Személyes adatok feldolgozásának típushelyzetei, avagy adatvagyon-minőségbiztosítás

(Typical cases in processing personal data or quality assurance in data asset management)

Pitlik László, Szani Ferenc, Veres Péter Apertus Nonprofit Kft.

Kivonat: A dokumentum célja annak demonstrálása egy valós, de anonim adathalmaz esetét alapul véve, milyen adatvagyon-gazdálkodási kérdések merülnek fel automatikusan egy új munkavállalóban és/vagy egy új adatbázis megjelenésekor? Az itt közölt listák elemszáma elvileg végtelen nagy lehet.

Kulcsszavak: adatvagyon-gazdálkodás, minőségbiztosítás, GDPR

Abstract: The aim is, to demonstrate, what kind of questions concerning data asset management are realized in the head of a new employee and/or in case of a new data source based on an anonym data set. The list being initialized here and now can be seen as unlimited sets of elements.

Keywords: data asset management, quality assurance, GDPR

# Bevezetés

Tételezzük fel, hogy egy új dolgozó és/vagy új adatvagyon találkozása megtörténik. A feladat triviális: az adatvagyon megértése: vagyis az értelmezés kereteinek felállítása, a létező szabályok betartottságának számonkérése az adattartalmakon, ill. újszerű értelmezési lehetőségek felvetése. A példa kedvéért az adatvagyon legyen sok millió rekordnyi, ill. tartalmazza az alábbi mezőket:

