

2.2. A KRTK KAPCSOLT ÁLLAMIGAZGATÁSI PANELADATBÁZISA – ADMIN4

SEBŐK ANNA

A Közgazdaság- és Regionális Tudományi Kutatóközpont (KRTK) Adatbankjában¹ 2022 nyarán létrejött a legújabb, sorrendben a negyedik Kapcsolt Államigazgatási Paneladatbázis, az Admin4. A kapcsolt államigazgatási paneladatbázisok, így az Admin4 (ami a 2003–2021 közötti időszakot öleli fel) is, adatintegrációs eljárással készültek, anonimizált módon, ugyanakkor egyéni szinten tartalmazzák a magyar lakosság felének egészségügyi, oktatási, munkaerőpiaci és munkanélküliségi adatait, valamint a magyar vállalatok Bértarifa-felvételben szereplő tulajdonságait. A létrejött adatbázisok nem tartalmazzák természetes azonosítókat és háztartástáblát, ugyanakkor kutatási szempontból egyedülállóan részletesek. A korszerű nemzetközi tudományos célú adatkezelő rendszerek kritériumainak megfelelően, az adatok biztonságos szerverkapcsolattal, kontrollált körülmények között, kizárólag kutatási céllal kerülnek felhasználásra.

Az adatbázis korábbi hullámai (az Admin1, Admin2 és Admin3) széles körben ismertek és elismertek a hazai és nemzetközi munkaerőpiaccal, egészség-gazdaságtannal foglalkozó kutatók között, és a közgazdaságtan egyéb területein. Erről részletesebben Köllő János ír jelen Közelpép 2.1. alfejezetében.

Admin4

Az államigazgatásban keletkező információk kutathatósága nagy múltra tekint vissza Magyarországon. A több nyilvántartás együttes vizsgálatát célzó kutatási adatbázisok létrehozásának lehetőségét és módját eddig a döntés-előkészítéshez szükséges adatok hozzáférhetőségének biztosításáról szóló 2007. évi CI. törvény rögzítette (Scharle, 2019). Az egyesített adatbázisok előállításához szükséges adatintegrációt a fenti törvényben rögzített költségvetési szervek vezetői indíthattak (tudományos célból az MTA elnöke), amelyet technikailag kizárólag a Nemzeti Infokommunikációs Szolgáltató Zártkörűen Működő Részvénytársaság (NISZ Zrt.) hajthatott végre.

A Kapcsolt Államigazgatási Paneladatbázis összes hullámának alapelve, hogy az adott összekötés időpontjában fellelhető és összekapcsolható összes kutatási relevanciájú regisztert egyesíti. Így a legutóbb összekapcsolt Admin4-ben a Nemzeti Egészségbiztosítási Alapkezelő (NEAK), a Magyar Államkincstár (MÁK), az Oktatási Hivatal (OH), a Nemzeti Szakképzési és Felnőttképzési Hivatal (NFSZH – SZIR) és a Technológiai és Ipari Minisztérium (TIM) egyéni és vállalati szintű adatai lettek összekötve. Az összekapcsolásnak köszönhetően összefüggéseikben is kutatható adatkörök a következők.

1 Korábban: Magyar Tudományos Akadémia Közgazdaság- és Regionális Tudományi Kutatóközpont (KRTK) Adatbankja.

Egészségügyi terület: laccím adatok (regionális szinten), társadalombiztosítási jogviszonyra vonatkozó, közgyógyellátási, háziorvosi, járóbeteg- és fekvőbeteg-ellátási, halálozási, vénnyírási, gyógyszerkiadási, pénzbeli, CT/MRI- és fogászati ellátásra, valamint nagyértékű beavatkozásokra vonatkozó adatok, egyéni szinten.

Munkaerőpiaci terület: munkavállalói, munkaerőpiaci, közfoglalkoztatási és munkaerő-kiközvetítési adatkörök, egyéni szinten.

Szociális transzferek területe: nyugdíjfolyósítási, családi pótlék, pénzbeli ellátás, munkanélküliségi, munkaerőpiaci programokhoz kapcsolódó adatok, egyéni szinten.

Oktatási terület: felsőoktatási képzési, felsőoktatási jogviszony, köznevelési jogviszony, érettségi, Országos Kompetenciamérés adatai, egyéni szinten.

Vállalati terület: társasági adó (tao) bevallásából és a Bértarifa-felvételből származó adatok, vállalati szinten, ugyanakkor egyénekhez köthetően.

Az adatok biztonságos összekötése

Az adatintegrációs eljárás alapja a különböző államigazgatási regiszterek egyéni szintű, esetleg valamely egyéb egyedi azonosítóval rendelkező egységek szerinti összekötése. Az adatintegráció eredményeként a különféle adminisztratív regiszterek tartalma együttesen válik elemezhetővé, egyénenként vagy vállalatokként, mégis anonim módon. Az adatgazdánál megtalálható adatokat kapcsolati kódok alapján kötik össze, majd törlik a természetes azonosítókat. Kapcsolati kód lehet minden olyan egyedi azonosító, amely több adatgazdánál is fellelhető.

