Fatalin László

Adat-eszméletek

Az adat fogalmához alábbiakban vázolt némileg felszínes gondolattöredékek megerősítik bennem azt az érzést, hogy oktatásunkat tematikailag is intenzívebben kellene fejlesztenünk elsősorban a multidiszciplináris fogalmak irányába, még az „örök igazságokat" közvetítő matematika esetében is. Az alábbiakban érintett ismeretek szinte kivétel nélkül véges halmazokra és relációkra vonatkoznak, melyekkel matematikaórán legfeljebb elvétve találkozhatunk. Számos olyan fogalmat definiálunk (nemcsak matematikából), amelyekkel a hétköznapi életben az emberek döntő többsége sohasem találkozik. Tudom, hogy e fogalmak elsajátítása során szellemünk csiszolódik, de talán a fejek kiművelésekor mind didaktikai, mind egyéb szempontokból célszerű lenne a jól használható, praktikus fogalomrendszerek kialakítására nagyobb hangsúlyt helyezni. E feladat persze nehéz, hiszen a multidiszciplináris fogalmakhoz kötődő ismeretanyag és annak módszertani feldolgozása kiforratlan, de a várható eredmény reményében a „kontárkodás" vádját vállalva is érdemes próbálkozni.

A szakirodalom olvasása során egy-egy tanulmány gondolati mélységeibe alászállva gyakran tévelygek kavargó szakkifejezésekbe burkolt bölcsességek homályában. Valaha apám segített rendet teremtenem a bennem eluralkodó káoszban. Módszere irigylésre méltóan egyszerű és megbízható volt. Szakmai szempontból dilettánsnak minősíthető, ámde józan, lényegre törő kérdéseivel, észrevételeivel vezetett rá az éppen aktuális rendszer logikájának felismerésére. Ma már én bombázom kérdéseimmel lányaim közül az éppen közelemben levőt. A kérdésfeltevés nehézsége ilyenkor abban rejlik, hogy nem tetszeleghetek a tudálékos szakember pózában, nem bújhatok el a szakma kialakult védőpajzsa mögé, hanem hétköznapi magyar nyelven kényszerülök megfogalmazni gondolataimat magamnak is. A hosszú évek során e gyakorlatok egy kis szabályrendszert alakítottak ki. Nem afféle „Ki mit tud", avagy „Kérdezz-felelek" játékszabályok ezek. Magunk között „Mit jelent az a szó, hogy..." nevet adtuk e játéknak, utalva arra, hogy szinte bármilyen problémakörhöz közeledve, már a kezdetekben így fogalmazódnak meg kérdéseink.

Egykori cégemnél, az egyik nagy közszolgáltató vállalatnál egyszer azzal bíztak meg, hogy informatikai rendszerük műszaki alapadatbázisának kialakítását irányítsam. Sok okos instrukciót kaptam – hiszen már évek óta kísérleteztek e feladat megoldásával sikertelenül –, melyeket megpróbáltam szakmai ismereteimmel kiegészíteni. A szakszerűségbe vetett hitem, majd reményeim is szertefoszlottak, mivel legnagyobb megdöbbenésemre javaslataim rendre betonfalakba és -fejekbe ütközve ignorálódtak. Fokozatosan úgy elbizonytalanodtam, hogy egy gyenge pillanatomban tizenkét éves lányomhoz fordultam, a szokásos játékunkba az adat szót helyettesítve. Zongoraórájára rohantában egy pillanatnyi gondolkodás után a következő, meglehetősen zanzásított választ bökte felém: „Az adat azt jelenti, hogy valamiről valamilyen tényt tudunk. Van olyan adat, amit jobb

eltitkolnunk, általában adathiányban szenvedünk, egyébként meg nem foglalkozunk ilyen szavakkal az iskolában." Nyelvészkedő, banális és evidens válasza őt magát sem elégítette ki, ezért a későbbiekben még legalább tucatszor visszatértünk ehhez a gumicsontunkhoz is és próbáltuk megfejteni az adatfogalom rejtelmeit. Várakozásom szerint elvi szempontból eleve kudarcra ítélt vállalkozásba fogtunk, ugyanis meggyőződésem, hogy egy klasszikusan zárt elmélet keretébe éppen lényegükből fakadóan nem szoríthatók be a multidiszciplináris fogalmak. E metafizikai hipotézisem ezúttal sem rendült meg, ugyanakkor néhány apró eredmény egyvelege a szűkre szabott megértés örömével ajándékozott meg bennünket.