1. felhasználó/személy ID-ja
   * mértékegység: dimenzió nélküli alfanumerikus karaktersor (amennyiben a jel egy fajta beszédes kód, akkor a jelen belüli szakaszok képzésének szabályai, a jelrészletek értelme)
   * definíció: egyedi azonosító
   * hasznosság: a felhasználóról több táblában, több adatbázisban nyilvántartott adatok személyhez köthetőségének garantálása
   * potenciális kockázatok/problémák/többletfeladatok:
     + az egyediség megsérülése esetén személyek keveredhetnek, ill.
     + adatszivárgás esetén már az ID illetéktelen kezekbe kerülése is lehetővé teszi további látszólag anonim adatbázisok célirányos értelmezését
     + …
2. a felhasználó email-címe
   * mértékegység: dimenzió nélküli alfanumerikus karaktersor
   * definíció: egyedi azonosító (a user és a domain együtt), hiszen adott levelező rendszeren belül azonos felhasználó azonosítóval rendelkező egyedek nem lehetnek
   * hasznosság: személyes kapcsolatfelvétel lehetőségének azonnali rendelkezésre állása
   * potenciális kockázatok/problémák/többletfeladatok:
     + amennyiben nem világos, milyen célra is adta át a felhasználó az adatkezelőnek az email-címet, akkor felmerülhet a jogtalan tartalmak kommunikálásának kockázata (vagyis elvárás, hogy az email-címhez kapcsolódó rendelkezési joga a felhasználónak világos legyen, ne sérüljön – eleve az email-cím használatának általános jellege vagy specifikusságai nyilvántartásra kerüljenek úgy, hogy az email-cím tulajdonosa ezen beállításokat bármikor módosítani legyen képes úgy, hogy csak azt és csak úgy tudja módosítani, amihez ténylegesen joga van (pl. nem tagadhatja meg a munkavállaló a munkáltató általi üzenetek beérkeztetését spam-szűrővel),
     + pl. jogosult-e egy munkavállaló felé a munkáltató akkor is közvetíteni a munkáltató számára üzletileg releváns reklámot továbbítani, ha ez az email-cím pl. gmail-es domain-en van?
     + lehet-e egy munkavállaló vs. munkáltató által érintett munkakapcsolatban nem munkáltató által fenntartott levelezőrendszerben deklarált email-címe a munkavállalónak (vö. pl. nemzetbiztonsági kockázat-e, ha közjegyzőknek gmail-es email-címe a „hivatalosan” kommunikált email-cím)
     + ha a felhasználói nevek képzése belátható logikát követ, s nem a felhasználói egyediséget tükrözi, akkor már egyetlen egy cím kiszivárgása esetén is sok felhasználó akkor is elérhető, ha ezeknek csak a neve ismert
     + …
3. a felhasználó anyja neve
   * mértékegység: alfabetikus karaktersor
   * definíció: bármilyen kultúrából származó bármilyen névadási szokásjog visszatükröződhet a használt karakterkészletnek megfelelően
   * hasznosság: a személyi számhoz hasonló (korlátozottan használható) univerzális egyedi azonosítók hiányában a személy egyértelmű meghatározásának egyik eleme
   * potenciális kockázatok/problémák/többletfeladatok:
     + nem konszolidált névváltozatok esetén az egyértelműség elvész
     + minden nem egyedi azonosítót jelentő, de egyedi azonosításra használt attribútum esetén is előfordulhat nem kívánatos együttállás (vagyis ha egymástól két/több különböző személy pl. azonos nevű, azonos anyja nevű, azonos napon és településen született)
     + …
4. a felhasználó neme
   * mértékegység: egyjegyű szám, vagy betűkód (pl. F/N, ill. 1/2)
   * definíció: a felhasználó elvileg születési anyakönyvben nyilvántartott státusza
   * hasznosság: az egyedi azonosítást segítő adatréteg (vö. pl. anyja neve, stb.)
   * potenciális kockázatok/problémák/többletfeladatok:
     + az orvostudomány mai állása szerint már a nem-váltás lehetséges, így a születési nem és az aktuálisan valós nem között eltérés lehet anélkül is, hogy ez adatrögzítési hiba lenne
     + …
5. a felhasználó születési dátuma
   * mértékegység: dátum (valamilyen konvenció preferálása mellett: pl. ÉÉÉÉ.HH.NN)
   * definíció: az adott közösség által deklaráltan használt naptárrendszer logikája szerinti (anyakönyvezett) dátum
   * hasznosság: a személyazonosítás egyik rétege (vö. anyja neve, stb.)
   * potenciális kockázatok/problémák/többletfeladatok:
     + a nem magyarországi születési adatok esetén a dátum-konvenció quasi automatikus kínálja fel a hónapok és napok keverhetőségét az 1-12-es értékek esetében (pl. 01.12. <> 12.01).
     + …
6. a felhasználó születési országa
   * mértékegység: alfabetikus karaktersor (lehet benne szóköz, kötőjel, stb.)
   * definíció: a felhasználó születési helyének (település) közigazgatási kötődése a születés pillanatában ország szinten
   * hasznosság: a személyazonosítás egyik rétege (vö. anyja neve, stb.)
   * potenciális kockázatok/problémák/többletfeladatok:
     + országok megszűnhetnek, átneveződhetnek
     + területi egységek más (korábban is létező, változatlan nevű) országokhoz kerülhetnek[[1]](#footnote-1)
     + …
7. a felhasználó születési városa (települése)
   * mértékegység: alfabetikus karaktersor (lehet benne szóköz, kötőjel, stb.)
   * definíció: a felhasználó születési helyének közigazgatási kötődése a születés pillanatában település szinten
   * hasznosság: a személyazonosítás egyik rétege (vö. anyja neve, stb.)
   * potenciális kockázatok[[2]](#footnote-2):
     + települések eltűnhetnek/megszűnhetnek
     + átnevezhetők
     + keletkezhetnek
     + egymás részeivé válhatnak nevük megőrzése mellett
     + …
8. a felhasználó állampolgársága
   * mértékegység: alfabetikus karaktersor (elvileg nem illik, hogy azonos legyen az állampolgárságot nyilvántartó ország nevével)
   * definíció: a születés pillanatában létező, állampolgárságot kínáló entitások valamelyikéhez való kötődés kifejezése
   * hasznosság:
   * potenciális kockázatok/problémák/többletfeladatok:
     + nem magától értetődő, hogy a születési hely és az állampolgárság összefügg (pl. magyar szülők gyermeke Németországban születve is magyar állampolgár, de már az USA-ba tartó repülőn születő gyermek lehet USA-polgár?)
     + lehet egy személy többszörös állampolgárság által érintett, mely jelenséget az állampolgárság kezelése esetén kényszerűen kezelni kell az egyértelműség garantálása miatt – hiszen pl. lehet, hogy adott munkaköröket többes állampolgársággal rendelkezők nemzetbiztonsági kockázatok vélelme miatt nem tölthetnek be…
     + az állampolgárság a személy akaratától függetlenül is változhat – mely jelenség látszólag nem adattárolási kérdés, ellenben masszívan elemzési feladat, ha a többes állampolgárság (nemzetbiztonsági) kockázati indexét kell becsülni esetlegesen – ott már nem mindegy, vajon egy személyes jellegű adat (akár névváltozás) nyomás alatt, vagy szabad akaratból jött létre, ill. mi volt az adat-opciók sorrendje?
     + …
9. a felhasználó felsőfokú végzettségnek léte
   * mértékegység: státuszváltozó pl. egy karakterből álló jel (I/N, vagy 1/0)
   * definíció: A nyilvántartásba vétel pillanatában felsőfokúnak számító képzettséget igazoló bizonylat létének (ellenőrzésének, bemutatottságának, másolatlétezésének?) tényét rögzítő jel.
   * hasznosság: Egy fajta származtatott attribútum, mely csoportképzésre (vö. szociológiai, munkavégzési jogosultsági csoportok) alkalmas a teljes sokaságon belüli összefüggések értelmezését támogatandó.
   * potenciális kockázatok/problémák/többletfeladatok:
     + adott ország esetén sem igaz, hogy minden időszakban egy adott bizonylat felsőfokú képzettségnek számított, ami ideális esetben a mező/attribútum-definíció végletekig részletes kifejtettségével adott célú adattárolás esetén egyértelműen kezelhetővé tehető
     + több testület esetén sem feltétlenül igaz, hogy egységes szabályok alapján kell ezen jelenséget értelmezniük
     + a nemzetközi bizonylatok kapcsán nem elég maga a bizonylat, hanem egy fajta ekvivalencia bizottsági határozat is kell, ill. csak ez kell a hazai nyilvántarthatóság érdekében
     + a képesítés/végzettség speciális esetekben akár vissza is vonható (foglalkozástól való eltiltás)?
     + a felsőfokú képzettség lehet primer módon megszerzett és következhet más státuszok eléréséből (pl. nyelvvizsga vizsga-alapján vs. külföldi végzettség alapján)
     + …
10. a felhasználó által elvégzett képzési programok nyilvántartási száma
    * mértékegység: tetszőleges hosszúságú alfa-numerikus jel
    * definíció: egyedi azonosító
    * hasznosság: a képzési program központi nyilvántartásba vételét igazoló egyedi azonosító
    * potenciális kockázatok/problémák/többletfeladatok:
      + bonyolult (hosszú, speciális írásmódokat tartalmazó) kódok kézi felvitele esetén az elírás veszélye nagy, így a konszolidálás gyakorisága szintén jelentős lehet
      + ha egy képzési program egy személyhez kapcsolva jeleni meg a riportban, akkor a képzési program megléte a személy attribútuma, de a képzési programok is lehetnek objektumok, nem csak a felhasználók, mely kihívás mentén a teljes adatvagyon újraszervezendő, ami feldolgozási idő igényes és tárhely-igényes lehet
      + …
11. a felhasználó által elvégzett képzési programok neve
    * mértékegység: alfa-numerikus karaktersor (hosszú szöveg)
    * definíció: Az egyedi azonosítóval ellátott képzés tartalmának szöveges leírása/címe/neve
    * hasznosság: Az egyedi kód értelmezési félreértéseinek feloldását támogató szöveg
    * potenciális kockázatok/problémák/többletfeladatok:
      + esetlegesen sorszámozott képzések esetén pl. a római számozás téves használata (IV. <> VI.)
      + zárójeles név/cím-részletek következetességének hiánya: pl. ha nincs zárójelben az, hogy (e-elearning), akkor az automatikusan jelent-e valami mást (pl. kontakt, vagy blended) – mindezen terminológiai/nevezéktani kihívások ideális esetben a mező/attribútum-definícióban kell, hogy tisztázásra kerüljenek – vagyis az ideális definíció soha nem a beszédes mezőnév, soha nem egy rövid mondat, hanem egy hosszú, minden egyes értelmezési kockázati forrásra kiterjedő, részletes értelmezés! Az ideális mező-definíciók ez által rengeteg kulcsszót illik, hogy tartalmazzanak, s ez a meta-kereső hasznosságát minden további erőforrás-áldozat nélkül is nagyságrendekkel növeli.
      + …
12. a felhasználó által elvégzett képzési programok keletkezésének/első hozzáférhetőségének éve vagy a képzés felhasználó által való abszolválásának/felvételének éve
    * mértékegység: év
    * definíció: évszám, amikortól az adott képzés a kiképzendők számára hozzáférhető vagy amikor a képzés az adott felhasználó által felvételre került vagy amikor az adott képzés az adott felhasználó által abszolválásra került[[3]](#footnote-3)
    * hasznosság: a képzések egyediségének jelzése, ill. a képzések kronologikus rendjének alapadata
    * potenciális kockázatok/problémák/többletfeladatok:
      + az esetlegesen sorszámozott témakörök esetében szabad-e, hogy egy nagyobb sorszámú képzés előbb keletkezzen, mint egy kisebb sorszámú képzés?
      + lehetséges-e azonos képzésnév több évszámmal ellátva?[[4]](#footnote-4)
      + …
13. a felhasználó által elvégzett képzési programokról szóló leíró rekord keletkezésének dátuma
    * mértékegység: dátum
    * definíció: a leíró adatok keletkezésének dátuma
    * hasznosság: timestamp-jellegű adatok esetén a történések valós kronológiája válik értelmezhetővé
    * potenciális kockázatok/problémák/többletfeladatok:
      + ha egy leíró rekord hamarabb keletkezik, mint a képzés éve adat tartalma, az lehet-e akceptálható?[[5]](#footnote-5)
      + ha azonos képzéseknek eltérő a rekord-keletkezési dátuma, az tolerálandó-e?[[6]](#footnote-6)
      + …
14. a felhasználó által elvégzett képzési programok teljesítettségének ténye
    * mértékegység: egy karakternyi jel (1/0, vagy I/N)
    * definíció: a képzés teljesítettségének deklarálása, ahol a nem teljesítettség jele a terv szerint vállalt, de még le nem zárt aktivitásra utal
    * hasznosság: státuszváltozó a képzés lezárásának sikerességéről
    * potenciális kockázatok/problémák/többletfeladatok:
      + a hiányzó adat nem azonos az el nem végzés deklarációjával[[7]](#footnote-7)
      + vannak-e a képzések között előfeltételi/egymásra-épülési szabályok[[8]](#footnote-8)
      + mit jelent egy olyan rekord, ahol a teljesítettség adata (még?) hiányzik?
      + …
15. a felhasználó munkáltatója
    * mértékegység: alfabetikus karaktersor (hosszú szöveg)
    * definíció: a munkáltató megnevezése utal a területi egységre és a működési módra
    * hasznosság: a munkáltató humán erőforrás-ellátottságának egyenszilárdsága vizsgálható[[9]](#footnote-9)
    * potenciális kockázatok/problémák/többletfeladatok:
      + adott személy, aki munkáltatót vált, miként értelmezendő?
      + különösen akkor, ha a munkahelyváltások nem egymást követik, hanem kihagyásokkal tarkítottak?[[10]](#footnote-10)
      + …

