szoftverszempontból olyannyira rugalmas, hogy a komplex együttműködésnek az egész magyar könyvtári horizontot beleértve sincsen számottevő akadálya.

A világ bármely ma számítógépesedő könyvtára számolhat azzal, hogy az egész Internet-közösség eléri OPAC-ját. Ez ma több millió felhasználót jelenthet. A különféle metakommunikációs eszközökön keresztül pedig „gyerekjáték" bármely könyvtárra rátalálni. Egy távoli OPAC pedig képvisel valamit magán túl is: egy intézményt, egy várost, vagy akár egy országot is. Ez a nyilvánossági fok óriásira növeli a rendszertulajdonosok felelősségét.

Az „egyetlen rendszer" vitát lezárva ma már arra kell törekedni, hogy viszonylag kis számú, a feladatokhoz jól illeszkedő, kedvező feltételek és garanciák mellett szoftver kerüljön minél több magyar könyvtárba.

A Networkshopban összejött könyvtárosok, könyvtári informatikusok – úgy tűnik – hasonlóan gondolkodnak, nyilvánosan kimondva (lásd a megfogalmazott projektötleteket), vagy a folyosói beszélgetések szintjén is megfogalmazódtak teendőink. A rendszerek összehangolását megkönnyítendő (1) gyakorlati egyezségre kell jutni a használni óhajtott magyar karakterkészletet illetően, valamint (2) a tényleges adatcsere céljainak megfelelő, gyakorlati felhasználásra alkalmas HUNMARC ügyében. (3) Tisztázni kell a Nemzeti Könyvtár és mások által nyújtott gépirekordszolgáltatás helyzetét, és (4) igyekezni kell egy legalább lekérdezési szinten elérhető – s egyre több könyvtári OPAC-ot tartalmazó – „központi lelőhely-katalógust" létrehozni, akár valóságos adatbázisban, akár lekérdező felület szintjén. (5) Rövid távon üzembe kell helyezni egy online Nemzeti Periodika Adatbázis rendszert, amelyet egyre több könyvtárnak kell online „töltenie" is. Erre épülhetne később a magyarországi föllelhetőségű folyóiratcikkek könyvtárközi rendszere, akár az amerikai UNCOVER mintájára. (6) A Magyar Elektronikus Könyvtár projekt kapcsán tisztázni kell az elektronikus szöveg feldolgozási problémáit, a copyright kérdésétől a nyilvántartáson át az indexelésig. (7) Újonnan kialakuló információs adattáraink nem lesznek jól elérhetőek, ha a korszerű metainformációs eszközök (Gopher, WWW, WAIS, Mosaic stb.) hazai működését nem hangoljuk össze, ill. ha ebben a koordinációs munkában nem veszünk részt. E munkákban fokozottabban együtt kell működnünk a számítógépes hálózatosokkal, mint olyan szakembereknek, akik elsősorban a hálózaton megjelenő információk tartalmi vonatkozásaiért felelősek.

Mindebből persze az is következik, hogy nem szabad hagyni, hogy a hálózathoz értő könyvtárosok „maguktól" legyenek. Az új követelményeknek megfelelően át kell alakítani a szakmai képzést az „egyszerű" könyvtárosképzés szintjén is, de főként a könyvtári informatikus képzés keretében. Posztgraduális szinten törekedni kell a speciális rendszergazda, a system's librarian típusú könyvtáros-számítógépes szakemberek kiképzésének megindítására is. A meglévő szakemberállomány rendszeres át- és továbbképzésére is meg kell találni a lehetőségeket, hiszen a kihívás óriási, mivel a „hagyományos, konzervatív" tudást egy új ismerethalmazzal kell összeegyeztetni, és – néha – összebékíteni.

Bakonyi Péter azzal fejezte be a Networkshop '94 zárszavát: „Jövőre reméljük találkozunk a Networkshop '95-ön!" Könyvtárosként ehhez azt tehetjük hozzá, hogy nekünk is ott a helyünk, mégha kezdetben idegenül mozogtunk is kicsit, hiszen ma már világosan látszik, a hálózati információtárolás, -feldolgozás és -visszakeresés ugyanúgy a mi területünk marad, mint a szép régi „papíralapú világban".

Kokas Károly
(JATE Egyetemi Könyvtár)

Csontváz van a szekrényben: adatbázisok hibái*

Bevezetés

A legjobb online és CD-ROM szakemberek jó ideje panaszkodnak a sok adatbázisban megtalálható minőségi hibákra. Jobb minőség-ellenőrzést sürgetnek, és gondos vizsgálatokon alapuló esettanulmányokat közölnek illusztrációképpen. Főleg a pontatlanságokra és következetlenségekre összpontosítanak: az előírásokra, a következetlen helyesírásra, a hibás adatokra, arra, hogy egyes adatmezők szemétládaként szolgálnak, befogadva minden olyan adatelemet, amely a többi adatmezőbe nem illik bele.

