



ADMINISZTRATÍV ADATBÁZISOK INTEGRÁCIÓJÁVAL KAPCSOLATOS KÉRDÉSEK



SZÉCHENYI 2020



MAGYARORSZÁG
KORMÁNYA

Európai Unió
Európai Szociális
Alap



BEFEKTETÉS A JÖVŐBE

Készült az EFOP-3.4.5-VEKOP-17-2017-00001 azonosítószámú, *Rendszerszintű fejlesztések és hozzáférés bővítését szolgáló ágazati programok a felsőoktatásban* kiemelt projekt DPR Pillér megvalósítása során.

Szerző: Kiss László

Felelős kiadó:

Dr. Gloviczki Zoltán

© Oktatási Hivatal, 2020

Az Oktatási Hivatal kiadványa az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg. A kiadvány adatainak bármilyen formában történő nyilvános közzétételéhez vagy felhasználásához a kiadó előzetes írásbeli hozzájárulása szükséges.

Adminisztratív és kutatási célú adatok, adatkezelés

Az adminisztratív és a kutatási célú adatfelvétel és adatkezelés között komoly módszertani különbségek vannak. Az adminisztratív adat igazgatási céllal gyűjtött adat, jellemzően érvényesül vele kapcsolatban a célhoz kötöttség elve, másodlagos felhasználási lehetősége korlátozott, egyedi engedélyekhez kötött. Az ilyen típusú adatbázisba történő bekerülést jellemzően jogszabály írja elő, a bekerüléshez jogok és kötelezettségek társulhatnak. „A legelterjedtebb meghatározás szerint az adminisztratív adatforrás közfeladatot ellátó szervezet által fenntartott adatgyűjtemény, amit azért gyűjtenek és használnak, hogy eleget tegyenek a közterhekkel, (nyilvántartási, engedélyezési, jogosultsági) eljárásokkal, szolgáltatásokkal, juttatásokkal kapcsolatos feladataiknak. Az adatokat tehát eredetileg jog és kötelezettség keletkeztetésével, az egyénnel szemben valamilyen joghatás kiváltásával összefüggésben gyűjtötték.”¹

Fontos, hogy az adminisztratív célú adatbázist az érintett személyi kört illetően (elméletileg) a teljeskörűség jellemzi, teljes mértékben lefedik (le kell fedniük) a célcsoportot. (Ennek az elvnek a sérülését jól példázza a Közigazgatás megszűnését követően bevezetni szándékozott Kormányzati Személyügyi Szolgáltató Rendszer kiüresedése.)

Az ENSZ 5. statisztikai alapelve kimondja, hogy statisztikai célokra bármilyen adatforrásból – statisztikai adatfelvételekből vagy adminisztratív nyilvántartásokból – származó adatok felhasználhatók. A statisztikai szervezeteknek a forrás kiválasztásánál figyelembe kell venniük az adatok minőségét, időszerűségét, a költségeket és az adatszolgáltatókra háruló terheket.

Az adminisztratív adatbázisok másodlagos célú felhasználását jelenti az esetleges kutatási tevékenységbe történő bevonásuk, kutatás során elvégzett elemzésük. A másodlagos felhasználás ugyanakkor szakpolitikai célokat is szolgálhat, a szakpolitikai döntéshozás támogatásában is fontos lehet.

Az adatvédelmi elvek szem előtt tartása értelemszerűen rendkívül fontos az adatbázisok kezelése, átadása, felhasználása során, ugyanakkor számos technikai jellegű problémát is felvet. A 2003. évi XLVIII. törvény értelmében „személyes adat bármely meghatározott (azonosított vagy azonosítható) természetes személlyel (a továbbiakban: érintett) kapcsolatba hozható adat, az adatból levonható, az érintettre vonatkozó következtetés. A személyes adat az adatkezelés során mindaddig megőrzi e minőségét, amíg kapcsolata az érintettel helyreállítható. A személy különösen akkor tekinthető azonosíthatónak, ha őt - közvetlenül vagy közvetve - név, azonosító jel, illetőleg egy vagy több, fizikai, fiziológiai, mentális, gazdasági, kulturális vagy szociális azonosságára jellemző tényező alapján azonosítani lehet.