Egy kifejezés, egy fogalom használatakor az első lényeges tennivalónk, hogy lehetőleg minél pontosabban tisztázzuk jelentését. A hétköznapi fogalmakból az ún. explikációs folyamaton keresztül kristályosodnak ki az egzakt tudományos fogalmak, melyeket többnyire egy-egy definíció szűkös keretében próbálunk meg körülírni. A definíciók a mögöttes elméleti keretek ismerete nélkül általában meglehetősen üres jelentéstartalommal bírnak. Az adat fogalmával is ez a helyzet. A szótárak, lexikonok meghatározási mélységét meghaladó információkhoz a megfelelő szakirodalmi kutatások vezethetnek el bennünket. Az adat fogalmának elemzésével az utóbbi évtizedekben egyre több cikk foglalkozik. Az adatelmélet rohamos fejlődésének kulcsa abban rejlik, hogy átfogó módszereket dolgoz ki az adatgyűjtésre és értékelésre, valamint a különböző kutatási módszerekkel gyűjtött megfigyelések elméletileg és gyakorlatilag is egységes és logikus osztályozási rendszerezésére törekszik. Eredményei általánosak, a tudományágak bármelyikében használhatók, amivel hozzájárul a viselkedő rendszerekben alkalmazható „mérési eljárások" alapjainak megteremtéséhez, az empíriák pontosabb leírásához. Néhány klasszikus tudományágban, így például a fizikában ma már többnyire meglehetősen egyértelműnek tűnik az adat fogalma, míg a fiatalabb tudományágak (pl. fiziológia, pszichológia, szociológia, néprajz, közgazdaságtan, politika, szervezéstan stb.) még napjainkban is küzdenek az empíriák egzaktabb, adatszerű leírásával. Ezért nem meglepő, hogy e tudományágak szakemberei intenzívebben foglalkoznak az általános adatelmélet kiépítésével. Az adatelmélet, mint tudományközi jellegű elmélet első átfogó tárgyalása is a pszichológus *C. H. Coombs*-tól származik (1964-ből). Az adat fogalmára általa adott definíció nem kijelölő, hanem szabályozó jellegű, így első ismerkedésre nehezen megemészthető. Coombs felfogásában az inger–reakció viselkedőrendszerekre vonatkozó empirikus adatok nem a közvetlen viselkedést jelentik, hanem viszonylagosak, lényegük az ingerek és egyedek közötti relációban áll. E reláció az adat hétköznapi fogalmában is kifejezésre jut. Az adat mindig valamiről valamilyen tényt közöl, azaz valamilyen objektumhalmaz és egy tulajdonsághalmaz közötti relációt jelenít meg. E meghatározás formálisan egzakttá tehető ahhoz hasonlóan, mint a matematika halmazelméleti megalapozása, ugyanakkor szükségképpen fellép a tyúk és a tojás elsőbbségének dilemmája is, azaz a halmaz és a reláció prioritási kérdése. (Az oktatásban jelenleg a halmaz fogalmát tekintjük elsődlegesnek, bár már magának az alapfogalomnak a kialakításakor, a nulladik kritériumban is fellép az „elem-e" reláció, hiszen minden objektumról el kell tudnunk dönteni, hogy eleme-e a halmaznak vagy sem. A reláció fogalma a matematika tantárgyban később is elvétve, alapvetően csak rendezési, illetve ekvivalenciatípusként fordul elő.) E definíciós problémakör kísértetiesen hasonlít a mechanika klasszikus newtoni axiómarendszerének dilemmájára, miszerint a második axióma $F=m \cdot a$ Euler-féle, illetve az $F=dI/dt$ eredeti Newton-féle megfogalmazásában e törvény még egyszerű, kijelölő típusú definíciónak sem tekinthető, hiszen egy összefüggéssel két különböző mennyiségi fogalmat nem lehet meghatározni. (A tanítási-tanulási folyamatban persze ezt a rendszer szempontjából egyfajta a priori erő és tömeg, illetve impulzus fogalmának kialakításával hidaljuk át, így a második axióma már törvényként is felfogható.) A különböző didaktikai interpretációk több-kevesebb sikerrel teszik

emészthetővé pszichikumunk számára az ilyen típusú fogalomrendszerek dilemmáit. Coombs definíciójának feltétlenül érdeme, hogy nem kerüli ki e problémakört és rendszerszemléletéből adódóan a relációs komponensre helyezi a hangsúlyt.

