A FREETEXT SZÖVEGES INFORMÁCIÓKERESŐ RENDSZER

Erdős Iván—Biszak Sándor
Infoker Kisszövetkezet

A professzionális személyi számítógépek elterjedése új távlatokat nyithat a magyar könyvtárak számára is. A legnehezebb feladatot a megfelelő szoftver kiválasztása jelenti. A hazai fejlesztések két irányban indultak el. Többen az UNESCO által készített szöveges információkereső programot, a Micro-ISIS-t használják, míg mások a dBASE III relációs adatbázis-kezelőt. Ez utóbbi, szemben a Micro-ISIS-szel, nem igazi információkereső rendszer, viszont egyszerűsége és főleg programozhatósága révén a könyvtárakban is nagy népszerűséget szerzett. Egyetértünk azokkal, akik azt állítják, hogy a dBASE III alkalmatlan szöveges adatok tárolására. Véleményünk szerint (és az eddigi tapasztalatok is ezt támasztják alá) a Micro-ISIS hazai alkalmazhatósága is nehézségekbe ütközhet. Ezért úgy döntöttünk, hogy létrehozunk egy saját fejlesztésű, valódi szöveges információkereső rendszert. Egyrészt létre akartunk hozni egy teljes egészében hazai fejlesztésű rendszert, amely versenyképes a Micro-ISIS-szel, másrészt az általunk készített magyar szabadalmi adatbázis is elérte azt a mennyiséget (jelenleg 40 000 bejelentés adatait tartalmazza), amely már egyre nehezebben kezelhető az eredetileg választott dBASE II adatbázis-kezelővel. *FREETEXT* névre keresztelt szöveges információkereső rendszerünket MS-DOS operációs rendszer alatt, TURBO PASCAL programnyelven írtuk.

A rendszer legfontosabb tulajdonságai:

- Változó hosszúságú mezők, rekordok kezelése
- A felhasználó által definiálhatók az adatbázisok, az adatbeviteli és megjelenítési formátumok
- Invertált fájlszerkezet
- Invertálási típusok: szavas, kijelöléses, teljes mezőre
- Ismételhető mezők
- A felhasználó által definiált stopwordlista
- Keresés során csonkolás, ÉS, VAGY, DE NEM operátorok használata
- Rögzítéskor beállítható alapértelmezések
- Rendezés tetszőleges mezőkre
- Hajlékony beviteli és megjelenítési formátumnyelv, amely változatos formátumok készítésére alkalmas
- A nyomtatást befolyásoló adatok (pl. hasábszám, hasábszélesség, betűtípus) állíthatók.

Áttekintés, fogalmak

Mint minden adatbázis, a szöveges adatbázisok is *rekordokból* állnak. Egy rekord állhat pl. egy könyv, folyóirat adataiból. A rekordok *mezőkből* épülnek fel. Egy-egy mező lehet egy könyv címe, szerzője stb. A mezőket az adatbázis létrehozásakor határozzuk meg. Ekkor adjuk meg az *adatbázis nevét* is. A mezők adatait, jellemzőit az *adatszótár* tartalmazza. A felhasználó által definiált mezők mellett mindig létezik egy, a rendszer által generált állandó mező, amely minden rekordot egyértelműen azonosít. Ez a *belső sorszám,* amely 1-től induló, folyamatosan növekvő sorszám.

Lássuk, a mezők mely adatait tárolja az adatszótár!

Mezőnév: Max. 20 karakteres név, az adott mezőre a program használata során mindenütt ezzel a névvel hivatkozunk.

Invertálási típus: Szöveges adatbázist azért hozunk létre, hogy abban különböző szempontok szerint keresést végezzünk, az eredményeket különböző formátumban megjelenítsük. A keresést az ún. invertált fájlon keresztül végezhetjük. Az invertált fájl rendezetten tartalmazza a szavakat, kifejezéseket, amelyeket kereshetünk. Ezeket együttesen *keresőkulcsoknak* nevezzük. A keresőkulcsok mellett az invertált fájl azoknak a rekordoknak a belső sorszámát is tartalmazza, amelyekben az adott keresőkulcs előfordul. Rendszerünkben 4-féle *invertálási típust* különböztetünk meg.

