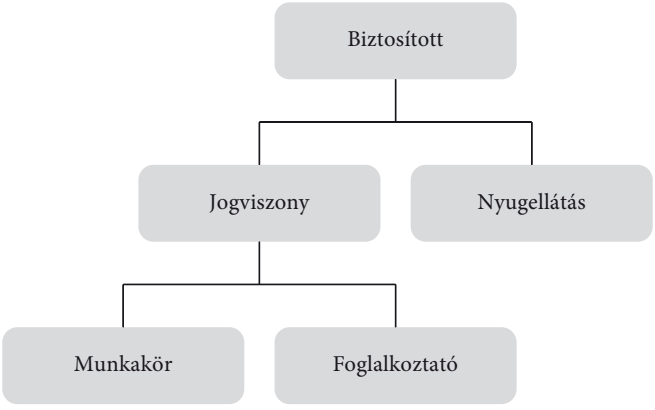
KÖZGAZDASÁGI SZEMLE, LXII. ÉVF., 2015. DECEMBER (1359–1366. o.)

Mikroszimulációs nyugdíjmodellezés adattárház támogatásával

Az Országos Nyugdíjbiztosítási Főigazgatóság újonnan létrehozott adattárháza[1] a nyugdíjbiztosítás jogosultsági adatbázisaiból egyesít statisztikailag fontos információkat, amelyeket az anonim statisztikai és mikroszimulációs elemzések elvégzéséhez elérhetővé tesz a nyugdíjmodellezési igények alapján kialakított adatpiacon *(1. ábra)*.

1. ábra
Az adatpiac egyszerűsített szerkezete



Az adatstruktúrában nem beazonosíthatóan egyéni szinten szerepel minden olyan biztosított, akinek létezik legalább egy elektronikusan rögzített jogviszonya a nyugdíjbiztosítási nyilvántartásban. Ez utóbbi tartalmazza az elérhető teljes magyarországi életpálya munkaviszonyait és egyéb jogszerzéseit. A munkaviszonyok kiegészítő adatai a foglalkoztatók és a munkakörök listája. Amennyiben a biztosított ellátott, a

[1] Az Országos Nyugdíjbiztosítási Főigazgatóságnál 2015 közepén zárult le az uniós támogatásból megvalósult VS/2013/0132 számú projekt. Ennek részeként európai összehasonlításban is korszerű módon, adattárház-alapú fejlesztés biztosítja a MIDAS_HU mikroszimulációs előrejelzésének bázisadatait.

A modellezést lehetővé tevő adattárházat az Omnit Solutions szakemberei készítették a NAS Kft. nyugdíj-nyilvántartási rendszer szakértőinek közreműködésével.

A kézirat első változata 2015. július 22-én érkezett szerkesztőségünkbe.
DOI: http://dx.doi.org/10.18414/KSZ.2015.12.1359

hozzákapcsolódó nyugellátási adatokat is elérhetők az adattárházban. A végeredmény nemcsak az adatelemzést könnyíti meg a nyugdíjszakma számára, hanem biztosítja, hogy az előre jelzési modell mindig friss adatokkal működhessen.

A modellezés alapjául szolgáló adathalmazt korábban az informatikai szakterület egyedi lekérdezések és beszámolók formájában támogatta. 2013-ban a komplex adatrendszer kialakítására elkészült a személyes adatoktól megfosztott jogviszony–biztosítotti adatbázis, azonban az állomány folyamatos frissítése továbbra is nehézkes maradt. Így merült fel az igény egy adattárház kialakítására, amelynek az igényspecifikációja alapjául az egyszeri adatkimentés szolgált. Az eredeti adatmodellt több ponton finomították, továbbfejlesztették, valamint felmerült az automatikus, informatikai szakemberek nélkül végezhető adatfrissítés igénye is.