Nem lehet elégszer hangsúlyozni a lábjegyzetekben már több ponton is érintett gondolatot: A mezők/attribútumok leírásának egy-egy projektben/cégben a fenti mintát minimálisan követnie kell, sőt, a definícióknak olyan részletességűeknek kell lenniük, melyek alapján minden értelmezési félreértés kizárható.

Az attribútumok leírását ki kell egészíteni jelen dokumentum folytatásaként bemutatásra kerülő részletekkel is:

* milyen ellenőrzések kapcsolódnak az adott attribútumhoz annak érdekében, hogy a téves adatok feltárásra kerülhessenek?
* milyen rendszeresen (lehetőség szerint automatikusan) előállított riportokban vesz részt az attribútum?
* milyen elemzési feladatok által érintett az attribútum?
* milyen objektumokhoz kapcsolódik az attribútum, amennyiben több objektum is ismert az adatbázisok/dokumentációk alapján (vö. adatvagyon-gazdálkodási kézikönyv - <http://miau.gau.hu/miau/kofop/AVG_KK_v1.docx>)?

Mivel a rendelkezésre álló háttér-dokumentációk alapján a felhasználók adatait egy ún. képzési referens tölti fel, így az alábbiakban minden ellenőrzési megállapítás egyben a referensek munkájának egy fajta kritikájaként is értelmezhető. Amennyiben a referensek ID-ja is ismert lenne, úgy az egyes konkrét munkatársak közötti munkaminőség-különbség válna értelmezhetővé a lehetséges hibajelenségek, mint attribútumok alapján a referensek, mint objektumok a mindenki-lehet-másként-egyforma elv alapján anti-diszkriminatív és egyben objektív módon rangsorolhatók a referensek relatív viselkedési normájának levezetése és az ettől való eltérések irányának és mértékének alapján.

# Ellenőrzési feladatok

Az ellenőrzési feladatok olyan lehetőség szerint operatív feladatok, melyek egy adatvagyon minőségét hivatottak felmérni. Az előző fejezet az adatvagyon értelmezését volt hivatott segíteni. Ebben a fejezetben check list jelleggel újrafogalmazásra kerülnek azok a feladatok, melyek elvégzése kötelező/ajánlott/hasznos – a mindenkori döntéshozó saját értékítélete szerint. A listák nem teljesek, így a dokumentum ezen fejezete felhasználható egy belső képzés gyakorló/vizsga-feladataként is, ahol a kiképzendőktől elvárásra kerül, hogy a bevezetés gondolataiból keressenek az egyes mezőket érintően további operatívan elrendelhető, lehetőség szerint teljesen automatizálható ellenőrzési szempontokat:

## Általános elvárások

* Személyes (egyedi azonosításra használható) adatok esetén elvárható, hogy ne legyen adathiány.[[11]](#footnote-11)
* Minden, egyedi azonosításra használható adat írásmódja legyen konszolidált (vö. dátum, nevek, ékezetek, speciális betűk pl. SCH, kötőjelek, stb.)
* Státuszváltozók (pl. felsőfokú végzettség) esetén soha/sehol/semmikor ne legyen adathiány, vagyis a hiányzó adat nem feleltethető meg az igenlő válasz nemleges párjának, hanem az adathiány az adathiány, azaz nem ismert, van-e felsőfokú végzettsége annak, akinek sem igen, sem nem nem szerepel a neve mellett.
* …

## User-ID (1)

* Minden ID egyedi azonosító-e?
* Minden ID azonos logika szerint lett-e képezve?
* …