Kevesebb szó esik a hasonlóan fontos, de még kellemetlenebb következményekkel járó láthatatlan hibákról, az adathiányokral. Ilyen hiányról beszélhetünk, ha egy gyakran használt, biztosnak tekintett adatelem (a kiadás éve, a dokumentumtípus, a dokumentum nyelve, osztályozási jelzete stb.) a rekordok számottevő részéből hiányzik. Az ilyen hiányok gyakran releváns rekordok elveszítését eredményezik a keresés

* E kétrészes cikkel *Jacsó Péter*, aki a cikk leadásakor a University of Hawaii vendégdocense (visiting associate professor) volt, elnyerte az 1993. évi *Excellence in Writing* díjat. A díjat a University Microfilms International (UMI) cég alapította, és évente adják ki az információs szakma legjobbnak ítélt publikációjára. A díj átadására minden év decemberében Londonban, az International Online Information Meeting bankettjén kerül sor.

során, máskor félrevezető és drága eredményre vezetnek a találatok rendezésekor. Az online keresés egyik szépsége, a keresés különféle szempontok kombinációjával történő finomítása otrombaságba torkollhat azzal, hogy a hiányos rekordok rejtve maradnak.

Ez a cikk arra tesz kísérletet, hogy olyan keresési megoldásokat és trükköket gyűjtsön össze, amelyekkel „felfedezhetjük a csontvázat a szekrényben". Az egyik cél az, hogy felkészüljünk a defenzív keresésre, a másik cél arra ösztönözni az adatbázis-értékelések szerzőit, hogy értékelésükbe ilyen típusú vizsgálatokat is iktassanak bele. Az online keresések magas költségei sok módszer használatától visszariaszthatnak bennünket, a CD-ROM adatbázisok használattól független költségei azonban bátorítólag hathatnak.

I. rész: HIÁNYOK

Hogy egy adatmező meglétére vagy hiányára hogyan kereshetünk, az függ a keresőrendszer sajátságaitól, az adatelemek indexelési módjától, és az adatbázis-készítő előírta konvencióktól.

A lehetőségek terén az egyik végletet a DIALOG online és CD-ROM keresőrendszere jelenti, amelyben az adatmezők többsége prefix indexelésű, és ezekben mód van a **PY=?** típusú teljes csonkolásra, így könnyen meghatározhatjuk, hogy egy-egy adatmező hány rekordban található meg.

A másik végletre az EBSCO adatbázisok többsége szolgálhat például *(Magazine Article Summaries, Academic Abstracts, Facts on File),* amelyekben a teljességet nem is vizsgálhatjuk, mert alig van mezőspecifikus index, teljes csonkolásra nincs mód, a kikereshető találatok számát pedig a szoftver (legalábbis annak 1992 őszén élő változata) 10 000 rekordra korlátozza.

A legtöbb keresőrendszer lehetővé teszi a teljesség vizsgálatát, legalábbis egyes mezőkre. Ezek azután jelzésül szolgálhatnak arra, hogy milyen teljességet remélhetünk a többi adatmezőtől. Nem feledkezhetünk meg persze arról, hogy egyes adatmezők jogosan hiányozhatnak. Nem minden dokumentumnak van szerzője, nem minden folyóirat rendelkezik ISSN-számmal. Ha az elsődleges dokumentumon nem szerepel a kiadás éve, az adatbázis-készítő vagy meg tudja azt határozni, vagy sem. (Az utóbbi esetben persze betehet egy speciális kódot, jelezve a hiányt.) Egyes adatbázisokban jogosan hiányzik a nyelv, ha a dokumentum angolul van, ilyen például az *ERIC.* Az *NTIS* kinyilvánítja, hogy a kiadás országát kihagyja, ha ez az ország az USA. Ha azonban sok adatrekordból hiányoznak például a deszkriptorok, a SIC-kódok vagy a dokumentumtípus, a hanyagságra utal.

Az összrekordszám meghatározása

A teljességi vizsgálat alapja annak a meghatározása, hogy hány adatrekordot tartalmaz összesen az adatbázis. Ez adja az összehasonlítás alapját az összes további eredmény számára. Ez egyszerűen hangzik, de a valóságban nem mindig az. Az adatbázis dokumentációja és a reklámanyagok csak közelítő adatot nyújtanak, az is sokszor elavult.

Az ideális megoldás az, amelyet a *Computer Select* adatbázishoz használt Bluefish keresőrendszert nyújt. Ez egy bevezető képernyőn az ötrészes adatbázis minden egyes szekciójáról megadja, hogy az az adatbázis adott változatában hány rekordot tartalmaz.*

Más adatbázisokban az aktualizálás adatmezőből kereshető ki az összrekordszám. Ez az adatmező (jele rendszerint **UD**) azt tartalmazza, hogy mikor építették be az adott rekordot az adatbázisba, és rendszerint automatikusan generálja az adatbázis-építő szoftver. Néhány példa az összrekordszám keresésére:

Dialog OnDisc	**S UD=?**
Wilsondisc	**F(DA) 8: OR 9:**
SPIRS PsycLIT	**F UD=0000–9999**
SPIRS LISA	**F DA>0**

Nem minden adatbázisban találunk aktualizálás adatmezőt. Egyes ilyen esetekben célhoz jutunk más adatmezővel. Például a *Books in Print* vagy az *Ulrich's Plus* CD-ROM változatában eredményes a **KW=$** keresés, hiszen legalább egy kulcsszó minden adatrekordhoz tartozik.