- Nem történik invertálás, a mező tartalma alapján nem kereshetünk vissza.
- *Szavas invertálás.* A mező minden egyes szava bekerül az invertált fájlba, így a mező minden szavát visszakereshetjük. Szónak tekintünk minden olyan karaktersorozatot, amelyet *szóhatároló jelek* vesznek körül. A szóközön kívüli szóhatároló jeleket a felhasználó adja meg saját igényei szerint, tárolásuk egy, az adatbázishoz tartozó reprezentációs fájlban történik. Szóhatároló jelek lehetnek pl. a . , ; : ? ! ( ) stb. Ugyancsak ebben a fájlban találhatók az ún. *stopwordök.* Ezek azok a szavak, amelyekre nem óhajtunk invertálni. Megadásuk ugyancsak a felhasználó feladata. Stopwordök lehetnek névelők (az, a, egy), kötőszavak (és, vagy) és általában minden olyan szó, amely nagy számban fordul elő, de nem hordoz információtartalmat, így tárolása az invertált fájlban csak az adatbázis helyszükségletét növelné feleslegesen.
- *Kifejezésre invertálás.* Ebben az esetben a teljes mezőtartalomra történik invertálás. Jól használható szerző típusú (szerző, feltaláló, újító) mezőknél, vagy különböző osztályozási mezők (ETO, NSZO stb.) esetén. A mező típusa emellett lehet *ismételhető* is (lásd alább). Nem ismételhető mező esetén a teljes mezőtartalomra, ismételhető esetén az elválasztójelek között szereplő valamennyi mezőértékre történik invertálás.
- *Kijelöléses invertálás.* A mezőből azok a karaktersorozatok kerülnek az invertált fájlba, amelyeket a felhasználó által definiált jelek közé zárunk. A kijelöléses invertálás nyitó és záró jelét ugyancsak a reprezentációs fájlban találhatjuk. Ilyen típusú invertálás jól használható hosszú, szabad szöveget tartalmazó mezőknél, ahol a szavas invertálás túl sok helyet foglalna.

További mezőjellemzői csak az invertált mezőknek vannak.

Indexsorszám: 0-tól induló sorszám, azt adja meg, hogy az adatbázishoz tartozó hányadik indexfájlba kerüljenek az adott mező keresőkulcsai.

Indexhossz: 1—40 közötti egész szám; a keresőkulcs hosszát adja meg az indexfájlban. Ez tehát fix hosszúságú, ha az adott keresőkulcs ennél hosszabb, a fennmaradó rész levágásra kerül. Valamennyi invertálási típusnál pontosan meghatározza a tárolandó kulcs hosszát.

Egyediség: Ha a mező egyedi, az adott mezőben minden érték csak egyszer fordulhat elő, ezt a program bevitelekor, módosításakor ellenőrzi, és nem enged már létező értéket ismételten bevinni.

Ismételhetőség: Ha a mező ismételhető, az adott elválasztójelekkel határolt mezőértékek mindegyike bekerül az indexfájlba, ellenkező esetben csak a teljes mezőtartalom.

Az adatbázishoz kötelezően tartozik még egy *beviteli* és egy *megjelenítési formátum,* utóbbi az alapértelmezés szerint megjelenítési, előbbi az adatbeviteli formátum. Adatbevitelkor, ha nem választunk másikat, az ebben leírt formátum használatos. Egy adatbázishoz tetszőleges számú beviteli és megjelenítési formátum tartozhat.

Az adatbázisban az invertált fájlok alapján kereshetünk. Az adott feltétel(ek)nek megfelelő rekordokat *találatoknak* nevezzük, azokat *találati halmazokba* gyűjtjük, rájuk sorszámokkal hivatkozhatunk.

Az adatbeviteli formátum szerkezete

Az adatkarbantartás során az adatbázis adatai a képernyőn jelennek meg az aktuális formátum szerint. A formátum-menüpont segítségével ez bármikor átállítható, ekkor az új formátum nevét kell megadni. Ha már létezik ilyen nevű formátum, az aktuálissá válik, ha még nem, új adatbeviteli formátumot hozhatunk létre, és ez lesz az aktuális. Lássuk a beviteli formátum adatait.