Az adminisztratív adatbázisok felhasználhatósága

Az adminisztratív adatbázisok strukturális okokból fakadó adathiánya vagy torzítása a szakirodalom által gyakran említett probléma. Mint egy elemző tanulmány megfogalmazza: „Az adminisztratív adatbázisok mindegyike rendelkezik vakfoltokkal, azaz nem teljes körűen rögzített adatsorokkal. Másrészt nyilvánvaló, hogy a társadalomkutatás számára elengedhetetlenül fontos objektív adatok jó része sem tartozik az adminisztratív nyilvántartások célkörébe – gondolhatunk itt elsősorban a származási adatokra. Más esetben a hiányt az okozza, hogy változószinten elérhető ugyan adminisztratív adatbázisból a kutatás számára szükséges adat, ám tartalmában nem illeszkedik a vizsgálat céljaihoz. Jellemzően ilyen eset a regionális vizsgálathoz elengedhetetlen lakóhelyi adat,

¹ Gárdos Éva: Adatok és kezelésük a hivatalos statisztikában. Educatio, 2015/3. 30.

amely adminisztratív megközelítésben a hivatalosan bejelentett, állandó lakcímet jelenti, míg a kutatás számára a tényleges, életvitelszerű tartózkodás helye lehet informatívabb. Összességében a fő problémát nyilván az jelenti, hogy adminisztratív adatok felhasználása esetén a kutatói érdeklődésnek utólag kell hozzáidomulnia a rendelkezésre álló adatok által kínált információtartalomhoz és nem előre tervezetten és strukturáltan gyűjteni be azt, primer kutatás során.”²

Az adminisztratív adatbázisok statisztikai célú felhasználásának következő erősségei és gyengeségei a legfontosabbak:

- Lefedettség: a célpopulációt a közigazgatás igényei határozzák meg.
- Tartalom: a közigazgatás szempontjából igényelt adatokra szorítkozik.
- Fogalmak, definíciók: a közigazgatás igényei határozzák meg.
- Területi, kisterületi elemzések: előállításuk az egyedi rekordok beazonosíthatóságának függvénye.
- Minőség-ellenőrzés: a közigazgatási hatóság végzi, a kulcsváltozókra koncentrálnak, egyéb változókat figyelmen kívül hagyhat.
- Stabilitás: jogi, eljárásbeli módosulások befolyásolják.³

Az adminisztratív adatok kutatási célú használatával kapcsolatos további probléma a bizonytalan adatminőség; az adminisztratív adatbázisokban lévő torzítások kiszűrésének szinte egyetlen módja a más adatbázisokkal való összevetés.

Adminisztratív adatbázisok 1. – A Magyar Államkincstár Központi Illetményszámfejtő Rendszere

Az adatösszekapcsolás megvalósítása a döntéselőkészítéshez szükséges adatok hozzáférhetőségének biztosításáról szóló 2007. évi CI. törvény hatálya alá tartozik.

A KIR, vagyis a Központosított Illetményszámfejtési Rendszer, a jelenlegi rendszer egy korábbi változata. A KIRA (Központosított Illetményszámfejtési Rendszer Alkalmazás) 2015. november 1-jével került bevezetésre, ez a KIR továbbfejlesztett verziója. Jelenleg az ÁSZ, az NMHH és az Országgyűlés Hivatala kivételével minden költségvetési szerv alanya az illetményszámfejtési rendszernek, kapcsolódásuk kötelező. A kimaradt intézmények központosított rendszerbe vonása is folyamatban van. A HM és a Honvédség is részét képezi az illetményszámfejtési rendszernek, csak más formában. Más, költségvetési kapcsolattal rendelkező szervek is kapcsolódhatnak a központosított rendszerhez.

A kötelezően csatlakozó intézmények száma kb. 13.000 db, az önként csatlakozók száma 300 db alatt van. Jelenleg 1 millió fő bérszámfejtését látják el a központi illetményszámfejtési rendszerben. Ahogy valaki adókötelessé vagy statisztika kötelessé válik, bekerül a rendszerbe. A GYES-en lévők is tartalmazza a rendszer, mint fizetés nélküli jogviszonnyal rendelkezőket.

