hogy az a Tájékoztatási főosztály feladata, a megyeszékhelyeken pedig kutatószobai feltételekkel működő gépek telepítését kezdeményezik, megkönnyítve ezzel a vidéki kutatók, elemzők munkáját és hozzáférését az adatokhoz.

A három korreferens közül elsőként *Dr. Kovács Katalin*, az MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Regionális Tudományi Intézet tudományos osztályvezetőjének nevében *Koós Bálint*, az intézet tudományos munkatársa fejtette ki kollégáival egyeztetett közös véleményét. A legfontosabb elemként a határidőt emelte ki, hivatkozva az EU-s fejlesztési tervekre, amelyek 2014–2020 közötti projekteket kívánnak támogatni. Ezért a 2014-től kezdődő projekteknél elengedhetetlen a használható adatok gyors megjelenése. A területi adatokra reflektálva megkérdezte, hogy lesznek-e számlálókörzeti adatok, továbbá az informatikai háttér kérdése is felmerült, vagyis milyen programcsomaggal lehet majd elemezni az adatokat. A második korreferensi véleményt *Kabos Sándor*, az ELTE docense, módszerfejlesztéssel foglalkozó szakértő osztotta meg, aki a térbeli statisztikai modellezéssel, autokorrelációval kapcsolatos kérdéseit vetette fel. Többnyire olyan észrevételeket fogalmazott meg, amelyek a következő népszámlálás tervezéshez kapcsolódó javaslatokként rendkívül értékesnek bizonyulhatnak. A harmadik korreferens, *Rácz Attila*, a Szegedi Tudományegyetem tanársegédje az egyetemisták és a főiskolai hallgatók néhány szempontját is felvillantotta. Szerinte a település mint elemzendő területi egység túl nagy, a számlálókörzeti egység lenne erre a megfelelő. Kitért a konferenciabeszélgetésen már korábban említett kutatói kör meghatározásának kérdésére

is. Továbbá megerősítette a korábbi előadásokból azt a kérést, hogy az adatok legyenek minél előbb elemezhetők és a kutatói összefogást sürgette. A minnesotai modellel összefüggésben itt is megjegyezte, hogy jó lenne, ha az adatok nyers adatbázis formájában letölthetők lennének például a KSH-honlap népszámlálási aloldaláról, így az egyetemi hallgatók gyakorlati órák keretében találkozhatnának és elemezhetnék a népszámlálási adatokat.

A műhelykonferencia zárása előtt a népszámlálási adatok feldolgozásával kapcsolatos kérdésekre *Kovács Marcell*, a KSH Népszámlálási főosztályának osztályvezetője válaszolt. A népszámlálási kérdőívek szabadszavas kitöltéssel rögzített részeinek feldolgozását 2012 decemberére fejezik be, és megerősítette, hogy a népszámlálási kötetek 2013 márciusától kezdve jelennek majd meg. A népesség visszavezetésével kapcsolatos kérdésre azt a választ kaphattuk, hogy ez folyamatosan zajlik. A területi adatok (geokódolt állomány) óriási értéket jelentenek, itt a legalacsonyabb szint várhatóan a tömb lesz majd.

A délutáni részt Dr. Harcsa István zárta, megköszönve az előadók és a hallgatóság részvételét, és reményét fejezte ki, hogy tovább folytatódik a párbeszéd a kutatók és a (statisztikai) hivatal között.