A képernyőn minden mező egy-egy bekeretezett ablakban helyezkedik el. A formátum tartalmazza az ablak bal felső sarkának koordinátáit, az ablak szélességét és feliratát. Az ablak 3 soros: egy alsó, egy felső keret a felirattal és egy adatsor. Ha ebben a mező tartalma nem fér el, az ablak többsorossá válik, tehát mindig annyi sorból áll, amennyi az adatmegjelenítéshez szükséges. Az ablakok tetszés szerint átlapolódhatnak, egymásba érhetnek. A felirat tetszőleges, célszerűen a mezőnév vagy annak rövidítése. Az input formátumhoz a fentiek mellett a bevihető érték hossza és kötelezettsége tartozik. Előbbi 1 és 253 közötti érték, és azt adja meg, hogy milyen hosszú lehet maximálisan az adott mező. Ennél hosszabb értéket bevinni nem lehet. Ha a mező fix hosszúságú, üresen hagyható, vagy kötelező ilyen hosszban kitölteni (a kitöltés pl. 10 hosszúságú mezőnél nem hagyható abba 6 karakternél). A kötelezettség értéke igaz vagy hamis. Az előbbi esetben a mező kitöltése kötelező, az utóbbiban nem.

Az új formátum megadásakor a fenti adatokat kell meghatároznunk. A mező megadása a képernyő jobb oldalán megjelenő mezőnevek közötti választással lehetséges. Az éppen kiválasztott mező más színű, a le- és felfelé mutató nyilakkal sétálhatunk ezen a listán. Ha 17-nél több mezőnk van, a következő, ill. előző lapra a PgDn, ill. PgUp billentyűkkel

léphetünk. A mező kiválasztása az ENTER billentyű hatására történik meg. ESC hatására nem történik mezőkiválasztás, hanem befejeződk az input formátum megadása. Az input formátum később bármely szövegszerkesztővel módosítható. Egy adatbázishoz tetszőleges számú formátum fájl létezhet, ezek nemcsak formájukban különbözhetnek, hanem adattartalomban is. Lehet pl. dokumentumtípusonkénti beviteli formátum. Elképzelhető, hogy ugyanazon adatbázisban tartunk cikkeket és szabadalmakat, ekkor készülhet egy formátum a cikkek, egy másik a szabadalmak bevitelére. A mezőhossz és a kötelezettség nem az adatszótárban meghatározott mezőtulajdonság, hanem adott beviteli formátumhoz kapcsolódik.

Adatbevitel

A bevitel almenü segítségével új rekordokat tölthetünk az adatbázisunkhoz az érvényes beviteli formátum alapján. A képernyő első sorában a betöltés alatt álló rekord belső sorszámát láthatjuk, ez eggyel nagyobb, mint a legutoljára betöltött. A betöltés alatt, és minden olyan későbbi esetben is (módosítás, lekérdezés), amikor a képernyőn látható ablakokban szerkesztünk, a következő szerkesztőkaraktereket használhatjuk:

ENTER	: Lépés a következő mezőre
↑	: Lépés az előző mezőre
Ctrl G,Del	: Törlés a kurzor pozíciójában
BackSpace	: Törlés a kurzor előtt
←, →	: Kurzormozgatás
Ctrl F,End	: Ugrás a mező végére
Ctrl A, Home	: Ugrás a mező elejére
Ctrl Y	: A teljes mezőtartalom törlése
ESC	: A szerkesztés befejezése

Ha egy ablak megtelik, magassága eggyel nagyobb lesz, míg el nem érjük a maximális hosszát. Ha a mező kötelező, nem tudunk továbblépni, míg nem töltjük ki. Nem tudunk kilépni a mezőből akkor sem, ha az fix hosszúságú és nem töltöttük ki teljes hosszában (vagy nem hagytuk teljesen üresen).

A bevitt adatok azonnal bekerülnek az adatbázisba, és az invertált fájlok karbantartása is megtörténik. Ha valamelyik mezőt egyedinek definiáltuk, és ennek a mezőnek már létező értéket adtunk, hibaüzenetet kapunk; a rekord nem kerül betöltésre.

Módosítás

A hibásan bekerült vagy megváltozott rekordok módosítását az aktuális beviteli formátum szerint lehet elvégezni. A módosítandó rekordok kiválasztása tetszőleges szempontú kereséssel történik. A feltételeknek megfelelő rekordok közül az első kerül a képernyőnkre (ha volt ilyen), méghozzá azok az adatai, amelyeket a beviteli formátumban megadtunk. Módosításkor használhatjuk a bevitelnél használt vezérlőbillentyűket.

Törlés

A feleslegessé vált adatok törlését végezhetjük a segítségével. A törölni kívánt rekordok köre egy kereséssel hozható létre, hasonlóan a módosításhoz. Ha vannak a feltételnek megfelelő rekord(ok), az(ok) a képernyőn megjelennek. A rekord megtekintése után választhatunk, hogy törölni kívánjuk-e valóban a rekordot az adatbázisból. A törlés és az invertált fájlok karbantartása is azonnal megtörténik.