A KIRA a HR rendszerekből interfészt fogad. A HR rendszer gazdája interfészt fejleszthet a fejlesztőjével. Az adatok azonban csak a KIRÁ-n belül érhetőek el, csak a belső, intézményi felhasználó érheti el a felületet. Ha más felhasználó (pl. HR rendszer fejlesztője) szeretne belépni, akkor azt csak az intézményi jelenléttel teheti meg.

A HR adatoknak (képzettség, végzettség) is van helye a rendszerben, de ezek a kék mezők, vagyis a feltöltésük nem kötelező. (A rendszer fejlesztésekor minden adatkérés esetében meg kellett adni,

² Veroszta Zsuzsanna: Adminisztratív adatok társadalomkutatási kezelése. *Educatio*, 2015/3. 6–7.

³ Gárdos Éva: Adatok és kezelésük a hivatalos statisztikában. i. m. 31.

hogyan az illetményszámfejtéshez szükséges-e, és csak ezek lettek a kötelezően kitöltendő, ún. piros mezők).

Minden szervezettel offline kapcsolatban vannak. Egyedül a MÁK Illetményszámfejtési Főosztálya látja az összes adatot. Szigorú, finomhangolt jogosultsági szinteket rendelnek az egyes intézményekhez, és azon belül az intézményi munkatársakhoz (országos, illetményszám helyi, intézményi jogosultság, illetve intézményi szinten vezetői jogosultság: csak olvasási jog, szakmai menedzser, munkaügyesek, számfejtők, számfejtéssel foglalkozók, gazdálkodási ügyintézők, lekérdezők, rendszergazdai jogosultság).

A KSH adatszolgáltatást, OSAP adatszolgáltatást, 08-as bevallást a kötelezően csatlakozott intézmények esetében a MÁK készíti el és szolgáltatja, míg az önként csatlakozott intézmények esetében átadja az intézménynek, aki továbbítja azokat az adatszolgáltatást kérők felé. A Kincstár aggregált adatokat csak adott területhez kapcsolódóan adhat át más szervnek (pl. költségvetés tervezés, beszámoló során az NGM-nek). Ha külső személy, pl. kutató kér adatot, akkor a kérelmét az adatvédelmi biztos ellenőrzi, és csak az ő jóváhagyásával adhatják át.

1. táblázat A MÁK KIRA-rendszerének főbb adatkörei

Szervi adatok	Államigazgatási szerv típusa
	Gazdálkodó törzsszáma=PIR szám
	Szervezeti egység (kód)
	Szervtípus
Személyes adatok	Személyi azonosító
	Besorolási osztály
	Besorolási fokozat
	Fizetési fokozat
	FEOR-szám
	Heti normál munkaidő
	Iskolai végzettség szintje
	Jogviszonytípus
	Kulcsszám
	Munkaidő
	Részmunkaidő
	Szakképesítés szintje
	Túlmunkaóra (túlóra)
	Végkielégítés
Illetményadatok	Alapilletmény
	Alapilletmény-eltérítés
	Céljuttatás
	Díjazás, illetmény
	Egészségkárosító kockázatokért járó pótlék
	Éjszakai pótlék
	Eltérő illetményalap
	Gépjármű-vezetési pótlék
	Helyettesítési díj
	Idegennyelv-tudási pótlék

	Illetmény
	Illetményalap
	Illetménykiegészítés
	Illetménypótlék
	Integrált ügyfélszolgálati pótlék
	Jegyzői pótlék
	Jubileumi jutalom
	Jutalom
	Képzettségi, munkaköri pótlék
	Nettó jövedelem
	Személyi illetmény
	Vezetői illetménypótlék

A MÁK KIRA rendszerének fontossága témánk szempontjából „saját jogán túl” abban is rejlik, hogy ez a rendszer adja a legfőbb alapját a legtöbb nagy volumenű hazai közszolgálati hr-területet érintő statisztikai adatközlésnek, így a KSH OSAP 2238-nak és az NGM – NFSZ Egyéni bérek és keresetek statisztikája adatfelvételnek is.