A KRTK Adatbank Kapcsolt Államigazgatási Paneladatbázisa, szemben az államigazgatásban létrehozott adat-összekötésekkel, nem teljes körű, a TAJ-nyilvántartásban szereplők 50 százalékos mintájáról szól (335/2007. [XII. 13.] kormányrendelet). A minta leválogatását a Nemzeti Egészségbiztosítási Alapkezelő mint a teljes magyar népességet mintegy lefedő nyilvántartó véggezte a 2003-ban TAJ-számmal rendelkezők állományából. Az alapsokaság leválogatása a fent felsorolt többi adatgazda számára is ismeretes kapcsolati kódok (TAJ-szám) listájának létrehozásával kezdődött. Az anonim adatkapcsolásért felelős NISZ Zrt. által kifejezetten az adott összekötéshez generált hash-algoritmus² segítségével, az alapsokaságot leválogató regiszter kezelője az eredeti kódokhoz egyedi, technikai azonosítókat rendelt. A következő lépésben a többi adatközlő is leválogatta az alapsokasághoz kapcsolódó adatait, és *hashelt* állapotban továbbadta azt a NISZ Zrt. számára, amely végül egyesítette és anonimizálta az adatbázist. Így a létrejött adatbázis nem tartalmaz természetes azonosítókat. A kapcsolt államigazgatási paneladatbázisok esetében az egyéni adatok összekapcsolásán (TAJ-szám-sokaságon) kívül a munkáltatói adatok is hasheltek: a foglalkoztatási adatszám segítségével a munkáltatók a Pénzügyminisztérium, az TIM és a MÁK adatait is egyesíti az adatbázis. Ezt

2 A hash-algoritmus egyirányú kódolási gyakorlat, amely a bemeneti adatból a következő feltételek teljesülése mellett képez kimeneti adatot: adott bemeneti információból mindig ugyanazt a kimenetet adja, valamint a kimeneti adat egyértelműen utal a bemeneti adatra, de a kimeneti adatból nem állítható elő a bemeneti adat. Az eljárásban a bemeneti adat legkisebb változása is teljesen más kimenetet eredményez. Hash-módszereket használnak tömörítési, jelszótárolási, keresési eljárásokhoz is. Esetünkben az eljárás az anonim technikai azonosítók létrehozását szolgálja.

az egyesítés után eredeti azonosítók (például a TAJ-szám és a foglalkoztatói adószám) nélkül, ugyanakkor a személyek adataival kapja meg az KRTK Adatbank nyers, kutatásra még alkalmatlan formában. Az adatok átvételét követően kezdődik azok kutatási kérdéseknek megfelelő tisztítása.

Az adatintegráció kényszerűen magában hordozza és tovább örökíti az adatszolgáltatók tartalmi és adatrögzítési hibáit. Az elemzéshez szükséges előzetes tisztítás és adatértelmezés során figyelembe kell venni, hogy az adminisztratív adatoknak célhoz kötött tartalma, terminológiája és struktúrája van, azaz kezdetben az egyes államigazgatási nyilvántartások logikáján belül értelmezhetők. E logika megismerése elengedhetetlen a kutatási célú feldolgozásához és elemzéséhez is.

Az adatösszekötést követően idősoros paneladatbázist készít a KRTK Adatbank, amely a legfrissebb összekötés során több száz nyers mező harmonizálásával jön létre. A szakszerű – majdnem húszéves időszakot átfogó – adattisztítás és harmonizálás hosszú időt vesz igénybe. Éppen ezért a jelenleg összekapcsolt adatbázis tudományos célú elemzésére leghamarabb 2023–2024 folyamán kerülhet sor.

Az adattisztítás

Az adminisztratív adatok itt részletezett tulajdonságai miatt a longitudinális és a keresztmetszeti konzisztenciavizsgálatok egyaránt részét képezik az adattisztítás folyamatának. A keresztmetszeti konzisztenciavizsgálat során a különböző adatkörök tartalmának adott időpontra vonatkozó összevetését végezzük. A felmerülő kérdések megismerésével képet kaphatunk az adatok korlátairól. Ezzel párhuzamosan, az adatok longitudinális ellenőrzésével az évről évre változó adatkörök tartalmát, változásait figyeljük meg, majd rögzítjük metaadatbázisban a teljes megfigyelési időszakra. A tisztítás során a fent részletezett megalapozó vizsgálatokat követően, elemzői döntések nyomán nyerhető ki a kutatási kérdéseknek megfelelő adattartalom az önmagukban csupán adminisztratív jelentésű cellákból.