Forrásrendszerek

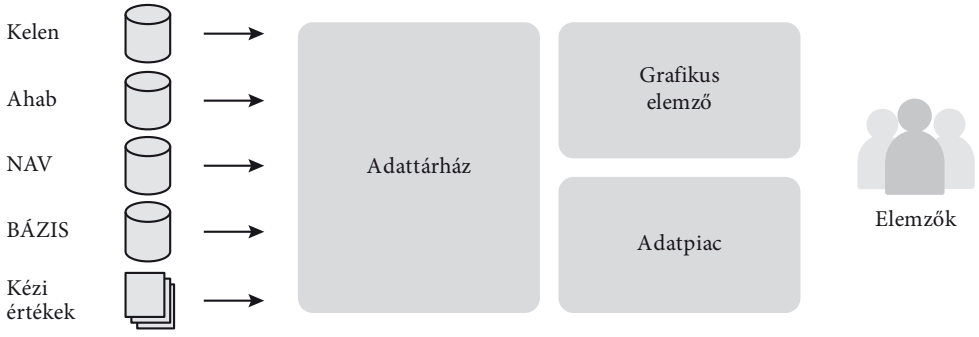
A mikroszimulációs adattárház a szükséges relációs táblázatokhoz közvetlenül a forrásrendszerek adatbázisából nyeri az adatokat olvasási hozzáféréssel. Első lépésként a forrásrendszerekből a szimulációs adatmodell előállításához szükséges összes táblázatot teljes aktuális tartalmával átmentik az adattárház feldolgozási területére.

A *biztosítotti* rekordokat két forrásból, a hiteles biztosítotti adatbázisból (ahab) és a Központi Elektronikus Nyilvántartásból (Kelen) töltik át. 1997-től közel teljesnek tekinthetők az elektronikusan tárolt magyarországi jogviszonyok. Ezek 2010-ig szintén a Kelenből kerültek be az adattárházba, 2010-től pedig a Nemzeti Adó- és Vámhivataltól érkeznek elektronikus formában.

A *folyósítási adatok* egy részét a Nyugdíjfolyósító Igazgatóság BÁZIS rendszeréből töltik be, melynek segítségével a biztosítottak teljes életútja végigkövethető lesz. A töltések során felhasznált kézzel összeállított paraméterállományok lehetővé teszik a kalkulációk finomhangolását és bizonyos törvényi változások követését.

A különböző forrásrendszerek közötti adatszinergiához a betöltés előtt meg kellett oldani az azonos személyhez tartozó biztosítotti és jogviszonyadatok összefogását. Ennek köszönhetően áll elő az elemzésre alkalmas komplex adathalmaz *(2. ábra).*

2. ábra
A komplex adathalmaz előállítása



Előkészítés

Az adattárház töltési, valamint kalkulációs folyamatainak kialakítását egy hónapos előkészítő munka előzte meg. A nyugdíjszakma és az informatikus mérnökök a felmérési fázis során pontosították a fogalmakat, számítási algoritmusokat, valamint az adatkörök elérhetőségeit a forrásrendszerekben. A következő főbb kalkulációs algoritmusokat határozták meg.

A SZOLGÁLATI NAPOK KISZÁMÍTÁSI ALGORITMUSA • A biztosított összes jogviszonyát egyben vizsgálva az egy adott hónapra jutó szakaszokat egyesíteni kell. Ha több jogviszony átfedésben van egy adott napon, akkor is csak egy szolgálati napnak számít az a nap. Az így képzett összesítésből el kell hagyni azokat a napokat, amelyek alkalmazásminőségük kódja szerint kieső idők (például fizetés nélküli betegszabadság). Ezek az ellátatlansági napok és típusok a jogviszonyoktól elkülönülve kinyerhetők a forrásrendszerek adataiból. Így végül éves bontásban előállnak a biztosítottak hónapokra jutó szolgálati napjai, amelyeket ki kell mutatni jogviszonyonként és összesen is.