Az objektumhalmaz és a tulajdonsághalmaz mögött valójában mindig valamilyen relációs struktúra húzódik meg. E relációs struktúrákban mindig tetten érhető a skatulyázási elv alkalmazása, azaz mind az objektumok, mind az egyes tulajdonságok élesen elkülönülnek egymástól. Ez tulajdonképpen az adat-egyértelműség nulladik feltétele. A skatulyázási elv pedig mindig egy ekvivalenciareláció feltételezését jelenti. Egy-egy tárgy színének megadásakor például a színtulajdonság mögött mindig található egy többé-kevésbé finom kidolgozású komódszerű „világkép", amelyben az egyes fiókok tartalma között nincs átfedés, ti. a piros, a kék, a zöld stb. fiókok tartalma élesen elkülönül egymástól. Az említett ekvivalenciareláció ilyenkor az „egyforma színű" reláció. Az egyformaság, az ugyanolyan, az egyenlőség szavak pedig értelemszerűen magukban hordozzák a tranzitivitási tulajdonságot, azaz $a=b$ és $b=c$ esetén az $a=c$ teljesülését is. A valóságban persze a helyzet lényegesen bonyolultabb, hiszen a nem érzékelhető, lényegtelennek tűnő apró eltérések felhalmozódása következtében a tranzitivitási szabály, az egymással helyettesíthetőség elve csak korlátozottan teljesül. Minden „színhez" tartozik ugyan egy-egy csoport, amely a vele „egyszínű" objektumokat tartalmazza, ugyanakkor tudjuk, hogy e csoportok nem különülnek el élesen egymástól, a határvonalak elmosódnak, átfedések vannak közöttük. Matematikailag ezt úgy fogalmazhatjuk meg, hogy a meglevő toleranciarelációt ekvivalenciával helyettesítjük. E közelítés gyakorlati szempontból megfelelő, ha az ekvivalencia kellően finom diszjunkt felbontást eredményez. Bizonyos értelemben meglepő, hogy a tudományos igényességű fogalommeghatározásokban direkt vagy indirekt formában az egyenlőség mindig manifesztálódik, hiszen ez egyben az emberi megismerés közelítő jellegét is garantálja. Gyanítom, hogy ennek hátterében biológiai felépítésünkből adódó okok is állnak. (Szervezetünk 10^9–10^{10} bit/s számú ingert fogad, amiből feldolgozásra csak 20–150 bit/s kerül. E több nagyságrendi különbség felhívja a figyelmet arra, hogy e nagyszámú inger szűrése és tömörítése, majd a feldolgozást követő kimeneti reakciók produkálása – melyek száma ismét kibővül a testi effektorok esetén 10^3–10^7 bit/s-ra –, csak kellően adekvát, megfelelően strukturált rendszerben valósulhat meg.) Az adatelmélet azon fejezetei, amelyek az adatfogalom tartalmi meghatározásával foglalkoznak, szükségképpen óriási nehézségekkel küzdenek. Multidiszciplináris megközelítésükből adódóan pedig a tárgyalási módnak még sematikusnak is kell lennie. E problémakör megoldásában valószínűleg a toleranciarelációk kezelési módszereinek pontosabb kidolgozása és alkalmazása jelent majd előrelépést.

Az adatosztályozási eljárások a fogalmi meghatározásnál lényegesen jobban kidolgozottak, hiszen e formális megközelítések esetén az adat és az informatikában használatos, általános definícióval szintén nem rendelkező információ fogalma egybeesik. Kiindulási alapként ilyenkor a kódolás-dekódolás és az átvitel problémaköre szokott előtérbe kerülni. Alaphipotézis gyanánt meglevőként elfogadunk egy, az adategyértelműség feltételeként megfogalmazott skatulyázási elvnek megfelelő struktúrát, melyet a kódolás során az ún. jelkódkészlettel próbálunk meg ábrázolni. Az ábrázolás fogalma persze itt is kétértelmű, hiszen az alkalmazott jel(sorozat)halmaznak van jel és jelentés értelmezése is.