Prefixes keresés teljes csonkolással

A DIALOG valószínűleg sok szempontból a legjobb szoftver, így a tesztkeresés szempontjából is. E szoftver használatának kellemetességét megkétszerezi, hogy ugyanazokat a lehetőségeket találjuk az online és a CD-ROM-változatban. Ebben a keresőrendszerben nagyon egyszerű a tesztkeresés, mivel lehetőségünk van a mezőnkénti keresésre. A kötelező adatelemek prefixszel kereshetők, pl. **LA=, PY=, DT=**. Az ilyen adatmezőkben teljes csonkolással is kereshetünk, vagyis minimális szótőt sem kell megadnunk. A *COMPENDEX* adatbázis online változatában például a következő eredményt adja egy ilyen keresés (a második oszlop a találatszám):

S1	**2820049**	**UD=?**
S2	**2819453**	**LA=?**
S3	**2805603**	**PY=?**
S4	**1436969**	**DT=?**
S5	**903928**	**TC=?**

A nyelv (**LA=**) mező hiánya nem egészen 600 adatrekordból nem jelentős az adatbázis többmilliós mérete mellett. A kiadás évének (**PY=**) hiánya több mint 14 000 rekordból már komolyabb probléma. A dokumentumtípus (**DT=**) 50 százalékos és a megköze-

* Az eredeti cikk minden példát, minden keresési módszert a képernyő tartalmát bemutató ábrával vagy ábrákkal illusztrál. Sajnálatos, hogy a tömörítvény szűk terjedelmi kereteibe ezek az illusztrációk nem férnek bele. – A ref.

lítési mód (**TC=**) 68 százalékos hiánya azt sugallja, hogy ezeket a mezőket óvatosan kell kezelnünk. Bár a nyomtatott dokumentáció jelzi, hogy a konferencia-előadások rekordjain és az 1985 előtti rekordokon nem található megközelítési mód (**TC=**) mező, ez sem igazolja a hiány nagy mértékét.*

Hasonló technikát követhetünk az OptiWare keresőrendszerben, így például a *Books in Print Plus*, az *Ulrich's Plus* és a *PAIS* adatbázisokban, valamint a nemzeti bibliográfiákban, legalábbis a szöveges adatmezőkben.

Aritmetikai keresés

Az aritmetikai műveletekkel aszerint kereshetünk, hogy egy mező tartalma kisebb vagy nagyobb-e egy megadott értéknél, vagy egy meghatározott intervallumba esik-e.

Ezt a megoldást használhatjuk például a SilverPlatter adatbázisokban. Így a *PsycLIT* adatbázisban **UD>0** és **PY >0** kereséssel egyaránt 333 920 találatot kaptunk, az adott időpontban ez volt az összrekordszám.

Az OptiWare keresőrendszert használó adatbázisokban ilyen módon kereshetünk egyes mezőkben, majd az eredményt a teljes csonkolásos prefixes kereséssel kapott összrekordszámhoz hasonlíthatjuk. Amikor az *Ulrich's Plus* adatbázisban a **KW=$, TI=$, CC=$** és **PC=$** keresések egybehangzóan 165 587-et adtak összrekordszámként, akkor a **CI>0**, illetve **PR>0** keresések azt mutatták, hogy mindössze 88 672 rekordban van példányszámadat, és 67 674 rekordban ár. Ezeket az adatelemeket tehát igen óvatosan kell kezelni. A Bowker cégnek világosan figyelmeztetnie kéne a felhasználót, hogy ezeket ne használja keresőmezőként.

Még rosszabb eredményt, mindössze 28 158 találatot ad az **LC=$** keresés, vagyis a Library of Congress osztályozási jelzete alkalmatlan a keresésre. Az adatbázis-előállító mentségére szól, hogy ezt megemlíti mind a kézikönyvben, mind a reklámanyagokban. Ez azonban nem segít az alkalmi felhasználón, aki nem fér hozzá a nyomtatott dokumentációhoz. Jobb lenne, ha ez egyáltalán nem lenne keresőmező. Ez az az eset, amikor a kevesebb több lenne. A Bowker felmérése szerint ezt az adatmezőt a felhasználók közül nagyon kevesen kívánják keresésre felhasználni, a szerző személyes tapasztalatai azonban ennek a felmérési eredménynek ellentmondanak.

Jelöld ki és keress

A továbbiak közül ez a legjobb módszer akkor, ha egy mező tartalma nem vehet föl mondjuk száznál többféle értéket, ezek mind megjeleníthetők az adatmező indexéből, és keresésre kijelölhetők. Fontos, hogy egyszerre több felvehető értéket lehessen keresésre kijelölni, különben nehézkessé válik az eljárás.

Ez a módszer használható például a Bluefish és a KAware kersőrendszerekben. Ha mondjuk a *Computer Select* adatbázisban a *Cikk típusa* adatmezőre állunk, megkapjuk a képernyőn ennek a mezőnek az indexét. Ezen a mező valamennyi lehetséges felvehető értékét egyszerre kijelölhetjük, így megkaphatjuk azon rekordok számát, amelyekben bármilyen tartalommal szerepel ez az adatmező. Az 1992. júliusi kiadásban az adatbázis-szekció 82 902 rekordjával szemben azt kapjuk, hogy csak 41 187 rekordban van megadva a cikk típusa. A többi rekordot átnézve látjuk, hogy vannak további cikktípusok is, amelyek az indexben nem szerepelnek, például a *Trend*. Ez bizony aligha megbocsátható gondatlanság az előállító részéről.