A KIR havonta adattáblákat szolgáltat a KSH-nak, amely azt azonos struktúra szerint kiegészíti a nem KIR-t használó költségvetési szervek adataival. A KSH egyedi rekordokkal nem rendelkezik, az adatok sztenderd adattáblákon érkeznek a MÁK-ból, az OSAP statisztikák táblái alapján tehát a sztenderd struktúrától eltérő bontások alkalmazására nincs lehetőség.

Adminisztratív adatbázisok 2. – A Nemzetgazdasági Minisztérium adatbázisa, a „bérstatisztika”

Az adatfelvétel az 1668 OSAP felvétel alapján történik. Ebben az egyéni bérek és keresetek adatait gyűjtik. Ez az adatgyűjtés már 30 éve fut, 2002 óta teljesen azonos módszertan szerint történik az adatok gyűjtése, az adatok ugyanolyan módon állnak rendelkezésre. A közigazgatásra vonatkozóan teljes lefedettséget biztosítanak az adatok, ezek 99,9%-át már a MÁK-tól kapják, annak illetmény és kereseti adatait veszik át a KIRÁ-ból.

Évente egyszer, májusban történik az adatkérés, de az adatgyűjtés módszertana biztosítja, hogy nem az adott hónap adatait kapják meg, hanem a teljes előző évre számított adatot. Az állománylétszám az adott évben meglévő létszám. A teljes közigazgatási rendszert lefedi az adatgyűjtés: közalkalmazottak, bírák, ügyészek, rendvédelmi szervek is szerepelnek benne, illetve az OH Mt-seit is tartalmazza. Közel 700 ezer fő kerül be az adatgyűjtésbe a közigazgatás részéről. A vállalkozások adatainak gyűjtésére mintavétellel kerül sor. Vállalkozások esetében 17 ezer szervezetre gyűjtene adatokat, 170 ezer foglalkoztatottakra. Az 50 fő felettiéknél kötelező az adatszolgáltatás, 50 fő alattiakra mintavételt vesznek, 20%-ot. Az egyénekenkénti adatok rendelkezésre állnak.

Az MNB, valamint kutatók felé is történik adatszolgáltatás. Kutatóknak együttműködési megállapodásuk van, más kutatóknak csak akkor adják oda, ha az NGM engedélyt ad rá. Emellett a megyéknek, munkaügyi központoknak is el kell küldeni, ezt használják az átlagkereset kiszámításához. Azért az NGM adja át ezeket az adatokat és nem a MÁK, mert ők feldolgozzák fel, a MÁK nem.

Adminisztratív adatbázisok 3. – A NAV adatai

Az adatkérőkkel mindig lényegre törően kell meghatározni, hogy milyen adatokra van szükségük. Kutatási célra jól meg kell találni a célt, és azután kérni az adatokat. A kutatókkal mindig megbeszéljük, hogy mik legyenek benne. A lekérdezendő időszak mindig kérdéses. Általában az április-május hónapokat szokták kijelölni, mint referenciahónapot.

Az egyszer már lekérdezett adatbázis közadattá válik, és onnantól bárki felhasználhatja. Nagyon fontos arra figyelni, hogy ne legyen visszakereshető az adat, tehát addig kell butítani, amíg nem lesz felismerhető.

Viszont: a lekérdezett adatbázisokat nem tárolják el, nincs rá kapacitásuk – csak akkor kapja meg más valaki is az eredeti kérvényezőtől kívül, ha ugyanakkor adja be az igénylést rá.

A NAV-nál a beérkező adatokat tárolják. Az önellenőrzésig minden adatot kezelnek. Itt javítás nem történik. Adattisztítást nem végeznek, azt az adózónak kell megtennie.