A jelek mint valamilyen fizikai mennyiség időbeli lefolyásának absztrakt felfogásában eltekintünk a jel anyagi-energetikai hordozójától és azt az általános jelszinttel mint szimbólummal helyettesítjük. A jelek egyik szokásos osztályozási szempontja szerint megkülönböztetjük azokat a jeleket, amelyek jelszintje csak diszkrét, illetve tetszőleges értékeket vehet fel. Az amplitúdó- és időtartományban is diszkrét jelek a digitális jelek, a folytonosak pedig az analóg jelek csoportját alkotják. A jel továbbításakor szükségképpen fellépő fizikai zavaró hatások, az ún. zajokkal szemben a digitális jelek természetükből adódóan lényegesen jobban védettek, illetve védhetők, sőt még az is elérhe-

tő, hogy a mintavételezési és kvantálási törvények betartásával a forrásoldalon elvégzett analóg-digitális, majd a felhasználónál a digitális-analóg átalakításból eredő hiba kisebb legyen, mint az analóg jel továbbításában fellépő hiba, ezért a digitális jelek egyre nagyobb teret hódítanak az alkalmazásokban. A kétállapotú rendszerek változatos fizikai realizálhatóságának köszönhetően többnyire a bináris rendszerek terjedtek el. (A jeleket persze más szempontok szerint is csoportosíthatjuk, a mérés- és irányítástechnikában általánosan használatos például a determinisztikus-szochasztikus és a bemeneti-kimeneti jelfelosztás is.) A digitális, illetve digitalizált jelek alkalmazásához illeszkedik az információtartalom jelenleg elfogadott definíciója is. Ez a jel(sorozatnak) mint ábrázolásnak a jelentés értelmezését próbálja megragadni és az információt mint tudásnyereséget, illetve megszüntetett bizonytalanságot számszerű formában, mennyiségként definiálja. (Meghatározásának módszere hasonlít a valószínűségi változó matematikai definiálásához, az információtartalom ebből a szempontból a valószínűségi változó analogonjának tekinthető.) Szemléletmódja alapvetően statisztikus, hiszen a kódszótár alapelemeinek, a kódábécének, azaz valamilyen $A=\{a_1,a_2,\ldots, a_n\}$ jelkészletnek a valószínűségi eloszlásából indul ki. Így minden a_i jelhez tartozik egy p_i valószínűség, átlagos előfordulási gyakoriság. A definíció szerint az a_i kibocsátásával a közölt információtartalmat az $I(a_i)=\text{ld } 1/p_i$ mennyiség határozza meg, ahol *ld* a kettesalapú logaritmust jelöli. Az információtartalom mértékegysége a bit, ami nem azonos a bináris rendszerekben honos ún. jelbittel, mely értéke szükségképpen egész szám. A teljes kódábécére kiterjesztve e fogalmat az információtartalom várható értéke a $H(A)=\sum_{i=1}^{n} p_i I(a_i)=-\sum_{i=1}^{n} p_i ld(p_i)$ alakba írható, amit a termodinamikai analógia alapján entrópiának nevezünk. Igazolható, hogy tetszőleges jelkészlet esetén az entrópia akkor maximális, ha a jelkészlet valamennyi elemének előfordulási gyakorisága egyforma, azaz n elem esetén $I(a_i)=H(A)=ld\ n$. A tízes számrendszerben mind a tíz számjegy előfordulási gyakorisága 1/10, így az egyes jelek információtartalma és a jelkészlet entrópiája is $I=H=ld10=3,322$ *bit*. (E megállapítás némileg pontatlan, mert nem veszi figyelembe a tizedesvesszőt és az előjelet.) Az európai írott nyelvek jeleinek átlagos információtartalma $I=H=ld\ 30=4,9$ *bit* lenne, ha minden betű azonos valószínűséggel fordulna elő, míg a különböző előfordulási gyakoriságok következtében az entrópia valójában csak 4 bit körüli érték. A kódszótár entrópiájának vizsgálatakor általában figyelembe kell vennünk, hogy egy újabb szimbólum kibocsátási valószínűsége függ az előzetesen kiadott szimbólumoktól is, azaz a rendszerben bizonyos memóriajelleg is található. Ezt a jelenséget a feltételes valószínűség segítségével vehetjük figyelembe a kódszótár entrópiájának meghatározásában. A szövegelemzések azt mutatták ki,