## Email-cím (2)

* Amennyiben egy felhasználónak egyszerre több email-címe is feltárásra kerül, akkor tisztázandó, mi ennek az oka: pl. más-más célokra (kommunikációs tartalmakat engedélyezve) más-más címeket adott-e meg a felhasználó?
* Vagy él-e minden egyes email-cím egyáltalán? (Ennek kapcsán tisztázandó szabad-e ellenőrző üzenetet küldeni minden megadott címre az esetleges nem-létezés automatikus feltárása érdekében?)
* …

## Anyja neve (3)

* Amennyiben adott személy kapcsán az anyja neve nem azonos, de pl. csak a keresztnevek számában tér el, akkor kezdeményezni kell a névkonszolidációt. Ilyen helyzet csak akkor fordulhat elő, ha a személy több adatbázisban is nyilván van tartva, vagyis az anyja neve redundáns módon tárolódik, s egyik helyen minden, a másik helyen csak az első keresztnév került felvételre.
* Amennyiben egy anyának több gyermeke is felhasználó, akkor is előállhat az a vélelem, hogy az egyik gyermek kapcsán az anya több (minden), míg a másik gyermek kapcsán kevesebb (egy) keresztnév került tárolásra az anya leánykori neve kapcsán. A felhasználók testvérként való vélelmezését erősíti az azonos családnév, s különösen az egy napon, egy helyen való születés.
* Amennyiben egy utónév-adatbázis(sal kapcsolat) kialakítható, akkor a női nevek létezését ellenőrizni kell annak érdekében, hogy potenciális elírások feltárhatók legyenek. Az idegen nevek esetén érdemes olyan státuszváltozót bevezetni, mely jelzi, hogy ezen neveket nem ajánlatos a magyar utónevek jegyzékével összevetni a jövőben sem.
* Vélelmezhetően kicsi/nulla az esélye annak, hogy valakinek csak egy tagú neve legyen (vagyis keresendő, van-e minden névben legalább egy darab szóköz). Ennek hiányában a névrögzítés potenciális hiányosságai vélelmezhetők.
* Egy személynek csak egyféle anyja-neve-adata lehet.
* …

## Neme (4)

* Egy személynek ma már nem csak egy neme lehet, de a nem-váltás adott országra érvényes nyilvántartási szabályaival (pl. egyben név váltás is? – Kiss Gabi? Gábor, vagy Gabriella?) részletesen össze kell vetni a lehetőségét annak, hogy igaz lehet-e: egy személynek csak egy neme lehet?
* …

## Születési dátum (5)

* egy embernek csak egy születési dátuma lehet
* minden dátum esetén a hónap/nap csere kockázatának feltárása nem magyar születési hely esetén
* …

## Település/Ország (6/7)

* Előfordulhat-e, hogy határmódosítások révén adott város más-más országokhoz (is) kapcsolódik? Ebben az esetben tisztázandó: mikor tekinthető egy ilyen anomália (az 1:1 reláció sérülése) hibának és mikor a történelem „fintorának”?
* Amennyiben az országok, városok neve nem egy egységes nevezék-tárból érkezik, akkor fel kell tárni a névváltozatokat és konszolidációs lépéseket kell tenni a változatok egységes terminológia szerinti kezelésének kikényszerítése érdekében.
* Egy felhasználónak csak egy születési városa lehet (kivéve, ha a település neve időközben változott?)
* Minden alfabetikus jelsorozat kapcsán ellenőrizendő az írásmód-változatok léte, s feladat ezek konszolidálása.
* …

## Állampolgárság (8)

* A többes állampolgárság elvileg egy időben jelent több kötődést, míg a rekordszerkezet szerint egy rekordban csak egyetlen egy állampolgárság kezelhető, ami kizárja a struktúra okán a többes állampolgárság egyidejűségét.
* Egy felhasználó esetén az egynél több állampolgárság váltás mindenképpen felveti a téves adatrögzítés gyanúját az összes rekord alapján, sőt, már egyáltalán az állampolgárság váltása is adatrögzítési minőséget érintő kockázatként értelmezendő.
* …

## Felsőfok (9)

* Üres rekordnak nem illene lennie, mert az adathiány nem azonos a még fel nem vitt adattal.
* Különösen akkor értelmezhetetlen az adathiány, ha más státuszváltozókban van igen/nem opció is…
* …

## Képzés-ID (10)

* Beszédes kódok esetén minden olyan kód feltárandó, mely adott szekvenciája egyedi, más kódokban elő nem forduló. Ha van kódszótár, akkor igaz-e, hogy ezen egyedi rétegek léteznek a kódtáblában (s természetesen a többszörösen előfordulók esetén is ellenőrizendő, létezik-e a kódrendszerben az a kódréteg)?
* Ha van a kódoknak belső logikája, akkor létrehozható-e az ún. ellenőrző összeg? (pl. adószám, TAJ-szám esetén)
* Előfordult-e, hogy egy képzést adott személy több, mint egyszer abszolvált?
* …

## Képzés neve (11)

* Minden képzés-ID-hoz csak egy képzés tartozik-e?
* (Hasznos lenne tudni: mi a szabályozása az esetleges mentességeknek? Vagyis legalább egy státuszváltozóra szükség lenne, mely adott felhasználó adott képzés metszeteként a mentességre utaló jelet tartalmazza – vagy maga a képzés teljesítettsége esetén lehetséges pl. igen/nem/felmentve opciók alkalmazása.
* A képzések kapcsán hasznos információ lenne a kötelező/választható/stb. státusz, mert ennek hiányában a képzésenkénti létszámadatok kevésbé értelmezhetők.
* …

## Képzés éve (12)

* Nem hiányozhat adat.
* Ha az egyes képzések egymásra épülnek, ill. van olyan képzés, melyet elsőként kell realizálni, akkor a dátumoknak ezen egymásra épüléseket vissza kell tükrözniük.
* Amennyiben a képzések teljesítésére az első képzéshez képest X idő áll rendelkezésre, akkor ellenőrizendő, vajon minden felhasználó esetén betartásra került-e ezen intervallum-korlát?
* (A képzések esetleges elhúzódása, eltolása ennyi attribútum alapján nem értelmezhető…)
* (A nem normatívan elbírált halasztási lehetőségek sem érhetők tetten, ha nem áll rendelkezésre a képzéssel kapcsolatos összes esemény és ezek időpecsétje…)
* Adott felhasználó mennyire egyenletesen teljesít (vö. szakmai életút kiegyensúlyozottsága)? Tetten érhető-e a határidő lejárta előtti teljesítési kényszer/pánik jelensége?
* …