Alapértelmezés

Ezen almenü segítségével lehetőségünk van a mezőknek alapértelmezést adni az interaktív bevitel gyorsítása érdekében. Lehetnek olyan mezők, amelyek tartalma ritkán (vagy egyáltalán nem) változik a bevitel során. Ezek ismételt begépelését takaríthatjuk meg az alapértelmezés megadásával. E mezők tartalma is módosítható a bevitel során, de ha erre nincs szükség, egyszerűen átugorhatjuk a mezőt. Ilyen mező lehet: a beviteli dátum, a rögzítő neve, de akár a szerző neve is, ha sok tétel van egy-egy szerzőtől.

Interaktív lekérdezés

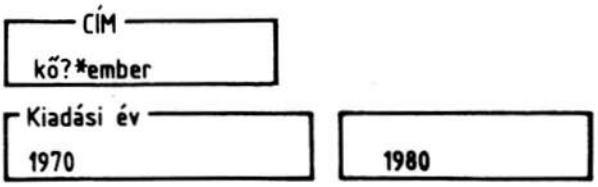
Szelektálás

A szelektálás segítségével hozhatunk létre találati halmazokat. A képernyőn a kereshető (tehát invertált) mezők ablakai jelennek meg. Ezekben az ablakokban adhatjuk meg a keresési feltételünket. Egyszerre több mezőre adhatunk meg feltételt, ezek logikai ÉS művelettel kapcsolódnak össze. Egy mezőn belül használhatjuk vagy a + (logikai összeadás, VAGY) a * (logikai szorzás, ÉS) műveletet. A ? a *csonkolás* jele, hatására az adott karaktersorozattal kezdődő szavakat kereshetjük meg. Minden mezőre két értéket adhatunk meg, egy alsó és egy felső határt. Így *intervallumkeresésre* van lehetőségünk.

A keresést nagyban segíti, hogy betekinthetünk az invertált fájlba, megnézhetjük a benne szereplő keresőkulcsokat. Az invertált fájlban pozicionálhatunk, annak az a része jelenik meg a képernyőn, ahol az általunk megadott kulcs előfordul (vagy elő-

fordulna, ha szerepelne az adatok között). A rendezett listán a ↓, ↑, PgDn, PgUp billentyűkkel navigálhatunk. Így gyorsan áttekinthetjük a lehetséges keresőkulcsokat, azok adatbázisbeli írásmódját, változatait. A kulcsok mellett azok előfordulási számát láthatjuk, innen már előre tájékozódhatunk, hogy hány találatot fogunk kapni a kérdésünkre. Segítséget ad a csonkolás pontos helyének a meghatározásához, és biztonsággal jelzi, ha az adott kulcs hiányzik az adatbázisból. Az ablakban a középső kulcs kiemelt színű, ENTER hatására ez kiválasztásra kerül, nem kell újra beírnunk.

A következőkben egy példát mutatunk a lekérdezésre. Jelentése: keressük azokat a dokumentumokat, melyek címében szerepel kő kezdetű és az ember szó. Ezenkívül kiadási éve 1970 és 1980 között van. Találatként kaphatjuk A kőszívű ember fiai című könyvet, amelyet 1978-ban adtak ki.



Eredményként egyrészt az egyes feltételeknek megfelelő dokumentumok számát, valamint az addigi összes feltételnek megfelelő dokumentumok számát kapjuk. Végeredményként, ha a találatok száma nagyobb, mint 0, a létrejött találati halmaz (SZET) sorszámát és a találatok számát kapjuk. Ha nincs találat, nem képződik találati halmaz. A keresést nem folytatjuk tovább, ha valamelyik feltételnél már nincs találat. A fenti kérdésre a gép válasza:

SZET	RÉSZ	TELJES	MEZŐ	ÉRTÉK
	30	29	Cím	kő?
	10	8	Cím	ember
	458	1	Kiadási év	1970–1980
0		1		

Kő kezdetű szó 30-szor fordul elő a címekben, ez azonban csak 29 dokumentumot jelent, tehát ebből egy dokumentumban 2-szer fordul elő. Az ember szó 10-szer szerepel, ebből 8-ban kő kezdetű szó is van. 458 tétel kiadási éve esik 1970 és 1980 közé, ebből 1 felel meg a fenti két feltételnek is. Ezzel létrejött a 0. találati halmaz, 1 találattal.