2. táblázat A NAV 08-as bevallásának adatkörei

Bevalló cég adatai	Adózó cég adószáma
	Adózó cég neve
	Székhely
	Levelezési cím
	A bevallás jellege
	A bevallás típusa
	A bevallás fajtája
	Kormányhivatal bevallása-e
Az egyes magán-személyek adatai	A biztosítási jogviszony időtartama
	Az ellátás folyósításának időtartama
	Az „alkalmazás minősége”
	Munkakör, foglalkozás
	Részmunkaidős, teljes foglalkoztatás, munkaidő-szervezési intézkedéssel csökkentett teljes munkaidő
	Biztosítás szünetelése, vagy munkabérrel ellátatlanság ideje, kódja
	Korkedvezményre jogosító munkakör
	Az egészségügyi dolgozók heti 48 órát meghaladó önként vállalt többletmunkavégzésének ideje
	Az arányos szolgálati idő naptári napjainak száma
	A biztosítási jogviszony megszűnésének időpontja
	gyermekgondozási díj/a gyermekgondozási segély folyósítása melletti munkavégzés időtartama
	Az egészségügyi szolgáltatási járulék NFA által átvállalt összege/Az egészségügyi szolgáltatási járulék összege
	A korkedvezmény-biztosítási járulék alapja
	a korkedvezmény-biztosítási járulék összege
	A felszolgálati díj után fizetett nyugdíjjárulék alapja

	a felszolgálati díj után fizetett nyugdíjjárulék összege
	A családi járulékkedvezmény alapja
	A családi járulékkedvezmény összege
	A társas vállalkozó személyes közreműködői díja
	Családi járulékkedvezményből érvényesíthető összeg
	Az egészségbiztosítási- és munkaerő-piaci járulék alapja
	levont (befizetett) természetbeni egészségbiztosítási járulék rész összege (4%) és/vagy a levonás elmaradásának oka
	A természetbeni egészségbiztosítási járulék terhére érvényesített családi járulékkedvezmény összege
	A fizetendő természetbeni egészségbiztosítási járulék rész összege
	Az egészségbiztosítási- és munkaerő-piaci járulék alapja
	A pénzbeli egészségbiztosítási járulék terhére érvényesített családi járulékkedvezmény összege
	A fizetendő pénzbeli egészségbiztosítási járulék rész összege
	Az egészségbiztosítási- és munkaerő-piaci járulék alapja
	A nyugdíjjárulék alapját képező jövedelem összege
	A nyugdíjjárulék-köteles jutalom összege
	A Nyugdíjfolyósító Igazgatóság által folyósított ellátásban részesülő személy által fizetendő nyugdíjjárulék alapja
	A nyugdíjjárulék terhére érvényesített családi járulékkedvezmény összege
	A fizetendő nyugdíjjárulék összege
	a gyermekgondozási díjra, gyermekgondozási segélyre vonatkozó adatok
	A GYET, ápolási díj, munka-rehabilitációs díj folyósításának, valamint az álláskeresési támogatás összege
	A nyugdíjjárulék terhére érvényesített családi járulékkedvezmény összege
	A fizetendő nyugdíjjárulék összege
	A biztosítási jogviszony megszűnését követő időre folyósított csecsemőgondozási díj, baleseti TP összege
	Teljes havi jövedelem (a részkedvezmény kiszámításához)
	A fizetendő szociális hozzájárulási adó összege
	Az adókedvezmény jogcíme
	A közfoglalkoztatás igénybevételének időtartama
	A közfoglalkoztatás keretében alkalmazott személy részmunkaidőben történő foglalkoztatása utáni tételes egészségügyi hozzájárulás számított összege
	A kedvezmény figyelembevétele nélkül számított szociális hozzájárulási adó alapja/összege
	A Rehabilitációs kártyával rendelkező személyek foglalkoztatásához kapcsolódó kedvezmény igénybevételének időtartama
	A számított szociális hozzájárulási adó alapja/összege
	A fizetendő szociális hozzájárulási adó összege
	A szakképzettséget nem igénylő munkakörben foglalkoztatott munkavállalóhoz kapcsolódó kedvezmény igénybevételének időtartama
	A kedvezmény figyelembevétele nélkül számított szociális hozzájárulási adó alapja/összege
	A legfeljebb 180 nap munkaviszonnyal rendelkező, 25 év alatti foglalkoztatott munkavállalóhoz kapcsolódó 27%-os mértékű kedvezmény igénybevételének időtartama
	A 25 év alatti foglalkoztatott munkavállalóhoz kapcsolódó kedvezmény igénybevételének időtartama
	Az 55 év feletti foglalkoztatott munkavállalóhoz kapcsolódó kedvezmény igénybevételének időtartama