## Rekord-dátum (13)

* Nem hiányozhat adat.
* A képzés éve és a rekord dátumának viszonya (azonos év – preferált / előbb-később gyanús).
* …

## Képzés teljesítettsége (14)

* Nem hiányozhat adat.
* …

## Munkáltató (15)

* Vannak-e konszolidációt igénylő elírások a hosszú szövegekben?
* Területi adatok kinyerése (megyék, járások) – minden megye és járás szerepel-e a kinyert adatok között?
* …

# Elemzési feladatok

## Általános feladatok

### Plauzibilitás tesztek

Az egyedi adat gyors-ellenőrzésének lehetőségei:

* dátumok legyenek dátumok (napok száma hónapokon belül)
* számértékek legyenek az értelmezési (max-min) intervallumon belül pl.
  + születési évszám élőnek vélt felhasználók esetén nagyobb, mint xxxx
  + születési évszám munkavállalónak vélt felhasználók esetén kisebb, mint zzzz
  + képzés évszáma legyen nagyobb-egyenlő az első képzés évszámával
  + leíró rekord dátuma legyen a rendszerüzemelés határain belül (akár az is lehet, hogy pl. leíró rekord csak munkanapra esően és csak 02.00-03.00 közötti érték lehet kötegelt adatfeldolgozás esetén)
  + …
* kényszerűen több tagból álló szövegekben legyen szóköz (pl. nevek)
* …

### Konzisztencia-tesztek

Több adat együttállási kényszereinek ellenőrzési lehetőségei:

* név és nem összefüggése (vö. férfi/női utónév-adatbázis)
* születési dátum és képzettség együttállása (hány éves kortól érhető el leghamarabb egy felsőfokú végzettség)
* születési dátum és képzések maximális száma
* ország és település együttállása (közigazgatási határok adatbázisa alapján)
* állampolgárság és születési hely/idő együttállása a hazai/nemzetközi jogi szabályok alapján
* többes állampolgárságok együttes létezhetősége (nemzetközi jogi szabályozás alapján)
* munkahely és domain-kötelezettség együttállása
* …

## Mező-specifikus feladatok

### User-ID (1)

* Az egyediség garantálásán túl más értelmezés ennyi attribútum összefüggésrendszerében nem releváns.
* …

### Email-cím (2)

* Adott munkáltató esetén milyen domain-ek fordulnak elő?
* A feltárt domain-ek mindegyikéről belátható-e, hogy ezek nem a felhasználó által létrehozott (privát) email-szolgáltatások annak érdekében, hogy az esetleges GDPR-felelősségi körök, ill. egyéb (pl. nemzet-)biztonsági kockázatok mérlegelhetők legyenek.
* Amennyiben csak privát domain van egy adott munkáltató esetén a felhasználók email-címeiben, akkor további feltáró munka keretében tisztázandó: mi ennek az oka? Van-e joga a munkáltatónak a munka-kommunikációt kiengedni nem saját felügyelete és felelőssége alá tartozó szolgáltatókhoz? (pl. áramlanak-e személyes adatok email-csatornákon át?) – Ezen vizsgálatok ismét csak az adatvédelmi felelősségek lehatárolását és a kockázatelemzést támogatják.
* Amennyiben egy munkáltató esetén a privát és munkáltatói domain-ek vegyesen fordulnak elő: mi az aránya a privát domain-eknek? Amennyiben nincs kivételkezelési szabályozás, csak úm. spontán sodródás, az ismét felvetik a kockázatkezelés elégtelenségének gyanúját, ill. a felelősségi körök átgondolatlanságának gyanúját.
* További feltárást igényel: A privát domain-ek által érintett felhasználók milyen jogosultság alapján térnek el a többi felhasználó céges domain-használatától? Minden privát (nem a cég hatáskörébe tartozó) domain használata kapcsán az erre is átkerülő tartalmakhoz kötődő felelősségek/kockázatok feltárása elengedhetetlen a GDPR logikáját követve.

### Anyja neve (3)

* Egy fajta szociológiai csoportképzés lehetőségét hordozza a többszörös keresztnevek léte, nem léte.
* …

### Neme (4)

* Adathiány esetén (mely a minőségileg aggályos adatrögzítés és a felelőtlen adatvagyon-gazdálkodás tipikus jegye – hiszen még az adattárolásban érintettek sem végeztek attribútum-ellenőrző lekérdezéseket?) a felhasználó nevének rendelkezésre állása segítheti a nem algoritmikus levezetését legalább is az anyakönyvezhető névvariánsok ismeretében – eltekintve az esetleges azonos-alakú, de eltérő nemhez is használható nevektől.
* Az email-címek esetén csak a teljes nevet sejtető felhasználó-azonosító vélelme esetén lehet adathiány pótolni, hiszen az email-címből kiolvasható név ugyanolyan elemzés alapja lehet, mintha a név maga áll rendelkezésre az előző pont értelmében és korlátai mellett.
* …

### Születési dátum (5)

* Asztrológiai attribútumok képzésének alapja.
* A választott képzések és pl. a csillagjegyek között elvileg lehet összefüggés (vö. ajánlórendszerek esetén speciális csoportképzési lehetőség).
* …

### Település/Ország (6/7)[[12]](#footnote-12)

* Nem magyar (településen született) felhasználók aránya.
* A nem magyar érintettségek egymáshoz képesti aránya.
* A nem magyar érintettségek időbeli alakulása (trendje).
* …

### Állampolgárság (8)[[13]](#footnote-13)

* A nem csak egy állampolgársággal rendelkezők száma,
* megoszlása a többszörösség értékei mentén.
* A nem csak egy állampolgársággal rendelkezők által érintett országok száma.
* A többes érintettség időbeli alakulása (trendje).
* …

### Felsőfok (9)

* Vizsgálandó, feltárható-e az adatok alapján összefüggés egyes képzések és a felsőfokú képzettség, mint adottság között. Itt csak a hibátlan klasszifikáció a valódi szabály létezésének indikátora.
* Amennyiben csak minimális szabályszegő objektumra derül fény, akkor ezen szabályszegést sejtető adatokkal szemben fenn áll a téves rögzítés gyanúja.
* …

### Képzés-ID (10)

* Az egyediség garantálásán túl más értelmezés ennyi attribútum összefüggésrendszerében nem releváns.
* …