A SZOLGÁLATI NAP ARÁNYOSÍTÁSÁNAK ALGORITMUSA • A szolgálati napok meghatározása után meg kell vizsgálni, hogy szükséges-e arányosítást alkalmazni. Három feltétel esetén van erre szükség:

– 1996. december 31. utáni időszakról van szó,
– nem teljes munkaidős a jogviszony,
– a jogviszonyhoz tartozó nyugdíjjárulék alapja kisebb az időszakra érvényes minimálbérnél. (Ha a nyugdíjbiztosítási járulékalap nem kisebb a minimálbérnél, akkor a nyugdíjjárulék alapja sem lehet kisebb.)

Az arányosítás úgy történik, hogy az eredetileg megkapott napok számát meg kell szorozni az időszakra jutó nyugdíjjárulék-alap és a minimálbér hányadosával.

AZ OSZTÓNAPSZÁMÍTÁS ALGORITMUSA • Osztónap, azaz jövedelemszerző nap, amely – 0-nál nagyobb arányban – szolgálati idő. Kivételek azok a napok, amelyeken a biztosítottnak bármelyik jogviszonyában van erre a napra vonatkozó olyan ellátatlansági ideje, amely speciális alkalmazás minőségének kódjával rendelkezik.[2] Ez utóbbi esetben az adott nap 0 mértékben számít osztónapnak.

Az osztónap mindig teljes nap, vagyis ha a megfelelő szolgálatiidő-nap nagyobb, mint 0, és a nap nincs kizárva az osztónapok közül, akkor arra a napra az osztónap szorzója 1. (Így előfordulhat, hogy több osztónap lesz, mind szolgálati időben töltött nap.)

A FŐ JOGVISZONY MEGHATÁROZÁSÁNAK ALGORITMUSA • A fő jogviszony meghatározása azt jelenti, hogy egy biztosított hány napot töltött fő jogviszonyként egy adott jogviszonyban. Ha nem volt párhuzamosan másik jogviszony, akkor

[2] 11: táppénz, 12: baleseti táppénz, 21: terhességi-gyermekágyi segély, 69: fizetés nélküli szabadság gyermekápolás, gondozás miatt.

értelemszerűen a fő jogviszonyban töltött napok száma megegyezik a jogviszonyban töltött napok számával. Ha létezik egy vagy több párhuzamos jogviszony, egymás mellett kell vizsgálni a jogviszonyokat, hogy a biztosított melyikben hány napot töltött fő jogviszonyként.

A jogviszonyok rangsorolására felállított komplex szabály figyelembe veszi:

– hogy a jogviszony teljes vagy részmunkaidős,
– a jogviszony időtartamát és kezdetét,
– egy napra jutó jövedelmet,
– pszeudó/valódi státusát: a jogviszonyok az alkalmazás minősége szerint két csoportra oszlanak: valódi és pszeudó jogviszonyokra. Összefoglalóan valódi jogviszonynak azt nevezzük, amelyben a nyugdíjbiztosítási járulék alapja tényleges munkavégzésből származik, minden egyéb jogviszony pszeudónak minősül.

Az adattárház megvalósítása

Az adattárház azt a klasszikus töltési módszert valósítja meg, amely először a forrásrendszerekből *kinyeri* az adatokat, *betölti* azokat a feldolgozási területre, majd elvégzi a *transzformációkat*, amely alapján előáll az elemzésre alkalmas adattárház és végül az adatpiac *(3. ábra).*