	A tartósan álláskereső személyek foglalkoztatásához kapcsolódó kedvezmény igénybevételének időtartama
	A számított szociális hozzájárulási adó alapja/összege
	A GYED-ben, GYES-ben vagy GYET-ben részesülő munkavállaló kapcsolódó kedvezmény igénybevételének időtartama
	A számított szociális hozzájárulási adó alapja/összege
	A doktori (PhD) vagy ennél magasabb tudományos fokozattal, vagy tudományos címmel rendelkező kutatók, fejlesztők foglalkoztatásához kapcsolódó kedvezmény igénybevételének időtartama
	A számított szociális hozzájárulási adó alapja/összege
	A nemzeti felsőoktatásról szóló törvény szerinti doktori képzésben részt vevő hallgatók vagy doktorjelöltek foglalkoztatásához kapcsolódó kedvezmény igénybevételének időtartama

A Központi Statisztikai Hivatal adatai

2013-ban kifejlesztették a KARÁT rendszert, amely dokumentálja a KSH adatátvételeit, biztonságos csatornát nyújt az adatátvitelre és a partnerek felé is egyszerűbb folyamatot biztosít.

Megfelelő dokumentációt kellene biztosítani az adatátvitelhez. Ehhez pontosan látni kellene az adatszolgáltató rendszertervét, illetve a metaadat leírásokat (az egyes mezők tartalmát, a mögöttes lévő fogalmakat). A metaadat leírásokat nem kötelező átadni a partner adatszolgáltatóknak.

Új statisztikai törvény készült, amely lehetővé teszi, hogy a KSH átvesse az adminisztratív nyilvántartások adatait. Amennyiben egy nyilvántartásban módosítás történik, azt dokumentáltan, KSH-val egyeztetve kell végrehajtani. Ez a gyakorlatban azonban még nem valósult meg teljeskörűen. A partnereknél rendelkezésre álló metaadat leírásokat át kellene vennie a KSH-nak.

A KARÁT csak azt biztosítja, hogy az adat formája megfelelő legyen, fájl formátumban és adatbázisba is betöltődjön. Az ADAMES adatelőkészítő rendszer biztosítja, hogy a logikai összefüggések megfelelőek legyenek. A végeredmény az, hogy olyan formátumban álljon elő az adat, hogy azt fel tudják dolgozni.

Az adatszolgáltatások különbözőek: van, ahol javít a KSH, de van, amit nem javít (pl. címek esetében az ADAMES rendszerben lehet tisztítani.) A kódlisták átkódolása is fontos feladat. Összefüggés ellenőrzések is történnek.

A KSH által használt kódokat néhány esetben az egész közigazgatás alkalmazza (pl. TEÁOR, FEOR). Sok esetben azonban nem frissítik a kódokat az intézmények.

A Magyar Tudományos Akadémia adatintegrációs tevékenysége, az Államigazgatási Adatgyűjtemény

Az államigazgatási adatgyűjtemény az OEP-ONYF-NAV-NMH-OH1 – adminisztratív - adatfájlok összekapcsolása által jött létre. Az állomány mintája a megfigyelt időszak kezdetén, 2003-ban 5-74 éves népesség fele. Az állományok a mintába bekerült személyekről demográfiai adatokat, valamint ellátásra és jogviszonyra vonatkozó információkat tartalmaznak.

A kutatás során készült szakmai interjú főbb meglátásai:

Kevés olyan adatbázis van, ahol időben követhetők az egyének a munkaerőpiacon. Vannak kapcsolt adatbázisok, vállalati szinten, de ezek alapján az egyén csak nehezen és nagy hibaszázalékkal követhető. Az MTA adatintegrációs adatbázisa esetében TAJ-szám alapján készül a minta, az adott korú népesség 50%-a lehet benne. A törvény tartalmazza, hogy 50%-osnál nagyobb mintát nem lehet venni. (Svédországban pl. teljes körű az integráció.) A NISZ hash-kódokat oszt ki az adatgazdáknak, azok a megfelelő változót meghashelik, és visszaadják a NISZ-nek, aki elvégzi az összekötést. A keletkező file alapja valamilyen jogviszony. Eleje dátum, vége dátum, és kérdés, mi van a kettő között. Az alapegység tehát nem az egyén, hanem a jogviszony, holott a legtöbb esetben arra van szükség, hogy egyénsoros legyen az adatbázis, tehát az egyén életútjának megfigyelésére legyen alkalmas.