Aritmetikai kereséssel további számottevő hiányokra bukkanhatunk ebben az adatbázisban. A céginformációs szekcióban a rekordok 66,8 százaléka tartalmazza a dolgozók létszámát, 43,1 százaléka az éves forgalmat, árinformációt a hardvertermékek szekciójában a rekordok 91,5 százaléka, a szoftvertermékek szekciójában mindössze 77,4 százaléka tartalmaz. A hiányok oka ezekben az esetekben nem az adatbázis-készítő hanyagsága, hanem az, hogy a kiadó nem képes beszerezni ezeket az adatokat az érintett cégektől. A tanulság azonban így is ugyanaz: legyünk óvatosak, ha ezekkel az adatelemekkel finomítjuk a keresésünket, sok, egyébként releváns rekord rejtve maradhat.

Ami a legzavaróbb ebben az egyébként kitűnő adatbázisban, az a javítására irányuló erőfeszítések hiánya. Mivel az adatbázis csak a legutóbbi 12 hónap rekordjait tartalmazza, csak el kellene határozni, hogy mostantól kezdve minden rekord kap érvényes cikktípuskódot. Az előállítónak nem kéne a rekordok százezreinek a visszamenőleges javításával küszködnie, a *Cikk típusa* mező kitöltése pedig nem éppen bonyolult feladat.

Böngészés a mező indexében

A mind online, mind CD-ROM-változatban ugyancsak széles körben használatos Wilsonline keresőrendszer szintén prefixes mezőnkénti keresést tesz lehetővé, de teljes csonkolásra nem ad módot, legalább egy karaktert ki kell írni. A *Kiadás éve* adatmező így is alkalmas a teljesség vizsgáltára, a **FIND(YR) 19:** keresés megadja mindazon rekordok számát, amelyekben a kiadás éve a *19* karakterekkel kezdődik. Ugyanígy használható a ProQuest szoftvert használó UMI adatbázisok esetében a **DA(19?)** parancs. A Wilsonline rendszerben az *yyddmm* alakot használó **DA** (a *Rekord bevitelének dátuma)* mező **FIND(DA) 8: OR 9:** formában használható az összrekordszám meghatározására.

* A dokumentumtípus (**DT=**) mezőt is csak 1982-ben vezették be részlegesen, és 1985-ben teljes körben. – A ref.

A többi mezőben ez a megoldás nem használható, mert nincs közös szótő. Amelyik mezőben azonban a lehetséges értékek száma korlátozott, ott valamennyi értéket megkaphatjuk az index kilistázásával (a Wilsonline rendszerben **NEIGHBOR** paranccsal), ezekre **OR** operátoros összekapcsolással elvégezve a keresést, megkapjuk a kívánt rekordszámot.

A *Wilson Business Abstracts* adatbázis 1991. december 26-i kiadásában például a **FIND(DA) 8: OR 9:** parancs az összrekordszámra 423 704 értéket adott. A mezőindexből kiindulva megkapjuk, hogy ezek mindegyike rendelkezik a három lehetséges rekordtípusérték valamelyikével, amelyiknek pedig *Cikk* a rekordtípusa közülük, az egyetlen kivétellel mind kapott *Tartalomtípus* kódot.

Érdekes, de nem dokumentált lehetősége a Wilson rendszernek, hogy a **NEIGHBOR*** parancs az egyesített index legelejére visz, ahol láthatjuk egyes adatmezőkről, hogy hányszor fordulnak elő az adatbázisban. Láthatjuk például, hogy az említett példa 423 704 rekordja közül 399 558 tartalmaz SIC-kódot és 92 279 tartalmi kivonatot. Míg az utóbbi érthető, hiszen kivonatot csak 1990 júniusa óta kapnak a rekordok, a SIC-kód gyakori hiánya arra int, hogy a keresésben ne hagyatkozzunk kizárólagosan erre.

Keresés ismert értékekkel

Ez nem túl kényelmes módszer, mivel ismernünk kell hozzá, és be kell vinnünk annak a mezőnek, amelynek a teljességét vizsgáljuk, valamennyi felvehető értékét. Emellett ez a módszer azokra a mezőkre korlátozódik, amelyek mintegy tucatnyi értéknél többet nem vehenek fel. Ilyen például a *Dokumentumtípus* mező. Míg a numerikus mezők jól vizsgálhatók az aritmetikai operátoros kereséssel (pl. ≤ vagy ≥), addig a szöveges mezők minden egyes lehetséges értékét külön kell bevinni.

A *PAIS* adatbázis SilverPlatter-változatában például a **PT=M OR PT=E OR PT=A** keresés szerint 331 406 rekordban található a *Publikáció típusa* adatmező, a **LA=E OR LA=F OR LA=G OR LA=I OR LA=P OR LA=S** keresés szerint pedig 331 397 rekord tartalmazza a *Publikáció nyelve* mezőt. Mindkét szám nagyobb, mint amit a kiadás évére végzett **PY>0** keresés ad (331 380).