Egy lekérdezési menetben 50 találati halmazt képezhetünk. A lekérdezés története a képernyőn követhető, ahol egyszerre 19 sor látszik. Ha a lekérdezés ennél több kérdést tartalmaz, a ↑, ill. ↓ billentyűkkel fel-le futtathatjuk a listát.

Szűkítés

Ezen almenü segítségével egyrészt a belső sorszám szerinti lekérdezésre nyílik lehetőségünk, másrészt adott találati halmaz találatait szűkíthetjük belső sorszám szerint. Ha még nem jött létre találati halmaz, és ezt a menüpontot választottuk, megadhatjuk a belső sorszám "tól-ig" értékeit. Segítségül megkapjuk, hogy mi a legnagyobb belső sorszám. Az ilyen típusú lekérdezés rögzítés-ellenőrzéskor lehet hasznos, amikor is egy adott belső sorszámnál nagyobb belső sorszámúakat kell ellenőrzés céljából kinyomtatni. Adott találati halmaz találatait is szűkíthetjük, ekkor a halmaz sorszámát és a szűkítés "tól-ig" értékét kell megadnunk. Ez hasznos lehet pl. témafigyelésnél. Adott belső sorszámnál történt lekérdezés után ugyanabban a témában csak az új tételeket kell keresni, megjeleníteni.

Kombinálás

A létrejött találati halmazokat ÉS, VAGY, NEM logikai operátorokkal kombinálhatjuk. A NEM kétváltozós, jelentése DE NEM. Egyszerre két találati halmazzal végezhetünk műveletet. Inputként az első találati halmaz sorszámát, a műveleti jel első betűjét, valamint a második találati halmaz sorszámát kell megadnunk. Eredményként, ha van találat, új találati halmaz képződik, valamint a találatok számát kapjuk meg.

A megjelenítési formátum szerkezete

A létrejött találati halmazok a Display vagy a Print paranccsal jeleníthetők meg a képernyőre vagy a nyomtatóra. A megjelenítés az aktuális formátum szerint történik. A program indításakor az alapértelmezés szerinti formátum az aktuális. Ez bármikor átállítható. Ekkor az új formátumot tartalmazó fájl nevét kell megadni. Ha nincs ilyen nevű fájl, létrehozhatjuk az új formátumot; ha már létezett, a formátum aktuálissá válik.

Lássuk részletesen a formátumnyelvet!

A formátum lényege, hogy megadhatjuk, mi jelenjen meg a mező *előtt, után,* mi jelenjen meg a mező *helyett,* ha a mező tartalma üres. Az adatbázis mezői mellett a belső sorszám is megjeleníthető. Az alapértelmezés szerinti formátumfájl létrehozása az adatbázis létrehozásakor történik, továbbiak a F/ormat almenüben hozhatók létre. A formátumfájl standard ASCII fájl, amely bármely szövegszerkesztővel módosítható. A formátumfájl szerkezete:

Mezősorszám	(0-mezőszám, közé eső érték)
Előtte	(Az itt leírt string jelenik meg a mező előtt, ha a mező tartalma nem üres)
Utána	(Az itt leírt string jelenik meg a mező után, ha a mező tartalma nem üres)
Tabulátor	(Ha a mező tartalma nem fér be az aktuális sorba, ennyivel beljebb kezdődik a következő sor)
ElőtteHiány	(Az itt leírt string jelenik meg a mező előtt, ha a mező tartalma üres)
UtánaHiány	(Az itt leírt string jelenik meg a mező után, ha a mező tartalma üres)
Helyette	(Értéke lehet mezősorszám vagy string. Előbbi esetben az itt megadott mező tartalma jelenik meg, ha a mezősorszámmal megadott mező üres. Ha stringet adtunk meg, az fog megjelenni, ha a mező üres)
HelyetteTabulátor	(Ha a "Helyette" mező nem fér el az adott sorban, ennyivel beljebb kezdődik a következő sor)

A soremelést (új sorban kezdést) \ jellel jelöljük. A megjelenítés során, a \ hatására soremelés történik. Egy példa a formátumfájlra:

2	Az adatbázis 2. mezője
\Cím :	ELŐTT kezdjünk új sort, és jelenjen meg a CÍM :
/	UTÁNA tegyünk / jelet. Ha a cím tartalma nem
9	fér el a sorban, a cím 2. sorát 9 karakterrel beljebb kezdjük
	Ha a 2. mező ÜRES, ne jelenjen meg sem ELŐTTE,
	sem UTÁNA,
	sem HELYETTE semmi
	és ekkor TABULÁLNI sem kell
----------	Elválasztójel (Tetszőleges tartalmú sor)
3	Az adatbázis 3. mezője
Kiadási év:	ELŐTT jelenljen meg a Kiadási év: szöveg
.	UTÁNA tegyünk pontot
	TABULÁLNI nem kell
(	Ha ÜRES a mező, ELŐTTE (
)	UTÁNA) jelenjen meg
év nélkül	HELYETTE az év nélkül szöveg jelenjen meg
	TABULÁLNI nem kell

A fenti formátum alapján a rekordok:

CÍM : Új módszerek a nagy ritkaságú fémek előállításához használt elektródák gyártására/ Kiadási év: 1984.
CÍM : Kémiai összefoglaló/ (év nélkül)

A formátumok létrehozásakor az input formátumnál már megismert mezőválasztás segítségével adhatjuk meg a megjelenítendő mezőket. A helyette érték megadásánál is megjelenik a mezőválasztás, itt is kijelölhetjük a menüből azt a mezőt, amely az üres mezőtartalom esetén megjelenítendő. ESC-vel jelezhetjük, hogy nem egy másik mezőt, hanem egy stringet akarunk a mező helyett megjeleníteni. Ekkor lehetőségünk van a string megadására.

A megjelenítési formátum mellett egyéb adatok is befolyásolják az outputképet. Ezek (zárójelben a lehetséges értéktartományt és az alapértelmezést adtuk meg):

Hasábok száma	(1–3,1)
Üres sorok a lap tetején	(0–20,0)
A sorok száma egy lapon	(10–80,19)
Üres sorok a lap alján	(0–20,0)
Sorok száma a tételek között	(0–10,1)
A hasábok közötti hely	(1–20,4)
A hasábszélesség	(25–80,70)
Betűtípus	(0–60,0)
Nyomtatás	(1–3,2)

Print (nyomtatás) történhet fájlba (1), nyomtatóra (2), mindkettőbe (3).

Az alapértelmezések a F/ormat menü segítségével megváltoztathatók, újabb változtatásig ezek lesznek érvényesek.

Megjelenítés nyomtatón, rendezés

A menüponttal egy adott találati halmazban található rekordokat jeleníthetünk meg az aktuális formátum szerint. A megjelenítés outputja nyomtató, fájl vagy mindkettő lehet. A nyomtatás laponként történik. Ha a megjelenítés fájlba történik, akkor standard ASCII fájl jön létre, amely bármely szövegszerkesztővel utólag szerkeszthető, ill. feldolgozható.

A nyomtatás alapértelmezésként a belső sorszámra növekvő rendezettségben történik. Lehetőség van arra, hogy tetszőleges mezőre rendezzünk. A találati halmaz sorszámának megadása után a rendezési ismérveket kell megadnunk. Ekkor a már ismert mezőválasztás menü jelenik meg, itt választhatjuk ki az első rendezési kulcsot. Meg kell adnunk a rendezési kulcs hosszát, a rendezéskor ennyi karaktert veszünk figyelembe. A hossz megadása után azt kell megadnunk, hogy ismételhető-e a mező. Lehetőség van tehát arra, hogy egy ismételhető mező minden értékére elvégezzük a rendezést. (Ekkor a listára egy tétel többször is bekerül!) Az első rendezési kulcs megadása után megadhatjuk, hogy mely mezőre történjen meg a rendezés, ha a fent megadott mező üres. Ezt ugyancsak a mezőválasztó menü segítségével tehetjük meg. Ugyancsak meg kell adni a kulcs hosszát és típusát. Megadhatunk második rendezési kulcsot is. Végezetül a rendezés irányát kell megadnunk, amely növekvő, ill. csökkenő lehet.

A rendezés során a karakterek sorrendjét a rendezési táblázat adja meg. A karakterek átváltása e táblázat alapján történik meg. Így a felhasználó maga határozhatja meg a karaktersorrendet. Lehetőség van a szabvány szerinti rendezésre, tehát az ékezetes betűk közül az *ö* és *ü* különböztetendő meg, az *a* és *á* vagy *e* és *é* pedig nem. De lehetőség van

arra is, hogy mindegyik ékezetes betűt megkülönböztessük az ékezet nélküli párjától, ahogy azt az igények diktálják. Rendezéskor figyelmen kívül hagyjuk az első helyen álló névelőket.