Az első körös adatbanki tisztítás a beérkezett adatok, tehát az adatszolgáltatók és a NISZ Zrt. munkájának áttekintésével kezdődik. Az átadott adatok ellenőrzése után, a nyers adatmezők iteratív átalakítása során különféle kutatási kérdéseknek megfelelő változók létrehozása következik, először csupán az egyes adatforrások regisztertartalmán belül.

Miután adatforrásonként megtörtént az adatok tisztítása, megismerése és harmonizációja, kezdetét veszi a szűkített, ugyanakkor a legfontosabb változókat összekapcsolva tartalmazó óriás adatbázis felépítése. Az így létrejött adatbázis az Admin4 esetében az egyének 2003 és 2021 közötti, havi szintű státuszait tartalmazza. Az összekapcsolást követően felmerülő adatbázis-inkonzisztenciákat megvizsgáljuk és kezeljük.

Ezáltal létrejön egy közelítőleg tiszta óriás adatbázis. Az összetett forrású változókat felhasználva minikutatások indulnak, annak érdekében, hogy a későbbi adatfelhasználáskor felmerülő potenciális hibákra, hiányosságokra és értelmezési korlátokra fény derüljön. Az ilyen módon felmerülő kérdések és megoldások szintén beépülnek a javított adatbázisba.

Ezt követi a harmonizáció kollektív fázisa, amelynek során a tudományos közösség különböző területeiről érkező szakértők lehetőséget kapnak a Kapcsolt Államigazgatási Paneladatbázis használatára, ezzel együtt a korábbi tapasztalataik tisztítási eljárásba építésére, annak tökéletesítésére. A kollektív harmonizáció fázisában gyűjtött információk (akár megírt programkódok) beépülnek a következő Admin-hullám első körös adatbanki tisztításaiba, organikus módon folyamatosan továbbfejlesztve azt. Hasonló módon, minden későbbi adathasználat során felmerülő kérdés, probléma és visszajelzés, valamint az azokra megírt programkódok is beépülnek az Admin-fájlok tisztító programjába.

Egyéni adatok védelme

A KRTK Adatbank Admin4 adatbázisa a magyar lakosság 50 százalékos véletlen mintáját vizsgálja. Az adatösszekötéskor a NISZ Zrt. ezt a mintát átnevesíti, tehát a TAJ-kódot, adószámot törzítja. Az adatgazdák nem adnak át természetes azonosítókat (név, születési hely és idő, anyja születési és utóneve.

A felfedési kockázat további csökkentése érdekében az összekötést végző szerv (Nemzeti Infokommunikációs Szolgáltató Zrt.) anonimizál is. Ennek során a különlegesnek számító, kritikusan alacsony számú eseteket tartalmazó kategóriákat (például különböző betegségek kódjait, iskolai végzettségeket) összevonja. Az adatbank ezután további kódolásokat, kategóriaösszevonásokat, törléseket végez az adatbázison a felfedési kockázat csökkentése érdekében. Az így kapott adatbázist KRTK Adatbank csupán biztonságos és zárt szerverkörnyezetben engedi kutatni, kizárólag tudományos háttérintézet által megbízott vagy megfelelő és ellenőrizhető tudományos célkitűzéssel rendelkező kutatók és szakdolgozók számára. A részletesebb adatokat igénylő kutatások pedig az adatbanki, kamerával megfigyelt kutatószobában folytathatók, ugyancsak a fenti feltételeknek megfelelően. Az egyéni adatok szakszerű védelme és a tudományetikai kritériumok közösen nyújtanak védelmet a felfedési kockázat ellen.

Az ellenőrizhető adathasználatra különös figyelmet fordító, tudományterületeken átívelő, független, adat alapú társadalomkutatási gyakorlat szükséges a szakpolitikák kidolgozásához, jól működő tudományos és államigazgatási szféra kialakításához, valamint az élet összes területére alkalmazható adateti- kai irányelvek kidolgozásához.

Kapcsolat

A KRTK Adatbankja hazai viszonylatban egyedülállóan szerteágazó témákban gyűjt *survey*- és adminisztratív alapú kutatási adatbázisokat, alakít ki és üzemeltet teljes körű mikroadatok vizsgálatára alkalmas kutatószobát, hoz létre regiszteralapú, adatintegrációs eljárással készült adatbázisokat, szervez STATA-képzéseket, valamint tart fenn kísérleti labort és ahhoz tartozó gépparkot. Általános tájékoztatást az adatbank@krtk.mta.hu email-címen nyújtunk, míg adatkéréssel kapcsolatosan az adatkeres@krtk.mta.hu email-címen lehet érdeklődni.

Hivatkozások

335/2007. (XII. 13.) [kormányrendelet](#) a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról szóló 2007. évi CI. törvény végrehajtásáról.

2007. évi [CI. törvény](#) a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról.

SCHARLE ÁGOTA (2019): [Hungary: a case study on improving access to administrative data in a low-trust environment](#). Megjelent: *Crato, N., –Paruolo, P.* (szerk.): Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design. Springer, 119–130. o.