Ugyanezt a keresést sokkal könnyebb elvégezni a *PAIS* OptiWare változatában, ahol nem kell tudnunk a mezők felvehető értékeit, mert teljes csonkolással kereshetünk. A próbakeresés a *PAIS* adatbázis meggyőző teljességét mutatja. A 331 406 rekord közül csak 9-ből hiányzik a nyelv kódja, 26-ból a kiadás éve, az OptiWare-változatban végzett **TI=$, SU=$, DT=P** és **JN=$** keresések szerint pedig egyetlen rekord nélkülözi a címet, 13 a témafejezetet, míg a 216 898 folyóiratcikk közül 2 a folyóirat nevét.

A UMI kiadásában megjelent *Resource One* adatbázisban a cikk hossza használható a teljesség vizsgálatára, mivel csak három értéket vehet fel: *length(short), length(medium)* vagy *length(long)*. Az elvégzett vizsgálat szerint minden rekord kap valamilyen értéket, tehát ezt a mezőt hatékonyan alkalmazhatjuk a dokumentumok hosszúság szerinti szelektálására.

Letöltés és megszámlálás

Vannak adatelemek, amelyek teljessége az eddig említett módszerek egyikével sem vizsgálható. Ilyenkor segíthet rajtunk egy szövegszerkesztő program. Ennek segítségével persze csak egy reprezentatív mintát vizsgálhatunk, azt is csak CD-ROM-környezetben. Válasszunk ki az adatbázisból valamilyen keresőkérdéssel egy ésszerű részhalmazt. Ez lesz a vizsgálati mintánk. Méretét az elérhető lemezterület vagy a szövegszerkesztő lehetséges állománymérete korlátozza. Töltsük le ezt a mintát egy adatállományba. Ha a CD-ROM keresőrendszer erre módot ad (DIALOG, Wilsonline, SPIRS, Compact Cambridge), akkor csak a vizsgálni kívánt mezőket töltsük le*, így kisebb a helyigény. Ezután a szövegszerkesztővel cseréljük ki a mezőazonosítót, akár önmagára (pl. **REPLACE text: DE: with text: DE:**). A szövegszerkesztő eközben összeszámlálja nekünk, hogy hány cserét hajtott végre, vagyis hány rekordban volt ilyen mező. Ha az érvényes érték nélküli mező megkülönböztetett jellel szintén benne van a rekordban (pl. a WILSONDISC adatbázisokban *SUB: not found*), az ezt tartalmazó rekordokat külön össze kell számlálnunk.

„Gyóntató" módszer

Ez a módszer azon alapszik, hogy egyes adatbázisok indexei speciális kóddal „vallják be", hány rekordban nem tartalmaz egy adatmező értéket. Ez elfogadható próbálkozás az előállító részéről, hogy enyhítse a hiányosság okozta problémákat.

Ideális példa erre a *Compact Disclosure* adatbázis, amelyben a prefixes mezők indexében *NA* érték jelzi a hiányt. Például az

1419 PC=NA
1766 SA=NA
1786 GP=NA

sorok megadják, hány adatrekordból hiányzik az *Elsődleges SIC-kód*, a *Nettó forgalom* és a *Bruttó nyereség*.

Más adatbázisokban esetleg csak néhány adatmezőre van ilyen indexsor. A *LISA* adatbázis DIALOG változata *PY=19XX* indexsorral adja meg, hány rekordból hiányzik a *Kiadás éve* mező. Ugyanennek az adatbázisnak a SilverPlatter változatában **PY=undetermined** keresőparanccsal kapjuk meg a kérdéses szá-

* és egy „biztos" mezőt, pl. a rekordazonosítót. – A ref.

mot. Ez a jelölés nem található a dokumentációban, így ez a gyónás arra emlékeztet, amikor csemeténk alig hallható motyogással vallja be, hogy rossz fát tett a tűzre. A *LISA* OptiWare változatából hiányzik is ez a lehetőség.

A Bowker adatbázisokban **PY=9999** kereséssel kaphatjuk meg, hányszor hiányzik a kiadás éve.

A gyónással óvatosaknak kell lennünk, lehet, hogy az adatbázis nem minden „bűnét" vallja be. A *LISA* adatbázisban például találhatunk olyan adatrekordokat is, amelyekben sem valódi évszám, sem *PY=19XX* érték nincs. Ezek száma szerencsére itt elhanyagolható. Máshol lehet a helyzet sokkal rosszabb. A *Books in Print* 1992. május-júniusi kiadásában 10 313 rekord tartalmazza a *9999* értéket a kiadás éveként, de 77 500 olyan rekord van, amely sem valódi kiadási évszámot, sem ilyen hiányt jelző értéket nem tartalmaz. Ez olyan, mintha fehér zászlót lengetve megadnánk magunkat, de közben egy Magnumot rejtegetnénk. A *9999* konvenció még a gyakorlott kereső éberségét is elaltatja.

„Természetes" módszer

A CD-Answer keresőrendszert használó adatbázisok, a *The Computer Archives,* a *Historical Abstracts,* az *America: History and Life* a lehető legegyszerűbb módon teszik lehetővé a teljességre irányuló keresést. A menünek az adatmezőnek megfelelő rovatába a **NONE** szót írhatjuk. Így közvetlenül megkapjuk, hány rekordban nincs értéke az adatelemnek.