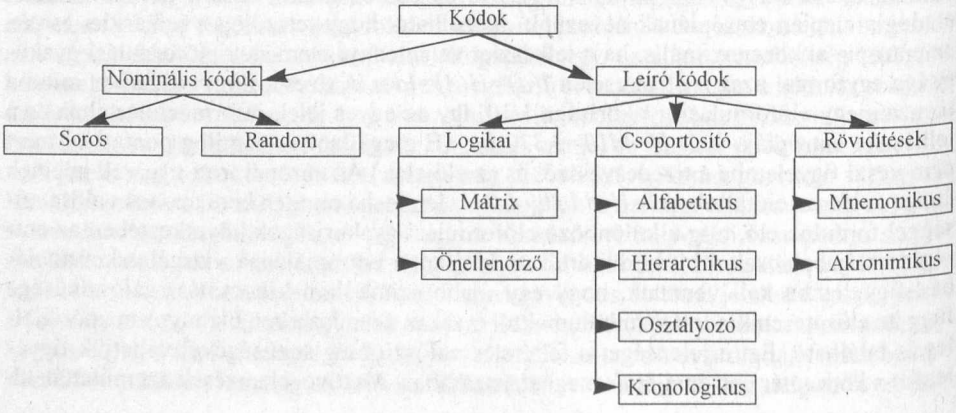
Az adatosztályozási eljárások a fogalmi meghatározásnál lényegesen jobban kidolgozottak, hiszen e formális megközelítések esetén az adat és az informatikában használatos, általános definícióval szintén nem rendelkező információ fogalma egybeesik. Kiindulási alapként ilyenkor a kódolás-dekódolás és az átvitel problémaköre szokott előtérbe kerülni. Alaphipotézis gyanánt meglevőként elfogadunk egy, az adategyértelműség feltételeként megfogalmazott skatulyázási elvnek megfelelő struktúrát, melyet a kódolás során az ún. jelkódkészlettel próbálunk meg ábrázolni. Az ábrázolás fogalma persze itt is kétértelmű, hiszen az alkalmazott jel(sorozat)-halmaznak van jel és jelentés értelmezése is.

hogy az európai írásbeliség entrópiája 1,5–2 bit között van. Ez egyben azt is jelenti, hogy az írásbeli üzenetek jelentős része nem hordoz információtartalmat, azaz redundáns. E viszonylag magas redundancia következtében egy szöveg többnyire még akkor is olvasható marad, ha minden második betű hiányzik. A matematikai képletek estében a redundancia sokkal kisebb, azaz a matematika kódrendszere lényegesen tömörebb, s megfejtése is nagyobb figyelmet igényel. A jó kódolással szemben támasztott követelményrendszer ellentmondásos. Az első feltétel, hogy egyértelműen lehessen dekódolni. Ehhez a kódolásnak ki kell elégítenie az ún. Fano-féle feltételt, mely szerint egyetlen kódszó sem egyezhet meg egy másik kódszó kezdetével. (A Morze-rendszerbe ezért kellett felvenni a szünetjelet is.) A kódolásba a tömörség rovására is célszerű beépíteni bizonyos redundanciát, hiszen csak így biztosítható, hogy a kódrendszeren keresztül folyó kommunikáció nagy megbízhatóságú legyen.

A kódszókészlet hibakorlátozásait vizsgálva kerül előtérbe a Hamming-féle távolságfogalom, amely két kódszó távolságán azon helyértékek darabszámát érti, amelyekben a két kódszó különbözik egymástól. A kódszókészletben fellelhető legkisebb távolságot a kódszókészlet Hamming-féle távolságának nevezzük és a továbbiakban *D*-vel jelöljük. A hibakezelésben megkülönböztetünk két módszert: a hibafelismerés csak hibajelzést generál; a hibafelismerés lehetővé teszi, hogy a hibajavítást közvetlenül is elvégezhessük. Ha legfeljebb *d* bitben bekövetkező hibát kívánunk felismerni és ezek közül legfeljebb *e* bitben bekövetkező hiba javítását is lehetővé tevő hibafelismerést kívánunk a kódrendszerben biztosítani, akkor a $D \geq d+e+1$ feltételnek kell teljesülnie. (A számítástechnikában változatos realizálási módot alkalmaznak, találhatunk pl. paritáselemes kódokat, arány-kódokat, ciklikuskódokat stb.) A gyakorlatban inkább tudatosan növeljük a redundanciát a tömörség rovására is, részben a megtanulhatóság, részben a hibakorlátozó hatása miatt. A kódolás a Shannon-féle kódolási törvények alapján egzaktabb módon is tárgyalható.