Rendezés esetén a megjelenítést *kiemeléssel* módosíthatjuk. Ez azt jelenti, hogy a rendezési kulcsot külön, kiemelve is megjeleníthetjük. Ennek megjelenítése csak akkor történik, ha változott az értéke. Kiemelésre csak az első rendezési kulcs kerül. (Ha ez üres volt, akkor az a mező számít első rendezési kulcsnak, amelyet helyette kijelöltünk.) A kiemelés esetén a megjelenítési formátumhoz hasonlóan megadhatjuk, hogy mit írjunk ki előtte, utána, és azt, hogy mi jelenjen meg helyette akkor, ha üres az érték. Végül meg kell adnunk, hogy hány karakterrel van kiemelve. Az "előtte", "utána" értékek tartalmazhatnak \ jelet a soremelés jelzésére. A kiemelés előtt és után mindenképpen egy soremelés történik. A rendezés, kiemelés alapján rendezett listákat, pl. szerzői bibliográfiákat készíthetünk. Az output formátumnál már megjelenített rekordok képe a kiemeléssel a következő lesz. (Most a szerző mező is megjelenítésre kerül.)

Berkes István:
CÍM : Új módszerek a nagy ritkaságú fémek előállításához használt elektródák gyártására/Berkes István, Nagy Zsolt; Kiadási év: 1984.
CÍM : Kémiai összefoglaló/Berkes István; (év nélkül)
Nagy Zsolt:
CÍM : Új módszerek a nagy ritkaságú fémek előállításához használt elektródák gyártására/Berkes István, Nagy Zsolt; Kiadási év: 1984.

A listában a szerző mezőre történt meg a rendezés ismételhető módon (tehát minden egyes szerzőre), és a rendezési kulcsot (a szerző nevét) 5 pozícióval emeltük ki.

Statisztika készítése

Még szöveges adatbázis esetén is gyakran felmerülő igény a különböző szempont szerinti statisztikák készítése. Ez sokféle lehet, pl.: hogyan oszlik meg a könyvállományunk kiadók, nyelvek szerint; hogyan változik ez a megoszlás évenként. Ezek mellett a statisztika az interaktív lekérdezést is támogathatja. Megkaphatjuk pl. adott találati halmaz kulcsszó szerinti megoszlását, amely ötleteket adhat a további kereséshez.

A statisztika bármely találati halmazra elvégezhető. Ha az egész adatbázisra akarunk statisztikát készíteni, akkor egy találati halmazba az összes tételt be kell tennünk. (Ez történhet pl. LIMIT segítségével.) A mező neve mellett a típusát kell megadnunk, amely kifejezéses, ill. szavas lehet. Utóbbi esetben a mező minden szava (a stopwordöket kivéve) bekerül a statisztikába, előbbiben a mező teljes tartalma (ismételhető mezők esetén az egyes előfordulások). A kulcshossz értékét is meg kell adnunk, ez dönti majd el, hogy milyen hossz esetén különböztessük meg a kulcsokat. A kulcshossz megválasztásával jól befolyásolhatjuk a hierarchikus osztályozási rendszereken végzett statisztika eredményét. Megadhatunk a fenti mező mellett egy másik mezőt, ez lesz a statisztika 2. változója. Ez fogja tovább bontani az első szempontot. Ha kiválasztottuk a 2. mezőt, annak hosszát kell megadnunk. Ez célszerűen egy dátum jellegű adat (kiadási év, bejelentés dátuma stb.). A statisztika eredményét a nyomtatón kapjuk meg. A lista formátuma:

	80	81	82	83	84	85	86	Össz
C07C	21	31	43	43	50	68	12	268
C07D	10	22	18	34	21	49	11	165
stb...								

Ez a példa egy szabadalmi rendszerben készülhetett volna egy találati halmaz Nemzetközi Szabadalmi Osztályozás (NSZO) mezője alapján. A kulcshossz 4 volt, E típusú. Kértek 2. változót, amely a bejelentés dátuma volt, a hossza 2. A tételek összértékre rendezetten jelennek meg, a 2. változó növekvő sorrendbe van rendezve.

Adatbázis létrehozása

Az adatbázis létrehozásakor a mezőjellemzőket kell megadnunk. A mezők megadásának végét e mezőnév üresen hagyásával jelezhetjük. A mezőjellemzők:

Mezőnév: tetszőleges, legfeljebb 20 karakteres név. A mezőnév-duplikáció nem megengedett.

Invertálási típusok:
nincs invertálás,
kifejezésre invertálás,
szavas invertálás,
kijelöléses invertálás.