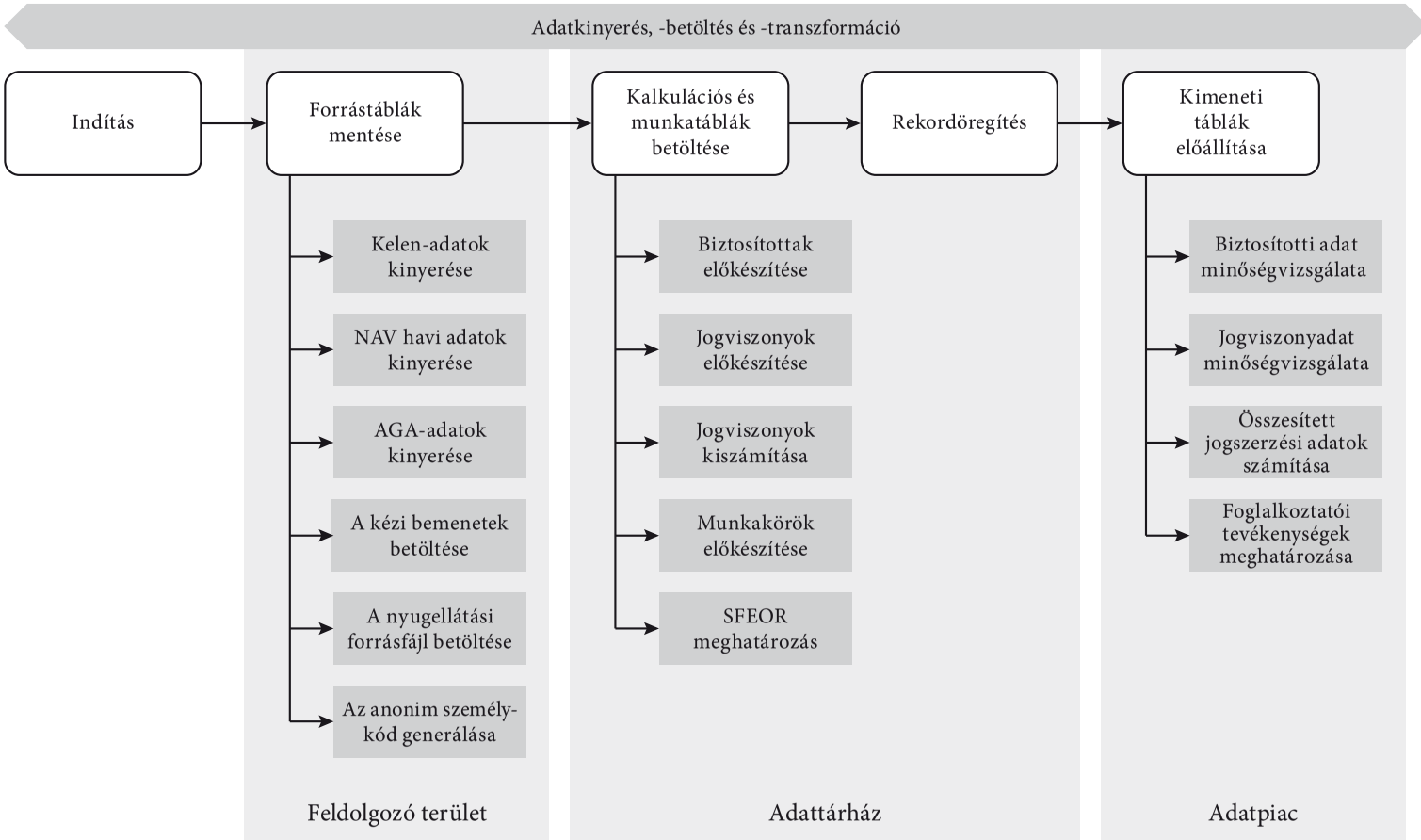
A felhasznált forrástáblák teljes tartalmát anélkül töltik át, hogy megvizsgálnák, melyek az új vagy módosult rekordok a forrásrendszerekben. A töltés során az aktuális teljes forrásadathalmazból előáll a munkatáblákban az elemzési modell aktuális állapota. Ezeket az eredménytáblákat összehasonlítják az adattárházban már korábban tárolt rekordokkal – forrásoldali egyedi kulcsok alapján párosítva. Az összehasonlításnak három kimenete lehet:

1. azokat a rekordokat, amelyek szerepelnek a munkatáblákban és nem találhatók meg az adattárházban, új rekordokként be kell szúrni;
2. azok a rekordok, amelyek szerepelnek a munkatáblában, valamint az adattárházban, és nem változtak, nem kerülnek tovább töltésre, megmaradnak eredeti formájukban az adattárházban;
3. azokat a munkatáblában és az adattárházban szerepelők rekordokat, amelyek módosultak, 2-es típusú SCD rekordöregítéssel tárolják.[3]

A teljes aktuális állapot előállítása és külön a változások tárolása ugyan valamivel több feldolgozási időbe kerül, mintha mindig csak a változások készülnének el, de több előnye is van. Az ősfeltöltés ugyanaz a folyamat, ami később a frissítéseket is végzi, hiszen a feltöltésnél előáll a komplett állapot, és mivel ősfeltöltés előtt az adattárház még üres, így minden rekord újnak számít, és bekerül a tárházba. Megbízhatóbb is, mivel a változásvizsgálat ebben az esetben az adattárházban történik, nem a

[3] Az SCD *(Slowly Changing Dimensions)* az az eljárás, amelynek segítségével nyomon követhetők a rekordok – tipikusan a dimenziótáblák – változásai az adattárházon belül. A 2-es módszer lényege, hogy egy rekord változása esetén létrejön annak egy újabb verziója, míg az eredeti sor nem kerül felülírásra, hanem lezárt érvényességgel megmarad a táblában.

3. ábra
Az adattárház töltési folyamata



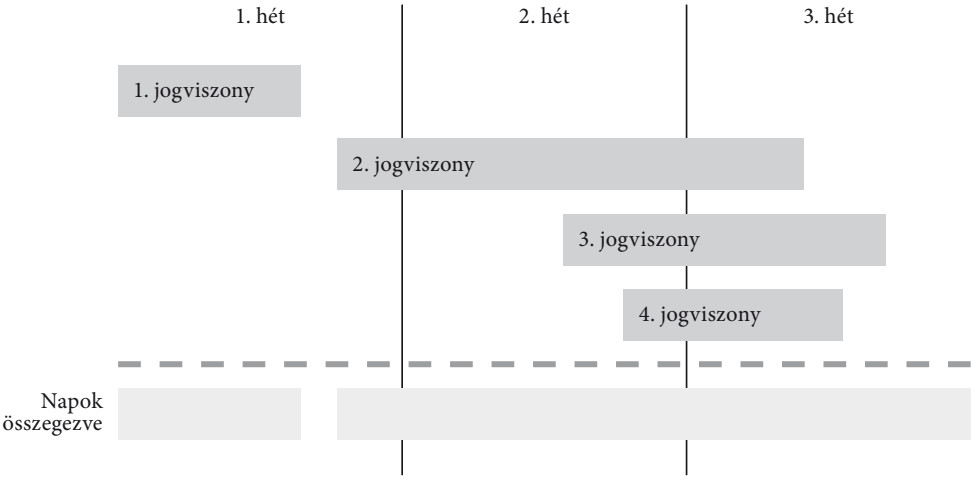
forrásoldalon ahol a historizálást (Δ-képzést) sokszor csak összetett vizsgálattal lehet megoldani. Szintén előnyös, hogy visszakövetés esetén a teljes adattartalom előállítása visszakövethető az egyes fázisokban elkészülő munkatáblákon keresztül a forrásrendszeri kiinduló táblákig.

Az eredménytáblák előállítása során a legnagyobb kihívást a félmilliárd rekordon elvégzett elemzésre előkészítő kalkulációk jelentették. A kezdeti próbálkozások során becsült több hónapos számítási időt sikerült fél napra csökkenteni, így a teljes töltés az adatok forrásrendszeri kinyerésével együtt csak néhány napot vett igénybe. Az optimalizáció többek között az adatbázis beállításainak finomhangolásából, táblastruktúra-átszervezésből, adatindexelésekből, feldolgozási egységek szétosztásából és párhuzamosításából állt.