### Képzés neve (11)

* A képzés nevének hossza (komplexitása) és az ezt választó felhasználók száma közötti korreláció?
* A képzés nevek hosszának alakulása időben (adott képzés-realizálási évben átlagos névhossz trendje).[[14]](#footnote-14)
* …

### Képzés éve (12)

* A felhasználók esetén a születési dátum, a képzések esetén a képzés éve olyan változó, mely alapján trendek határozhatók meg quasi minden más attribútumot érintően.
* …

### Rekord-dátum (13)

* A születés dátuma és a képzés éve mellett ez a dátum-típusú attribútum az egyetlen, mely napon belüli időpont-értelmezést is lehetővé tesz. Így vizsgálható pl. a munkaidőn belül/kívül született bejegyzések aránya.
* A munkaidőn kívüliség alapján vizsgálható a bármilyen jellegű (képzési adatokat érintő) hibák aránya a munkaidőben és az azon kívül született adatok esetén.
* Minden képzés név, képzés-ID elvileg csak egyetlen egy dátumhoz illene, hogy kapcsolódjon egy központi adatrögzítést feltételezve.
* …

### Képzés teljesítettsége (14)

* Amennyiben vannak olyan rekordok, ahol nemleges a teljesítettség, akkor a sikeresség és a sikertelenség okai (látszatkorrelációi) minden attribútummal szemben vizsgálhatók, ill.
* termelési függvények alkotható a várható siker becslésére.
* …

### Munkáltató (15)

* a munkáltatók által foglalkoztatottak kiképzettségének homogenitása
  + adott képzések meglétének aránya (0-1) a munkavállalók számához képest
  + adott képzésszámmal rendelkező munkavállalók száma
  + …
* …

# Riportok

Riportként értelmeződik minden olyan adat-együttállás, melyre viszonylagos rendszerességgel előre definiálható felhasználók illene, hogy rápillantsanak. A rendszeres riportok egy része kényszerűen az adatvagyon minőségellenőrzését érinti, hiszen addig, amíg a szükséges/megfelelő minőségi szint nem került elérésre más fajta riport kialakítása felesleges/kockázatos/értelmetlen:

* az egyes attribútumok select-distinct nézete, melyből kiderül, milyen egyedi tartalmak tárolódnak az adott szempont mögött, s melyik opció hányszor fordul elő egyáltalán a teljes sokaságot alapul véve (ez a nézet a konszolidálási gyanúk általános forrása)
* felhasználók száma mindösszesen
* felhasználók számának időbeli alakulása
* képzések száma mindösszesen
* képzések számának időbeli alakulása
* a dokumentum egyéb részeiben leírt adat-ellenőrzési anomáliák riportjai - pl.
  + adathiányok listája,
  + nem egyértelmű személyek listája
    - egy személy több születési időpont
    - egy személy több anyja neve
    - egy személy több születési hely
    - …
  + adott szempontból túl fiatal/idő személyek listája
    - felsőfokú képzettség megszerzésének minimálisan elvárt idejéhez képest
    - adott képzés megkezdhetőségéhez képest
    - …
  + speciális települések listája
    - több ország által érintett
    - …
  + minden felhasználó által felvett/teljesített képzések listája
    - a kötelező képzés hiányzó státusz-változójának becslésére
    - …
  + konszolidációt igénylő adatok listája
    - vélelmezhetően elírt településnevek, névváltozatok, adott nyelven és több nyelven
    - vélelmezhetően elírt ország-nevek, névváltozatok, adott nyelven és több nyelven
    - vélelmezhetően elírt nevek, névváltozatok, adott nyelven és több nyelven
    - …
  + …
* …

# Ajánlások

Ajánlásként olyan gondolatok kerülnek megfogalmazásra, melyek adott adatmennyiségen esetlegesen túlmutató következtetéseket érintenek, ill. a fentebb kényszerűen minőségbiztosítási túlsúllyal megfogalmazottak rendezése után merülhetnek fel:

* Státuszváltozók (pl. felsőfokú képzettség léte – igen/nem): tovább-felhasználás előtt ellenőrzéseket igénylő adattartalom:
  + Ilyen esetekben egyértelműen hiányzik a direkt kapcsolat a státusz ellenőrzéséhez.
  + Tudni kell, vajon a státusz manuális adatbevitel eredménye-e?
  + Ha igen, ki mikor vitte be az adatot?
  + Ha nem, milyen algoritmus fut a háttérben?
  + Ennek mikor volt az utolsó frissítése?
  + A frissítés előtti és utáni státusz-érték azonos-e?
  + Ha nem, lehet-e frissítési hibáról szó?
  + …
* OLAP-szolgáltatások potenciális objektumai:
  + felhasználók (pl. felhasználók által érintett képzések száma, dinamikája)
  + képzések (pl. adott képzés által érintett felhasználók száma, dinamikája)
* A felhasználók a rendelkezésre álló adatok alapján típusokba sorolhatók:
  + egyrészt a döntési fák módszertanát követve, amennyiben van következményváltozó (pl. kor – azaz adott kor elérése milyen tulajdonságkombinációkból vezethető le – ill. adott személy szakmai tapasztalatai/képzettsége alapján milyen korosztálynak feleltethető meg?)
  + másrészt a klaszterezési eljárások keretében önkényes csoportok képezhetők
  + harmadrészt tetszőleges fogalom-absztrakció (pl. tapasztaltság, kiegyensúlyozott szakmai életvezetés, stb.) létrehozható az alapadatok (és az ezekből esetlegesen származtatott adatok) valamint ezek absztrakcióval kapcsolatos irányultsága alapján (pl. minél több képzésen vett részt valaki, annál tapasztaltabbnak számít, még ha csak virtuálisan is találkozott esetekkel az is jobb, mint ha sehogyan sem)…

A típusba sorolás adja az alapot az ajánló rendszerek kialakításához: amennyiben az egyes típusokhoz tartalomfogyasztási statisztikák rendelhetők hozzá, akkor a csoportra leginkább jellemző tartalom azon csoporttagok számára is ajánlható, akikről ennek fogyasztása kapcsán még nem áll rendelkezésre visszaigazoló adat a rendszerben.

