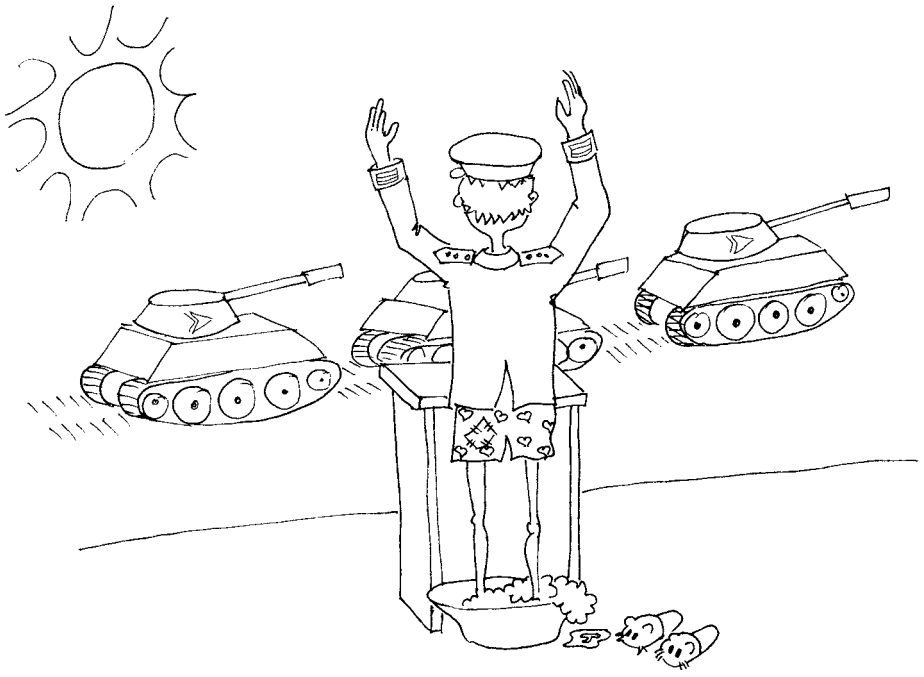
10. Mit jelent az adatminőség?

„Minden rendszerben olyan az adatminőség,
amellyel a rendszer még éppen üzemelni tud."



10.1. Az adatminőség üzleti jelentősége

Az informatika alaptörvényei közé tartozik az ún. GiGo-törvény: „Garbage in garbage out!", amely kimondja azt az igazságot, hogy rossz adatból csak rossz eredményhez lehet jutni. Valaki levelet kapott a kedvenc bankjától, amelyben nagy meggyőződéssel ajánlották egyik hitelkártyaterméküket. A dolog meglepő volt, mert az ügyfél már fél |217| éve rendelkezett ilyen típusú hitelkártyával, és nem értette, ez most mit is jelent. Talán elfelejtették? Vagy kiválthat még egyet? Esetleg megemelik a hitelkeretét? Mivel az elfelejtés nem tűnt túl valószínűnek, ezért arra jutott, hogy valószínűleg kiválthat még egyet, így néhány nap gondolkodás után be is ment az egyik fiókba a kártyáért. Ott azonban a szívélyes fogadtatás után kisebb tanakodás kezdődött, majd kiderült, hogy neki ez mégsem jár. Belső adminisztrációs hiba történt és elnézését kérik. Elég rossz szájízzel jött el az ügyfél, és komolyan megrendült a bizalma a bankban; vajon szabad olyan emberekre bíznia a pénzét, ahol ilyen esetek előfordulhatnak?

Bár a fenti eset lényegében anyagi kárt nem okozott, mégis jól példázza, hogy hogyan válhat egy egyszerű, jelentéktelennek gondolt adminisztrációs hiba súlyos presztízsveszteséggé egy intézmény számára. (A példában az ügyféladatok helytelen rögzítése miatt a hitelkártya- és a számlavezető rendszer nem tudott megfelelően összedolgozni, és a DM levelek nagy tömegű kiküldése során „elfelejtették", hogy az ügyfélnek már volt hitelkártyája.)