A számításokat bonyolítja, hogy az egy biztosítotthoz tartozó jogviszonyokat összességében kell vizsgálni, viszont ezek az egyes időszakokban átfedhetik egymást. Az átfedő eseteket bizonyos szempontok alapján összegezni kell (például a hónapon belül megállapított szolgálati napok esetén), vagy ki kell választani valamilyen rangsorolás szerint a legnagyobbat a fő jogviszonyban töltött napok kiszámításához.

A *4. ábra* bemutat egy példát az átfedő jogviszonyaira (egy jogviszony 1 rekord az adatbázisban, melynek az attribútumai közül egyik a jogviszony kezdésének időpontja, egy másik pedig a jogviszony végének időpontja).

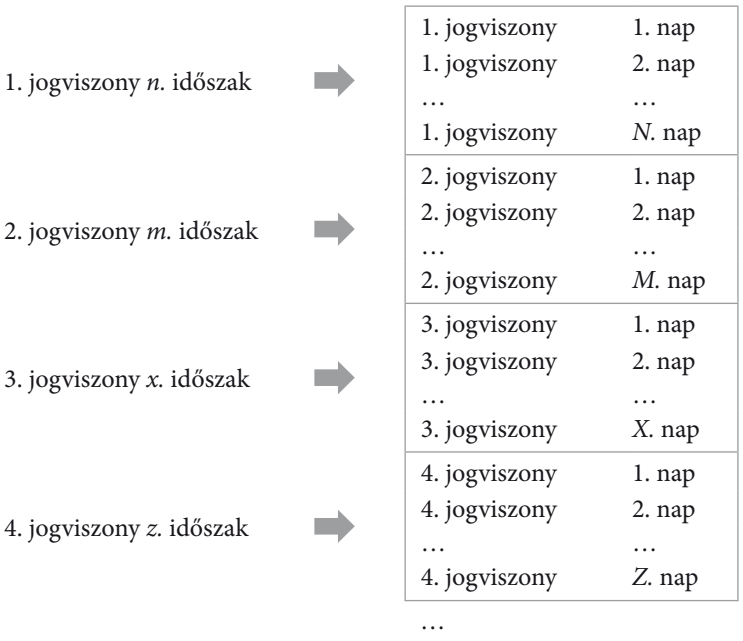
4. ábra
Példa átfedő jogviszonyra



Szekvenciális programozási nyelvvel kezelve az összegzés könnyedén megoldható tömbök segítségével, de relációs adatbázisokhoz használt strukturált lekérdező nyelvvel *(Structured Query Language, SQL)* már nem triviális feladat az átfedő időszakokat tartalmazó rekordokat összehasonlítani, illetve összegezni.

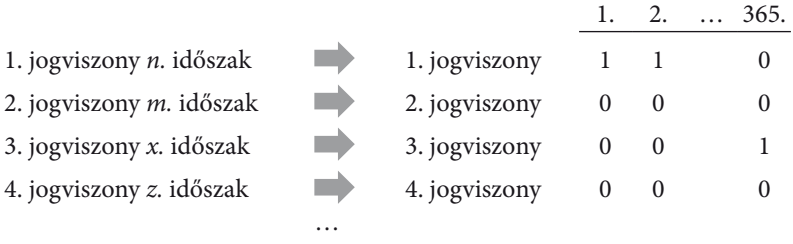
Az összehasonlítás elvégzésére az első megközelítés, hogy a jogviszonyokat tovább bontjuk napokra. Így előáll egy új tábla, amelyben egy jogviszony időtartamának összes napjára hozzáadódik egy dátumoszlop *(5. ábra).*

5. ábra
A jogviszonyok napokra bontása



Így naponként már könnyedén lehetséges az összegzés és az összehasonlítás aggregációs műveletek segítségével.[4] A végső megoldást hasonló elgondolás alapján alakították ki, de ennél nem új rekordok készültek naponként, hanem a jogviszony tárolt adatai mellé az év összes napjára 365 virtuális oszlop került feltöltésre az attól függő értékkel, hogy a jogviszony időszakába beleesik-e az oszlop által jelölt nap. Egy jogviszonyrekord mindig egy tárgyéven belüli időszakot jelöl, ezért lehetséges az év napjait oszlopként feltüntetni a jogviszonyok mellett (a szökőévet külön le kellett kezelni) *(6. ábra)*.