Havi felbontású adatbázist készítenek, egy egyén egy bizonyos hónapban mit csinált. (dolgozott, munkanélküli volt, hol dolgozott, mennyit keresett, kap-e nyugdíjat, transzfert, mennyi gyógyszert fogyasztott, volt-e kórházban, ki a munkáltatója – ehhez csatolt adatbázisban ott van a vállalat pénzügyi adatbázisa stb.) A havi felbontású file is nagyon nagy, de a létrehozásához létre kell hozni egy napi felbontású file-t, ebből lehet kiszámolni, hogy valaki az adott hónapban mennyit dolgozott összesen, mennyit keresett az egyes indentyisaiban összesen is (!). Van, amit úgy mérnek, hogy mit csinált a személy a hónap 15-én, van, amit az adott hónapban aggregáltak.

A munkaviszonyokat a vállalatok évente jelentik. Igaz, hogy ez le van bontva dátumokra. De ha a vállalat valamiért nem jelent, akkor az adott munkavállalóknak egy teljes év hiányzik. Vannak benne eltérő értékek, az egyik vállalkozó ezer Ft-ban jelent, a másik nem, ezeket tisztítani kell. Amit az adatgazdák elfelejtenek: változás van az új évben lekérdezett adatokban, fogalmi változás van. Az adatgazdák jellemzően nem rendelkeznek historikus metaadatbázissal, össze kell kaparászniuk a változásokat.

Tipikus hibák: OEP, MÁK – elkezdődik egy jogviszony, és nincs lezárva. (Pl. GYES 30-35%-a) Ok: máshogy zárják le, a kiinduló dátumot tartalmazó adat mellé nem rendelik oda a lezáró dátumot. FEOR: 1993-ban volt benne változás, aztán 1997-ben. Van, ami ezekből harmonizálható, és van, ami nem. A legutóbbi nem ilyen, két kódból néhányan átmentek egy harmadikba, ezt nem lehet lekövetni. 2 számjegyen már lehet. (Az elemzések legnagyobb része a végére marad az 1 számjegynél.)

Kérdések és felvetések

Adattisztítás és adatstruktúra

- ember- és időigényes feladat; hogyan lehet optimalizálni?
- mire terjedjen ki az adattisztítás? mit tekintünk az adattisztítás, és mit az adatkezelés/adatfelhasználás részének?
- adathiányok kezelése – hogyan történjen? valószínűségi összekötés vagy üresen hagyás a biztosabb? hosszú idősorok kialakítása és napi szintre bontott adatok készítése – van rá erőforrás?
- hogyan kerüljük el, hogy addig tisztítunk, míg már elkészül az újabb verzió, amit szintén tisztítani kellene?
- mikortól tekinthető használatra késznek az adatbázis?
- milyen lehetőségek vannak adattisztító munkacsoport létrehozására? külső partnerrel történő együttműködés lehetséges?
- pilot- vagy „mini” kutatások a felhasználhatóság javítása érdekében – lehetségesek? lehet ezeknek egyéb haszna is az adattisztításon túl?

- felhasználói visszajelzések, scriptek beépítése lehetséges? (persze, ehhez kellene felhasználók)
- hogy épüljön fel az adatstruktúra? minta: magadatbázis + csatolt adatbázisok; megfelelő lehet?
- hogyan oldható meg az eredeti, illetve munkaadatbázisok folyamatos rendelkezésre állása?

Diplomás pályakövetés adatintegrációs adatbázis pozicionálása

- ágazati háttérelmézések vagy kutatói hozzáférés és felhasználás?
- ki és milyen mélységben férhessen hozzá az adatbázishoz?
- ki dönt a kutatási célú felhasználás engedélyezéséről, ki dönt arról, hogy kutatásnak minősül-e egy adatkérés?
- hogyan történjen a kiszolgálás? mire van forrás, mire van lehetőség?