II. rész: PONTATLANSÁGOK ÉS KÖVETKEZETLENSÉGEK

Ha a keresésben felhasznált adatelemek minden rekordban megtalálhatók is, akkor sem lehetünk biztosak abban, hogy minden releváns rekordot megtalálunk. Túl gyakori az adatbázisokban a pontatlan vagy következetlenül használt adat. Az alább ismertetett módszerek azt célozzák, hogy szisztematikusan megvizsgálhassuk az adatbázisok pontosságát és következetességét, felkészülve ezzel a defenzív keresésre. CD-ROM környezetben az ilyen keresés nem kerül pénzbe, és csak kevés időt igényel, de online környezetben is bőven megtérül az ára azon, hogy megismerjük az adatbázis pontosságát és következetességét.

Az ilyen hibák kevésbé veszélyesek, mint a hiányok, hála a jól ismert és elterjedt hétköznapi gyakorlatnak, és az indexböngészés lehetőségének.

Az I. részben említett, a hiányok feltárására szolgáló módszerekkel többnyire a teljes adatbázist vizsgáljuk. A pontatlanságok és következetlenségek vizsgálatakor általában csak mintavétellel dolgozhatunk. Többnyire csak azokat az adatmezőket vizsgálhatjuk, amelyekbe az adatelemeket előírt kifejezések közül választják, vagy amelyek adata meghatározott értékhatárok közé esik.

A régi jó böngészés

A keresés tízparancsolatából az egyik parancs: keresés előtt böngésszünk. Ezt figyelmen kívül hagyni olyan, mintha anélkül ugranánk fejest egy tizenöt méteres szirtről, hogy előzőleg megnéznénk, milyen mély a víz.

Ha csak alkalmilag böngészünk is az indexekben (ahogy látogatóba érkező anyósunk úgy mellesleg végigfuttaja az ujját a szekrény tetején, megnézni, hogy nem poros-e), már akkor is képet kapunk arról, vajon elegendő gondot fordított-e az adatbázis-készítő a minőség-ellenőrzésre. Ha a H. W. Wilson adatbázisok bármelyik név- vagy kódindexébe belekukkantunk, meggyőződhetünk róla, hogy azok milyen következetesek, éles ellentétben az alábbi példák adatbázisaival.

Az alkalmi elírások szinte normálisak bármelyik adatbázisban, és sokkal könnyebben megbocsáthatók, mint a rekordok viszonylag nagy számát érintő hibák. Az utóbbiakra szolgál példaként a *Gale's Book Review Index,* amelyben a *Dokumentumtípus* mezőben 191-szer fordul elő helyesen a

DT=CHILDREN'S PERIODICAL

kifejezés, 277-szer a helytelen

DT=CHILDRENS PERIODICAL

forma. Hasonlóan elriasztó példa az *Economic Literature Index,* amelynek *Folyóiratnév* indexében találjuk a következőket:

7 JN=HOMG KONG ECONOMIC PAPERS
53 JN=HONG KONG ECONOMIC PAPERS
21 JN=INDIAN JOURNAL OF QUANTITATIVE ECONOMICS
37 JN=INDIAN JOURNAL OF QUANTITATIVE ECONOMIS
46 JN=JOURNAL OF ECOMONIC AND SOCIAL MEASURES
53 JN=JOURNAL OF ECONOMIC AND SOCIAL MEASURES
13 JN=POPULAITON RESEARCH AND POLICY REVIEW
42 JN=POPULATION RESEARCH AND POLICY REVIEW

Mindkét példa vigyázatlanságra és nagyfokú nemtörődömségre utal. Az ilyen hibák, amelyeket bármelyik elemista megtalálhatna és kijavíthatna, kétségessé teszik a többi adatmező minőségét is.

Ha a böngészés mellett még csonkolásra is van mód, az nagyban csillapíthatja gondjainkat. Egyes keresőrendszerek azonban (pl. a SPIRS és a Bluefish) a kérdéses adatmezőkben ezt nem teszik lehetővé. A SPIRS ráadásul sok fontos adatmezőben *(Dokumentumtípus, Kiadás éve, Országkód)* még a böngészésre sem ad módot. Ez a felhasználó cserbenhagyása,

hiszen csak találgathatja, milyen adatformátumok és lehetséges értékek fordulnak elő ezekben az adatmezőkben. Ez kétségtelenül segíti az adatbázis-készítőt abban, hogy a szemetet a szőnyeg alá seperje.

Bakugrásos böngészés

A helytelen és a helyes forma nem mindig szomszédos. Az *Economic Literature Index* adatbázisban például az **E DT=Journal of Econ** parancsra észrevétlenül maradna az említett helytelen *DT=Journal of Ecomonic...* forma, ha a DIALOG **EXPAND** parancsa nem adna két sort az indexből a kijelölt kifejezés *előtt*. (A helytelen és a helyes forma közé ékelődik még a *Journal of Econometrics*.)

Még rosszabb a helyzet a SPIRS keresőrendszerben, amelynek ömlesztett indexében az elírt forma több tucat képernyőnyi távolságban lehet a helyes formától, attól a cím, a szerző, a kivonat, a deszkriptor és a folyóiratnév mezők szavaival és kifejezéseivel elválasztva.