Az információtartalom fogalmának fentebbi kialakításában még számos nyitott kérdés van, aminek következtében e modellrendszer használhatóságának korlátait jelenleg nem tudjuk körvonalazni. Ismeretes, hogy a képi ábrázolások (pl. műszaki rajz) lényegesen tömörebb információhordozók, mint szöveges leírásuk, és az információtartalom fogalma inkább a verbális kommunikáció szekvenciális leírásához áll közelebb. A számítógépes grafikai rendszerek fémjelzik, hogy a képi ábrázolások, sőt még a mozgókép-megjelenítések is realizálhatók e fogalmi keretben, bár csak óriási memória- és sebességigény mellett. A különböző grafikai szoftverek intenzív fejlesztése és rohamos elterjedése során szerzett tapasztalatok előmozdítják az információtartalom fogalmának árnyaltabb megítélését, pontosabb újraértelmezését is.

Számítástechnikai megközelítésben az *adat=objektum+tulajdonság* definíció az *adat=objektum kódhalmaz+ tulajdonság kódhalmaz* típusú leírásként jelenik meg. A kódok osztályozására az ISO szabványtervezete a hetvenes években vázlatosan a következő felosztást javasolta:



E felosztás némileg mesterkélt, az egyes kódosztályok itt sem különülnek el egymástól. Az adatelméletben, különösen az adatgyűjtő és értékelő módszerek szempontjából a gyakorlatilag használható változó típusok osztályozási rendszere a mérvadóbb. A különböző szoftverek e tekintetben némileg eltérnek egymástól. Általában a következő egyszerű felhasználói típusok megengedettek: karakteres (Character, String,...); numerikus (Integer, Long, Real, Single, Double,...); logikai (Logical). Gyakran külön deklarálható dátum (Date), esetenként pedig halmaz, grafikus vagy egyéb speciális változó típusok is.

Egy adathalmaz mint *objektumhalmaz* × *tulajdonsághalmaz* reláció klasszikus leírási módjaként kínálkozik a relációtáblával való megjelenítése, ugyanakkor ennek közvetlen megadása óriási memóriaigénnyel bír. A memóriaigény jelentősen csökkenthető, amennyiben az *objektum* × *tulajdonság* reláció függvénnyel, azaz egyértelmű hozzárendelésként is leírható. Ilyenkor a tulajdonsághalmaz megfelelő kódolásával az egyszerű függvénytáblázatos megadási módot követhetjük. (A memóriaigény minimuma egy $n \times k$ méretű reláció megadásához az első esetben $2n \times 2k = 2n+k$ jelbit, míg a második esetben $k \times ((ldn(+1)$ jelbit.) Az ún. interpolációs táblázatokkal tovább csökkenthető az adathalmaz ábrázolásának helyigénye. (Az interpolációs táblázat tulajdonképpen az egyik őse az adatállományok indexelési eljárással való tömörítésének.) A tulajdonság kódolásával tetszőlegesen járhatunk el, a gyakorlatban előszeretettel használunk számokat kódként, aminek előnye a nyelvtől való függetlenségében és egyértelműségében rejlik. Ettől azonban e tulajdonság nem válik mennyiséggé, hiszen a nominális és rangsorskálához csak nominális kód képezhető. Ilyenkor célszerű a tulajdonságot (pl. azonosítók, cikkszámok...) továbbra is karakteres típusként deklarálni, hiszen a numerikus típusok körében elvégezhető műveleteknek semmilyen értelme sincs. A numerikus változótípusokat az intervallum- vagy arányskálával rendelkező tulajdonságokhoz fejlesztették ki.

Azonos objektumokra vonatkozó különböző függvénytáblázatok egyszerűsített megadására használható az összevont táblázat, az ún. blokk vagy tömb. Ebben a különböző tulajdonságokhoz tartozó kódok azonos típusú változóval egységesen vannak deklarálva. Megadásukhoz a tömb dimenzióját és típusát kell meghatározni. Az általánosabb táblázatnak mint strukturált adathordozónak mindig meg kell adni a részletesebb leírását is. A hétköznapi életben ez többnyire egy fejléccel történik, amelyben megadjuk, hogy az egyes oszlopokban milyen tulajdonság kerül megjelenítésre. Számítástechnikai alkalmazásokban a kódolás típusát is deklarálni kell, amivel kialakul egy ún. rekordszerkezet és a táblázat egyes soraihoz szimbolikus objektumként egy sorszám (record pointer) is csatolódik. Az azonos oszlopban levő cellák, mezők természetesen azonos típusúak, hiszen egy adott tulajdonsághoz tartozó kódokat jelenítenek meg. Egy adathalmaz n-oszlopos táblázatos megadása egy n-változós relációnak tekinthető. E felfogás használati előnye, hogy adatábrázolása emberközeli és áttekinthető, ami által egyszerűsödik az adatkezelés is.