* Az esetleges többszörös (kockázatos domain-t is megengedő) email-címhasználat elkerülését elősegítendő hasznos lehet az intézményi email cím kötelező használata és intézménytől független háttér-email cím használata annak érdekében, hogy a munkahelyváltások során a felhasználó önazonossága megmaradhasson.
* A GDPR elvárások értelmében adott felhasználó minden rekordjának fizikai törlését kezelő szabályzat kialakítása szükséges minden esetben.
* Részletes, minden elképzelhető helyzetre kiterjedő adatkezelési szabályzat kialakítása szükséges a GDPR szellemében, hogy személyhez köthető adatvagyonhoz csak olyanok kezelhessenek, akiknek megvan erre a jogosultsága.
* Az esetleges archív (rejtett, stb.) felhasználók esetén státuszváltozók bevezetése ajánlott annak érdekében, hogy az ún. aktív felhasználók listájában mindenkor csak azok legyenek benne, akik valóban odatartoznak. Az archív/rejtett állapot (pl. elhunyt felhasználó?) kérdésköre nem azonos a felhasználó kérésére, vagy más jogügylet hatására megindított fizikai törlés kérdésköreivel.
* Az adatvagyon megfelelően védett (jogi: pl. jogosultságkezelés, ill. technológiai: pl. betörésvédelem) részleteinek OLAP-szolgáltatássá alakítása segítheti a potenciális felhasználók edukálását, kiszolgálását.
* A cég/projekt jogi szakértőinek ajánlatos az egyedi és az összes mezők adatvédelmi szintjét megbecsülni és a kapcsolódó intézkedéseket az adatvédelmi tervezéshez, jogosultságkezeléshez feltárni.
* …

# Potenciális következmény-rétegek

* Abban az esetben, ha az adatkezelő közvetlenül érintett munkakörök és ezek felügyeletét ellátó munkakörök kezéből alapvetően ellenőrizetlen adatok kerülnek ki (pl. személyes adatok esetén 1%-ban nincs születési dátum, hely, anyja neve), akkor vélelmezhető, hogy az érintettek nem végeztek kellően gondos munkát, melynek következményei a keretfeltételek függvényében:
  + ha évek óta fennáll az anomália, akkor már a szervezet maga kell, hogy szemléletet váltson
  + ha új vezetés tárja fel az anomáliákat, akkor a felügyeleti/végrehajtásban régóta érintettek számára
    - az azonnali felmondástól
    - a figyelmeztetésen át
    - a kötelező képzések indításáig számos akció felmerülhet
* Az személyes adatokat tartalmazó adatvagyon tovább-feldolgozása kapcsán lehetőség van
  + a személyes adatok kitakarására, elleplezésére, törlésére,
  + mely akciók függvénye az, hogy bármilyen (pl. OLAP) szolgáltatás intraneten milyen adatkezelési szabályzat, milyen hozzáférési részletszabályok mellett milyen vezetői szint meg élesíteni…
* Gazdasági értelemben fontos döntési pontok
  + az anomáliák feltisztítása nélkül érdemes-e adatelemzést végezni?
  + érdemes-e pl. az Excel kb. 1 millió rekordos határát átlépve speciális elemzési környezet kialakítását felvállalni?
  + vagyis értékes-e bármilyen részhalmaz elemzése a teljes (hiányos és/vagy minőségileg rendezett) adatvagyonhoz képest?
* …

Tartalom

[Bevezetés 1](#_Toc504119134)

[Ellenőrzési feladatok 8](#_Toc504119135)

[Általános elvárások 8](#_Toc504119136)

[User-ID (1) 8](#_Toc504119137)

[Email-cím (2) 8](#_Toc504119138)

[Anyja neve (3) 8](#_Toc504119139)

[Neme (4) 9](#_Toc504119140)

[Születési dátum (5) 9](#_Toc504119141)

[Település/Ország (6/7) 9](#_Toc504119142)

[Állampolgárság (8) 9](#_Toc504119143)

[Felsőfok (9) 10](#_Toc504119144)

[Képzés-ID (10) 10](#_Toc504119145)

[Képzés neve (11) 10](#_Toc504119146)

[Képzés éve (12) 10](#_Toc504119147)

[Rekord-dátum (13) 10](#_Toc504119148)

[Képzés teljesítettsége (14) 11](#_Toc504119149)

[Munkáltató (15) 11](#_Toc504119150)

[Elemzési feladatok 12](#_Toc504119151)

[Általános feladatok 12](#_Toc504119152)

[Plauzibilitás tesztek 12](#_Toc504119153)

[Konzisztencia-tesztek 12](#_Toc504119154)

[Mező-specifikus feladatok 12](#_Toc504119155)

[User-ID (1) 12](#_Toc504119156)

[Email-cím (2) 12](#_Toc504119157)

[Anyja neve (3) 13](#_Toc504119158)

[Neme (4) 13](#_Toc504119159)

[Születési dátum (5) 13](#_Toc504119160)

[Település/Ország (6/7) 13](#_Toc504119161)

[Állampolgárság (8) 13](#_Toc504119162)

[Felsőfok (9) 14](#_Toc504119163)

[Képzés-ID (10) 14](#_Toc504119164)

[Képzés neve (11) 14](#_Toc504119165)

[Képzés éve (12) 14](#_Toc504119166)

[Rekord-dátum (13) 14](#_Toc504119167)

[Képzés teljesítettsége (14) 14](#_Toc504119168)

[Munkáltató (15) 15](#_Toc504119169)

[Riportok 16](#_Toc504119170)

[Ajánlások 17](#_Toc504119171)

[Potenciális következmény-rétegek 18](#_Toc504119172)