A fentiekhez hasonló esetek naponta előfordulhatnak a hétköznapi életben, számosságuk azonban előbb-utóbb elérhet egy olyan szintre, ahol már rombolja a vállalat imidzsét és megnehezíti az adatokkal dolgozók mindennapi életét. Ekkor megfogalmazódik a szándék, hogy ezeket a hibás adatokat valahogyan kezeljék, javítsák az adatok minőségét. Ezt pedig a sok apró hiba miatt leghatékonyabban adatbányászati eszközökkel el lehet érni.

A szakterület sokéves tapasztalata azt mutatja, hogy az adatminőség nem abszolút, hanem inkább relatív fogalom. Minden rendszerben olyan az adatminőség, amellyel a rendszer *éppen* üzemelni tud. Vagyis ha a rendszerünk különösebb probléma nélkül üzemel, akkor az adatminőségünk jó. A bajok sokszor akkor jelentkeznek, amikor változik a rendszer, változnak a követelmények, mert akkor változik az adatminőség is. Általában romlik, mert a változás által megemelkedett adatminőségi elvárásoknak az adatok már nem tesznek eleget. A fenti példában bevezettek egy új terméket, a hitelkártyát, az adatok minősége azonban arra már nem volt megfelelő, hogy a hitelkártya- és a számlavezető rendszert hibamentesen összeköthessék. Vagy ha csak arra használnak egy telefonszámot, hogy az ügyintézők felhívják rajta az ügyfelet, addig bármilyen formátumban beírva megfelelő a minősége. Ha azonban már arra is használni kívánják, hogy gépileg sms-t küldhessenek rá, akkor már formailag (kötőjelek, perjelek kivétele) és tartalmilag (országkód, körzetszám stb.) is egységesíteni kell; olyan formára kell hozni, amelyet az sms-t küldő program megkíván.

10.2. Az adatminőség problémái

10.2.1. Az adatokkal szembeni elvárások

A legegyszerűbben észrevehető probléma a mezőszintű adathiba. Minden mezőhöz
|218| megadható egy értelmes értékkészlet. Ha ebből kilóg az érték, akkor hibajelzést kapunk: például a negatív jövedelem és a 150 év feletti életkor ennek számít. A kitöltetlenség szintén hiba. A rekordszintű hiba már bizonyos összefüggéseket sérthet. Férfi pácienseken végrehajtott nőgyógyászati beavatkozáskóddal remélhetőleg csak adathiba miatt találkoztunk.

Ezek az adathibák úgy kezelhetők, hogy az üzleti szakértők által definiált elvárásokat dokumentáljuk, a rendszer részének tekintjük, és rendszeresen ellenőrző programokat futtatunk. Típushibák esetében létrehozunk metabázis alapú kritériumrendszert (pl. minden mező mellé rendelünk egy értékkészletet leíró táblát) generált ellenőrzőprogramokkal. Speciális esetekben a dokumentáció alapján egyedileg fejlesztett ellenőrzőprogramokat kell írni. A fejlesztés során felbukkanó többértelműségeket az üzleti oldallal kell egyeztetni, a feltételezéseket pedig dokumentálni kell.

Adatminőség-problémákat gerjeszthet egyrészt az adatok mennyiségéből eredő elégtelen működés, mint például nagy mennyiségben visszakapott értesítő levelek vagy a mindennapi munkát megkeserítő szintű belső *ügyfél-duplikáció*. Szintén problémákat okozhat, ha egy működő rendszert további funkciókkal szeretnék bővíteni (mint az említett telefonszám példájában). Az adatminőség-problémákat két alapvető kategóriába sorolhatjuk:

- Értékhiba. Ezekről a tételekről ránézésre tudni lehet, hogy hibásak, mert:
  - nem tartanak be valamilyenformai megkötést (például: adószámok esetén a rögzítés formája: a nyolc számjegy, kötőjel, egy számjegy, kötőjel, két számjegy),
  - értékkészlete nem egyezik meg az elvárttal (például: a város mezőbe írt információ nem szerepel a magyarországi települések jegyzékében),
  - valamilyen matematikai jellegű összefüggésnek nem tesz eleget (például: ügyfél születési ideje kisebb, mint 1900.01.01, súlya nagyobb, mint 300 kg).
- Összefüggésekből eredő adathiba. Ide tartoznak azok a hibák, amelyek önmagukon belül helyesnek látszanak, de egy viszonyítási ponthoz képest inkonzisztenciát okoznak:
  - Rekordon belüli összefüggéshiba (például: város nem konzisztens az irányítószámmal, utónév nem konzisztens az ügyfél nemével)
  - Rekordok közötti összefüggéshiba:
    - Technikai (egyediséget sértő, kulcsolást sértő, kötelező kitöltést sértő tételek)

- Duplikációk: amikor egy tétel (például ügyfél) többször szerepel a rendszerünkben, mint ahányszor kellene
- „Dummy" értékek: olyan értékek, amelyeknek nincs valódi információtartalmuk, azaz csak azért lettek beírva, mert rögzíteni kellett valamit (például: az adószám kötelezően kitöltendő érték, helyén "11111111" szerepel)

|219|

A duplikációk kezeléséhez nagyon hasonló problémakör az adatkonszolidáció, amelynek tipikus esetei az alábbiak:

- a cég törzsadatait egy vagy több külső forrásból származó adatbázisból átvett adatokkal szeretnék gazdagítani (például: KSH-adatbázisból a bevételi adatokat szeretnék beilleszteni meglévő céges törzsadataink mellé),
- cégen belüli adatintegrációra kerül sor, mely során a több, egymással párhuzamosan vezetett rendszer törzsadatait konszolidálni kívánják,
- a marketingkampány során megcélzott ügyfelek listájából ki kell szűrni azokat a tételeket, akik már ügyfelek.

A legkomplexebbek a rendszerszintű hibák, azaz amikor a rekordok összessége jelent hibát. Hogyan lehet ezeket a hibákat is diagnosztizálni? Ha minden elvárást dokumentálni akarnánk, az a rendszerterv sokszorosát tenné ki. Ezért inkább referenciapontok szolgálnak iránytűként.

10.2.2. Referenciapontok

A referenciapontok olyan ismérvek, amelyekhez képest az adatok eltérése hibának minősül.

Az adattárház gyakran valamilyen rendszer kiváltására készül (ugyanazt az információt hatékonyabban szolgáltatva). Ilyenkor könnyű helyzetben vagyunk, mert az eredeti rendszerekhez viszonyíthatunk. A forrásrendszerek is képesek bizonyos jelentéseket előállítani.

Segít, ha a szakértők véleményét kikérjük a teszteredmények kiértékeléséhez. Mindig van néhány termék, ügyfél vagy indikátor, amely illeszkedik a korábbi tapasztalatokhoz. Ezért nagyon hasznos, ha a fejlesztőkkel közösen kijelölnek néhány konkrét kimenetet, amely értékének jelentős eltérése valamilyen adathibára utal.

Referenciapont az adatok önmagukhoz mért folytonossága, kezdve az olyan nyilvánvaló dolgoktól, mint egy extraktum szokásos mérete, rekordszáma. Számos statisztika, diagram és eloszlás furcsasága árulkodhat azonban valamilyen adathibáról. Ez pedig már adatbányász gondolkodásmódot igényel. Nemcsak deklarált üzleti szabályoknak kell megfelelni (minden betétesnek van számlája), hanem tapasztalati szabályoknak is. (Egy kereskedelmi egységben egy kis falu adta a vásárlások 3%-át, ami egy elemzés során tűnt fel. Kiderült, hogy a pénztáros mindenkinek, akinek nem tudta pontosan az irányítószámát, a saját falujáét írta be.)