Különböző táblázatokat ún. indexállományokkal összekapcsolhatunk és így eredőben újabb táblázatokat állíthatunk elő. Ennek hátterében a relációk közötti kompozíció művelete áll. Az indexelési eljárás sok esetben az adathalmaz tömörítettebb ábrázolását teszi lehetővé. Egy vállalat éves fizetési listájának előállítása esetén a táblázatban minden dolgozó minden havi fizetéséhez kötődő adatának szerepelnie kell. Így a dolgozókhoz kötődő statikus adatok (név, beosztás, dolgozók száma) legalább tizenkétszer szerepel, azaz az adathalmaz ábrázolása redundáns. Lényegesen egyszerűsödik a helyzet, ha a dolgozók statikus adatait egy külön táblázatban, állományban tároljuk és egy kulcsmező, például a dolgozók számának segítségével kapcsoljuk ezen adatokat a kifizetési állományhoz. Indexelési eljárással egy alaptáblázatból különböző rendezettségű állományokat is létrehozhatunk memóriatakarékos módon. Egy könyvkatalógust készíthetünk cím, illetve szerzői név szerint rendezetten is. Ez tulajdonképpen két ugyanolyan méretű táblázatot jelent. Gazdaságosabb és gyorsabb megoldás, ha az egyik meglevő állományhoz egy kulcsmező megadva (l. cím szerinti sorrendben a rekordmutatóhoz) a szerző index-

állományban a szerzői névsormutató mellett a cím szerinti sormutatót jelenítjük meg. Az adathalmazok óriási mérete miatt a tömörített ábrázolásnak nem pusztán a kisebb memóriaigény elérése a célja, hanem az ezzel együtt jelentkező műveleti időigény lényeges csökkenése is. Természetes módon felvetődik a kérdés, hogy egy óriási adathalmazt hogyan célszerű táblázatokra és indexállományokra bontani. E problémakör klasszikus, némileg elavult heurisztikus megoldási módjaiban az adatokat különböző formális szempontok szerint osztályozták (input-output; statikus-periodikus-sztochasztikus stb.) és az egyes adatosztályokhoz dolgoztak ki módszereket. (Ilyen általánosan elterjedt módszer például a statikus típusú adatosztályra az ún. változások naplózásának a módszere.) A feladat egzakt, korszerű megoldásához azonban csak az adat-funkció reláció matematikai elemzése vezethet el.

Az eddigiekben az adatbázishoz tartó funkciókról, az adatkezelésekről nem volt szó érdemlegesen, pedig az adatra nyilván nem kincsként, hanem tőkeként van szükségünk, azaz nemcsak tárolni, hanem használni is kívánjuk az adatokat. Az adatbankok kialakítása ennek megfelelően történt, ami formálisan az *adatbank=adatbázis+adatkezelő nyelv* egyenlettel írható le. Az adatkezelő nyelvek a legáltalánosabb funkciókat realizálják. A relációs adatbázis-kezelés területén az egyik és egyben legjelentősebb szabványosítási törekvésként az SQL (Structured Query Language = Strukturált Lekérdező Nyelv) valósult meg, melynek fejlesztése a hetvenes évek közepén kezdődött az IBM égisze alatt. Ahogyan a relációs adatbázis-szerkezet meglehetősen rugalmas kereteket biztosít az adathalmazok tárolásához, ahhoz hasonlóan az SQL mint eljárásmentes nyelv is rugalmas kereteket nyújt az adatkezelési funkciók realizálásához. Az adatbázis-kezelés alapfunkciói között az adatbázisok létrehozása, karbantartása és szinte tetszőleges lekérdezhetősége mellett az adatbiztonságra is gondot kell fordítani. A hálózatok elterjedése, melynek legfőbb előnye az adatok megoszthatóságából fakad, különösen éles követelményként veti fel az adatok sérthetetlenségének kérdését. Ezt egy *adathozzáférési reláció x adatvédelmi jellemző reláció* eredő relációjával szokás megadni. (E fogalomkör is aránylag jól kidolgozott a számítástechnikában, amiben talán csak az a meglepő, hogy a világ első adatvédelmi törvénye csupán 1970-ben, Hessen német tartományban lépett életbe, pedig a személyes jellegű adatok védelmének már történelmi hagyományai vannak, pl. gyónási, illetve orvosi titok.) Az adatbiztonság fokozásához különböző technikai módszerek kerültek kidolgozásra (pl. biztonsági tartalékmásolatok; visszaállítási tartalékállomány; illetéktelen másolások ellen: speciális kódolások, önmegsemmisítések, spirális sávok kialakítása stb.), amelyekkel csak több-kevesebb sikerrel lehet az adatvédelmet biztosítani, ezért újabban az általános jogi védelmet is kiterjesztették mind a személyi adatokra, mind a szoftver termékekre.