1. Egy jól felépített informatikai rendszerben ezek a jelenségek már nem kockázatok, hanem masszív többletfeladatok: az aktuális országok tárolása és a világ valaha volt összes országának tárolása, ezek létének hatályossága, valamint ezen háttér-adatok összevetése a születési dátummal kezelhető és kezelendő feladatsor. A cégek/projektek szempontjából hasznos lenne egy országos/nemzetközi információs szövetkezés, mely működésének eredményeként a háttéradatok bárki számára ingyenesen rendelkezésre állnának… [↑](#footnote-ref-1)
2. Az országokról szóló háttéradatokhoz hasonló hazai/nemzetközi keretrendszer a települések esetében is szükséges/lehetséges lenne... [↑](#footnote-ref-2)
3. Tipikus (itt és most tudatosan kihangsúlyozott) példája annak, hogy csak a részletes definíció mellett lehetséges a potenciális „félre”-értések kizárása, ugyanis egy személy, egy képzés és egy évszám háromszögében számos értelmezési konstrukció racionális lehet. [↑](#footnote-ref-3)
4. Az egyes mezők egymáshoz való kapcsolata a definíciók szerves része kell, hogy legyen, ami tovább csökkenti a félreértelmezhetőségi potenciált. [↑](#footnote-ref-4)
5. Ha a képzésnek nem lenne keletkezési éve, de van első felmerülési éve, akkor az felfogható a képzés keletkezési évének, vagyis olyan fogalmak is életre kelnek, melyek akár elsődlegesen a definíciók hiányosságai miatt alapvetően félreértésként keletkeztek. [↑](#footnote-ref-5)
6. Az egyes mező-mező összefüggések közül a minőségi definíciók mentén kizárhatók azok, melyek nem tartoznak a tényleges értelmezési térbe. [↑](#footnote-ref-6)
7. Ismét kiemelendő: az adathiány értelmezése ideális esetben az adat definíciójának része. A kockázat fogalma akkor merül fel, ha megfelelő definíciók hiányában mégis előállnak adathiányos állapotok ahelyett, hogy egy igen/nem állapotpár mellé egy még-nem-került-feltárásra állapot is bevezetésre kerülne, vagyis az adathiány, mint olyan kizárása megtörténne központilag. Ezek a kérdések nem hagyhatók nyitva az elemzések megkezdésének pontjáig, különben az elemzések pl. teljesen felesleges dolgokat fognak vizsgálni [↑](#footnote-ref-7)
8. Az egyes képzések egymásra épülési rendszere alapján az adathiányok automatizmusokkal is feloldhatók, feltételezve, hogy a képzési rendszer racionális: vagyis: ha egy/több olyan tárgy már teljesítettként van nyilvántartva, mely előfeltétele esetén a teljesítettségről nem áll rendelkezésre semmilyen adat vagy ennek nem teljesültségét jelezné a státuszváltozó, akkor ezen hiány vagy téves adat „felülírása” automatikusan vagy manuális engedélyezéssel megtörténhet. Az ilyen típusú inkonzisztencia-gyanúk kapcsán a rendszer üzemeltetőire olyan részletes szabályok kell, hogy kidolgozásra kerüljenek, melyek lényegében a hibafeltáró és hibajavító robotok „hozzáférési” / „aktivitási” jogosultságainak felelnek meg. [↑](#footnote-ref-8)
9. Az egyenszilárdság egy absztrakt fogalom, közvetlenül nem mérhető: jelentése nem más, mint annak a potenciál indexnek a becslése, mely kifejezi, az egyes munkáltatókhoz (objektumokhoz) tartozó dolgozók átlagában mely képzés (mint attribútum) milyen arányban teljesült. A képzések sokszínűsége miatt esetlegesen szükség lehet konszolidációra (pl. A/B/C típusú képzések fogalmának bevezetésére). Amint az OAM statisztikai szinten felállt, feltehető immár a fogalom-alkotási kérdés a mesterséges intelligenciák számára: lehet-e minden munkáltató egyformán jól képzett dolgozókat foglalkoztató objektum? Az OAM kapcsán K.O. feltételek is megfogalmazhatók: mi az a minimális szint adott képzés átlagos rendelkezésre állása kapcsán, aminek mindenképpen teljesülnie kell? Ideális esetben a dolgozók és a munkáltatók közötti aktív/élő munkaviszony adatai is ismert, mert ennek hiányában csak azt lehet vélelmezni, hogy adott évben, ill. több év átlagában (a teljes adathalmazon) milyen képzés-teljesítettségi arányszámok közelíthetik az élő munkaviszonyt. [↑](#footnote-ref-9)
10. Már a szóhasználat (miként értelmezendő) is jelzi, a kockázatok/problémák/többletfeladatok egyik alapvető forrása a definíciók nem ideális részletezettsége! [↑](#footnote-ref-10)
11. Ez az elvárás természetesen relativálható: Egyes szakértői vélemények szerint Magyarországon a személyes adatok megszerzésére – egyes kivételes szituációkat kivéve – kizárólag önbevallás alapon van lehetőség, így ez túlzó akár elvárásnak tűnhet. Ráadásul a – nem kifejezetten ilyen célra létrejött nyilvántartások kivételével – az egyes személyes adatok tárolásának indoka nem tisztázott az egyes projektekben/cégekben, vagyis feltétlenül indokolt/indokolható, hogy egyáltalán ott és akkor bárki kezelje az adott adatot. Más megközelítésben azonban: ha már valaki bármilyen felhatalmazással és okból adatokat kér be a személy azonosítását illetően, akkor ezt érdemes kötelező mezőként online űrlapon bekérni (az adatközlő büntetőjogi felelőssége mellett) annak érdekében, hogy ha már egyszer kezelésre kerülnek adatok, akkor a személyes adatok ne tartalmazzanak lyukakat. Mindösszesen: ha egy személyes adattömb lyukas, akkor azonnal felvetődik ezen adatok kezelésének visszavonását kivizsgáló eljárás beindításának szükségszerűsége! [↑](#footnote-ref-11)
12. Amennyiben a területi kötődés jogszerűen és racionálisan nyilvántartásra kerül, úgy vélelmezhetően ennek információtartalmát ki kell aknázni. Az földrajzi objektumokhoz/objektumcsoportokhoz való kötődés (ahol a csoport már elemzői/döntéshozói absztrakció) arányai kockázati alakzatokat tárhatnak fel: pl. már az is egy fajta kockázat, ha pl. a nem anyaországi kötődésűek csoportja esetében pl. a szaknyelvi tudás egyenszilárdságát felmérve kiderül, hogy belső képzésekre lenne szükség a fogalmi rutin trenírozásához. A földrajzi kötődésű csoportok arányainak időbeni változása a kapcsolódó kockázatok/intézkedések tervezhetőségét segíti értelmezni. [↑](#footnote-ref-12)
13. A területi kötődéshez hasonlóan, amennyiben az állampolgárságot érintő adatok jogszerűen és szükségszerűen kerültek tárolásra, akkor ezek értelmezését is ki kell alakítani. A széles értelembe vett kockázatmenedzsment számára a default állapottól (csak 1 és csak magyar állampolgárság) minden eltérés a döntési fák generálása kapcsán racionális szabályképző erőtérként értelmezhető – még ha látszatkorrelációs formában is… [↑](#footnote-ref-13)
14. A karakterhossza minden szöveges tartalomnak a szövegbányászati rutin értelmében információértékkel bírhat: vö. <http://miau.gau.hu/miau/234/titanic.docx> [↑](#footnote-ref-14)