6. ábra
A jogviszonyok naptári évenként



Az így előállt táblában az átfedő jogviszonyokat alapaggregációs függvényekkel össze lehetett hasonlítani, és (mindössze 10 óra alatt) elvégezni a szükséges számításokat.

[4] Az ezzel a módszerrel kialakított optimális megoldás is két hét alatt tudta volna elvégezni a teljes adathalmazra a rekordok napokra bontását és összegzését, így szükségessé vált az algoritmus újragondolása.

Ellenőrzések

Az elkészült adatbázis ellenőrzését két lépésben végezték el.

1. A nyugdíjszakma és az informatikusok közösen meghatározták a teljes adathalmazt lefedő jogviszony- és biztosítotti típusokat. Az egyes típusokra kigyűjtött példákat összehasonlították az analitikus rendszerekben található forrásadatokkal, végignézve, hogy megegyeznek-e, illetve hogy a számított mezők helyesek-e.
2. A tételes ellenőrzést a teljes halmazra meghatározott összesítő számok összevetése követte a forrásrendszerek azonos mutatóival.

Továbbfejlesztési lehetőség

Az adattárház töltési rendszerét és táblaszerkezetét modulárisan alakítják ki, ami lehetővé teszi a kalkulációk egyszerű módosítását. Erre szükség lehet például olyan jogszabályváltozás esetén, amely a meglévő paraméterezési lehetőséggel nem követhető le. Ugyanígy a forrásrendszerek esetleges változásai esetén is (ide értve akár a forrásanalitika lecserélését is) elég a megváltozott bemenő adatokat a megfelelő integrált formára hozni, a számítások és az adattárház kimenetét, valamint az elemzési adatpiacot nem szükséges áttervezni.

Az adattárházak egyik nagy előnye, hogy a meglévő adatok és folyamatok módosítása nélkül lehetőség van az adatmodell bővítésére akár teljesen új adatkörök bevonásával. Az elkészült elemzési eszköz lehetővé teszi az egyéni szintű életpálya-követést és statisztikai elemzést, ami a nyugdíjszakma számára nagy előrelépés. Újabb információterületek bevonásával az adattárházba megvalósulhatnak olyan elemzések és kimutatások is, amelyek elkészítésére jelenleg kevés lehetőség adott.

*

A megvalósított rendszer az automatikus töltési folyamat során a forrásadatokat egységes formában, egy külön szerveren anélkül teszi elérhetővé, hogy az elemzők az adatgyűjtésekkel és a modellezéssel megterhelnék a forrásrendszereket. Az igényeknek megfelelően az adattárház biztosítja a visszatekintés lehetőségét, azaz egy korábbi állapot bármikor előidézhető akkor is, ha már frissebb adatokat is betöltöttek. Nagyon fontos ellenőrzési lehetőség, hogy az adattárházban tárolt rekordok a forrásanalitikáig visszakövethetők. Több, az adatok előkészítéséhez használt paraméter az elemzők által megadható, ezzel biztosítva a jogszabályváltozások követéséhez nélkülözhetetlen rugalmasságot.

A kialakítás során fontos szempont volt, hogy a modell bővíthető legyen. További adatkörök bevonásával a jövőben lehetőség lesz még összetettebb elemzések és tervezések megvalósítására.

Puskás Péter

Puskás Péter, az Omnit Solutions Kft. projektmenedzsere (e-mail: puskas.peter@omnit.hu).