Az elírás vagy következetlenség miatt egymástól távolra kerülő kifejezéseket deríthetjük fel a bakugrásos böngészés módszerével. Szemeljünk ki néhány olyan személynevet és intézménynevet, amelyekről valószínű, hogy következetlenül szerepelnek egy piszkos adatbázisban. Böngésszünk valamennyi sejthető névváltozat környezetében, amelyek egymástól távol lehetnek. Így például a *LISA* adatbázisban a *Chen-Ching-Chi* név négy változatára, és a vele nyilvánvalóan azonos *Ching-Chi-Chen* név további négy változatára bukkanunk. Az *Ulrich's Plus* adatbázisban a *John Wiley & Sons* vagy *John Wiley and Sons* kiadó hat névváltozatával, *Wiley & Sons* vagy *Wiley and Sons* kezdettel további tíz névváltozatával találkozunk.

Hogy helyesen és következetesen is lehet írni a neveket, azt a Wilson adatbázisok példája bizonyítja, így a *Library Literature* és a *Book Review Digest*.

Vannak esetek, amikor a legdefenzívebb kereső is reménytelen helyzetbe kerül. A *PAIS* adatbázis csaknem minden változatában az általánosan szokásos módon írták át az umlautos német magánhangzókat: az umlaut nélküli alapmagánhangzó után tett *e* betűvel. Van azonban egyetlen változat, a SilverPlatter-féle, amelyben a programozó úgy gondolta, hogy az *e*-t az alapkarakter *elé* kell tennie. Így az *Österreich* szóból például *Oesterreich* helyett (20 előfordulás) többnyire *eOsterreich* lett (2131 előfordulás). Ha ennek a szónak valamely változatával folyóirat, kiadóvállalat, cég vagy szerző neve kezdődik, az igen messze kerül a várható helyétől. Hogy ez nem véletlen, azt bizonyítja a *München* szó 100 előfordulása *Meunchen* formában, a *Börse* 171 előfordulása *Beorse* formában, és a *Geschäft* 268 előfordulása *Gescheaft* formában, szemben az egyszer sem található *Muenchen* és *Geschaeft* formákkal, illetve a *Boerse* forma egyetlen előfordulásával.

Megfelelések keresése

Bizonyos kódok egyértelműen meg kell feleljenek bizonyos szöveges mezők tartalmának, például az ISSN a folyóiratnévnek, a D-U-N-S szám a cégnévnek. Már elég szkeptikusak lehetünk ahhoz, hogy ezt az egyértelmű megfelelést ellenőrizzük. Válasszunk ki néhány ilyen párt, végezzük el mindkét tagjukkal a keresést, majd a kódkeresés találatai közül zárjuk ki a szöveges keresés találatait.

Az *ABI/INFORM* a legelső adatbázisok egyike volt, amelyekben néhány éve jelentős nagytakarítást tartottak. Amikor az MCI cégnévvel és a megfelelő D-U-N-S számmal keresést végeztünk, akkor ennek ellenére 12 rekordot találtunk 177 közül, amelyben ez a D-U-N-S szám más cégnév mellett szerepel. Ugyanilyen módon azonban valamennyi UMI adatbázisban hibátlan egyezést találtunk az ISSN és a folyóiratnév között.

Keresztutalások keresése

Előfordulnak jogos címváltozatok, névátírási változatok, megváltozhat egy folyóirat címe, országok, cégek neve, változhatnak a tezauruszok az új vagy részletesebbé váló terminológiának megfelelően. Ezek az esetek keresztutalások segítségével kezelhetők az adatbázisokban. Sok információkereső program elegánsan kezeli ezeket a keresztutalásokat, például a SPIRS, az OptiWare, a DIALOG Online és OnDisc, a Wilson szoftver böngésző módban, valamint a ProQuest újabb változata. Más programok nem nyújtanak megoldást, ilyen a Bluefish, és ilyen volt a ProQuest korábbi változata. A legveszedelmesebb az, ha egyszer van keresztutalás, máskor nincs. Ilyen például a *Magazine Article Summaries,* amelyben a *jog a halálhoz* és az *öngyilkosság* kifejezések között találunk keresztutalást, de a *jog a halálhoz* és az *eutanázia* kifejezések között nem.

A keresztutalások meglétét ilyen kiszemelt kifejezéspárokkal vizsgálhatjuk. Más példák erre a *Kampuchea* és a *Cambodia* országnevek, az *AT&T* és az *American Telephone and Telegraph* cégnevek.

Lehetetlen értékek

A kódolt mezők és sok numerikus mező hibátlanságát úgy is vizsgálhatjuk, hogy szántszándékkal hibás értékeket keresünk a mezőben. Ez a szoftver képességeitől és az indexek típusaitól függően többféleképpen történhet.

A numerikus mezőkben (pl. a *Kiadás éve, SIC-kód, Dewey-kód*) megkereshetjük a nem numerikus értékeket a **PY<0, SC<0, DC<0** kifejezésekkel. Ha egy mező tartalma betűvel kell kezdődjön, **tt>ZZZ** típusú kereséssel kapjuk meg a hibás értékeket. Tekintettel kell persze lennünk a jogos kivételekre, például a *SIC-kód* mezőben lehetséges *N/A* értékre, mondjuk a *Disclosure* adatbázisban.