Napjaink divatos és jól fizető üzleti vállalkozása a különböző szintű (vállalati) informatikai rendszerek létrehozása. A feladat elvégzéséhez rendelkezésre áll egy rugalmas kereteket biztosító adatbankszerkezet (a piacon több ilyen is található), és ezt kell a felhasználó igényeihez hozzáidomítani. E feladat nehézségének egyik oka, hogy a felhasználói kör multidiszciplináris szempontból nagyon alulképzett és ezt a vezető beosztásúak különösen nem szívesen vallják be, inkább tekintélyüket óvandó, lokális okoskodásokkal szakmai tudásukra hivatkozva hátráltatják a munka szakszerű elvégzését. A probléma másik oka, hogy mind az adatbázis-struktúra, mind a funkcióstruktúra adekvát kialakítása csak a véges relációk matematikai elméletében kidolgozott módszerekkel végezhető el megnyugtató módon. Ezen ismeretekkel aránylag kevés vállalkozó rendelkezik, így e piacon sok olyan kétes referenciákra hivatkozó, üzletileg sikeres amatőrrel találkozhatunk, akik üzleti vállalkozásaik során próbálják meg e szakmát autodidakta módon megtanulni. Találkoztam olyan informatikai céggel, amelynek prominens képviselői szerint elég az adathalmaz előállítása, a funkciók számukra nem is fontosak, pedig szakszerű tervezéskor mindig az adat–funkció reláció alapján kell mind az adat-, mind a funkcióstruktúrát

kialakítani. Az elemzés során a funkcióhoz kötődően az adathalmazra egy toleranciareláció adódik amelynek toleranciaosztályai alkotják az egyes alrendszereket és ezen osztályok közös részei jelölik ki az alrendszerek egymáshoz kapcsolódását. A funkciókra mint az adatok duálisára ugyanez elmondható. Ilyen módon elkészíthető az informatikai rendszer vázlata. Az adat-funkció reláció finomszerkezetének vizsgálatával a részletes rendszerterv is kidolgozható. E rendszerszemléletű megközelítés esetén természetesen egy sor okoskodó kérdés fel sem vethető. Az egyik ilyen tipikus problémakörként az adathiány–túlzott adatigény dilemmája szokott előkerülni, amit a dilettánsok homályos számítástechnikai lehetőségekre hivatkozva oldanak fel. Az adat–funkció struktúra ismeretében persze azonnal láthatók azok az adatok, amelyekhez értelmes funkció nem kötődik, és az is kiderül, hogy mely funkciók nem valósíthatók meg adathiány következtében. Szakszerűség hiányában ezen informatikai rendszerek beruházási költsége sohasem térül meg, sőt az adatbeviteli igények miatt a vállalat munkaerőigénye is megnövekedhet. Az utóbbi időben ezt a folyamatot sajnos egyre több tapasztalat is megerősíti. Ennek oknyomozata különösen a közszolgáltató szférában a tulajdonos és a menedzsment eltérő érdekeltségére vezet vissza, hiszen egy jól működő informatikai rendszer a vállalat jobb átláthatóságát biztosítja, ami a menedzsmentnek kisebb mozgásteret enged.

Az adatgyűjtés és értékelés módszereinek tárgyalása meghaladja e cikk kereteit, hiszen a felmérés–értékelés komplex kifejtése még felszínes megközelítésben is legalább ekkora terjedelmet igényelne. Talán érdemes meggondolni, hogy az itt vázolt gondolat- és ismerettöredékekből mennyi és milyen szétszórtságban szerepel a közoktatásban, jóllehet hétköznapi szempontból is kétségtelenül hasznos fogalmak kicsit tudományosabb színű, szemléletformáló megközelítéséről volt szó.