Ennek a módszernek kicsit bonyolultabb változata az intervallummal végzett keresés. Az *Education Library* adatbázisban például régi könyvek is szerepelnek. Így itt a **PY<1500 OR PY>1992** keresőkifejezéssel találjuk meg azokat a rekordokat, amelyekben a kiadás éve az elfogadható értéktartományokon kívül esik. Láss csodát, jóval több mint százezer ilyen rekordot találunk. Ez persze felettébb gyanús. Az 1992 utáni rekordok többségükben olyanoknak bizonyulnak, amelyek kiadási évként *199?* vagy *199-* szerepel. Az 1500 előttiek viszont majdnem mind olyanok, amelyekben a *Kiadás éve* helyett a *Copyright éve* szerepel *c1990, c1967* stb. formában. Súlyos figyelmetlenség volt ezeket így indexelni a CD-ROM-készítés során. Mivel a SPIRS adatbázisokban a *Kiadás éve* nem böngészhető, a felhasználók a rekordok közel felét elveszítik, ha a keresés során a kiadás évével korlátoznak*.

Következtetések

Ha bármelyikünk ilyen könnyen megtalálja a szemetet az adatbázisokban, miért nem végeznek hasonló vizsgálatokat az adatbázis-készítők, és miért nem lépnek a tapasztalatok nyomán? Részben a „kit érdekel?!" mentalitás miatt, részben a költségek miatt. A keresési eredményeket súlyosan eltorzító pontatlanságok és következetlenségek többségét azonban az előállító vagy a kiadó jelentéktelen költséggel könnyedén kijavíthatná. Ha sokan végzünk ilyen vizsgálatokat, és tudtára adjuk azok eredményét az előállítónak vagy a kiadónak, az talán arra ösztönözheti őket, hogy legalább a minimális javításokat végezzék el. Ha viszont panaszunk süket fülekre talál, akkor is legalább felkészülhetünk a defenzív keresésre. Ha pedig az ilyen vizsgálatok eredménye bekerül az adatbázis-bírálatokba, az a többieknek is tanulságul szolgál.

/JACSÓ P.: Searching for skeletons in the database cupboard. Part I: Errors of omission. = Database, 16. köt. 1. sz. 1993. p. 38–49.
JACSÓ P.: Searching for skeletons in the database cupboard. Part II: Errors of commission. = Database, 16. köt. 2. sz. 1993. p. 30–36./

(Válas György)

* A bemutatott próbakeresés szerint csak az egyharmaduk vész el, de az is iszonyúan sok. – A ref.

Az információtudomány eredete, fejlődése és kapcsolatai

Az információtudománynak három általános jellemzője van. (Számos szakterület osztozik rajtuk vele.) Először: az információtudomány *interdiszciplináris jellegű*, az egyéb területekkel való viszonyai azonban változóban vannak. Ennek a fejlődésnek még távolról sincs vége. Másodszor: az információtudomány szorosan *kapcsolódik az információs technikához*. A technika kényszerítő ereje az információtudomány felett is ott lebeg. Szélesebb értelemben ez hajtja a modern társadalom fejlődését az „információs társadalom", „információs korszak" vagy a „posztindusztriális társadalom" felé. Harmadszor: az információtudomány sok egyéb területtel együtt aktív és megfontolt *résztvevője az információs fejlődésnek*. Az információtudománynak komoly társadalmi szerepet kellett és kell játszania: a technika felett és azon túl jelentős társadalmi és humán dimenziói vannak.

E három jellemző vagy vezérmotívum keretében érthetjük meg az információtudomány múltját, jelenét és jövőjét, s azokat a kérdéseket, problémákat, amelyekkel szembenéz.

Eredet és társadalmi háttér

Mint sok más interdiszciplináris terület (pl. a számítógép-tudomány, operációkutatás), az információtudomány is a második világháborút követő tudományos és technikai forradalomban gyökerezik. Az új szakterületek kialakulásának folyamata, és a régiek interdiszciplináris kapcsolatainak kibontakozása semmiképpen sem fejeződött be. Az információtudomány ugyanazokon a fejlődési szakaszokon megy át, mint sok más terület.

Jelentős történelmi fordulatnak, az információtudomány lendítőerejének és valódi kezdetének tarthatjuk *Vannevar Bush: As we may think* című cikkét, amely 1945-ben az *Atlantic Monthly*ban jelent meg. Bush, a MIT tekintélyes tudósa, a II. világháborús amerikai tudományos erőfeszítések vezetője ebben az írásban (1) tömören meghatározta azt a lényeges problémát, amely már régóta élt sokakban; (2) olyan megoldást javasolt, amely összhangban volt kora szellemiségével, és stratégiailag is vonzó.

A probléma az volt (s ez alapjaiban máig is megmaradt), hogy a „rémisztő mennyiségű tudást hozzáférhetőbbé tegyük". Bush meghatározta az „információrobbanás" problémáját – az információ és annak rögzített formái szüntelen exponenciális növekedését, különösen a természet- és műszaki tudományok területén. Szerinte a fejlődő információs technikának kell megbirkóznia ezzel a feladattal. Egy MEMEX nevű gépet javasolt, amely képes a „gondolatok asszociációjára",