További mezőjellemzők csak abban az esetben adhatók meg, ha a mező invertálandó.

Indexhossz: 1 és 40 közötti érték, az indexfájlbeli kulcshosszt határozza meg.
Indexsorszám: az indexfájl sorszámát adja meg, amelybe az adott mező kulcsértékei kerülnek.
Egyedi: I esetén a kulcs egyedi, N esetén nem.
Ismételhető: I esetén a mező ismételhető, N esetén nem.

A mezőjellemzők megadása után a reprezentációs fájlt kell létrehoznunk. Előbb a szavas invertálás, majd az ismételhető mezők elválasztójeleit, végül a kijelöléses invertálás kezdő, ill. záró jelét kell megadnunk, ugyancsak itt adandók meg a stopwordök is. Következő lépésként az alapértelmezés szerinti megjelenítési, majd beviteli formátumot kell létrehoznunk.

A Micro-ISIS-szel szembeni előnyök

a) A Micro-ISIS legalapvetőbb hiányosságának azt tartjuk, hogy lezárt rendszer. Továbbfejlesztése a forráskód hiányában (amelyet tudomásunk szerint az UNESCO nem bocsát a felhasználók rendelkezésére) lehetetlen. Tehát amellett, hogy elkészítettünk egy általános rendszert, amely önmagában használható, megvan annak a lehetősége, hogy tetszőleges irányba továbbfejlesszük. Hosszan lehet sorolni azokat a feladatokat, amelyek általános megoldása szinte lehetetlen. Ilyen pl. az adatbevitelkor az adatok ellenőrzése, több adatbázis összekapcsolása egy rendszerré. Úgy érezzük, hogy professzionális, több tízezer, esetleg több százezer dokumentumot tartalmazó, adatbázisok készítésénél sokkal megnyugtatóbb, biztonságosabb egy teljesen kézben lévő, minden részében áttekinthető és alakítható program alkalmazása.

b) Részletesebben kell szólnunk a karakterkészlet problémájáról. Egy könyvtárban – nézetünk szerint – nélkülözhetetlen alapvető igények:

- A teljes magyar karakterkészlet kezelése mind a képernyőn, mind a nyomtatón, mind a klaviatúrán. Ez utóbbin azt értjük, hogy a klaviatúra feleljen meg a szabványos, magyar írógép-billentyűzetnek.
- A karakterek a magyar ábécé szerint kövessék egymást pl. az indexfájlban, a rendezés pedig a szabvány szerint történjék.
- Keresni lehessen az ékezetes betűkre is, megkülönböztethetők legyenek pl. a tor, tőr, tör szavak.

Ezeket az igényeket olyan fontosnak tartjuk, hogy hiányukban mindenféle gépesítést elhamarkodottnak ítélünk. Meglepődve tapasztaljuk, hogy a fejlesztések ezek megoldása nélkül indulnak be. Tudomásunk szerint nem készült el olyan ISIS-változat, amely a fenti igényeket maradéktalanul kielégítette volna. Jelenleg hiányos az ékezetes betűk kezelése (hiányoznak az *ő, ű* betűk, a meglévő ékezetes betűkre pedig nem lehet keresni). Ezeket a követelményeket programunk kielégíti.

c) Bár nagyon nehéz általános adatellenőrzési programot készíteni, vannak olyan részfeladatok, amelyek megoldhatók. Ilyen az egyediség ellenőrzése, amely talán az egyik legfontosabb ellenőrzési feladat. Említhetnénk szabadalmi adatbázisunkat is, ahol több mező egyediségére kell ügyelni. Megengedhetetlen, hogy az adatbázisba többször bekerüljön ugyanaz az alapszám vagy lajstromszám. Az egyediség ellenőrzése feltételezi, hogy a rögzített adatok azonnal bekerüljenek az adatbázisba, online karbantartás történik az invertált fájlon is. Ezek hiányoznak a Micro-ISIS-ből.

d) Rendszerünkben a felhasználó határozza meg, hogy az egyes mezők mely invertált fájlba kerüljenek. Ezzel szemben az ISIS-nél egyetlen invertált fájl van, ami nem mindig szerencsés megoldás.

e) A Micro-ISIS jelenleg maximum 32 000 rekordot képes kezelni. Bár ez látszólag igen nagy szám, véleményünk szerint ez a felhasználások során korlátként fog jelentkezni. Lehet ugyan részekre bontani az adatbázist, és több fájlban tartani az állományt, de úgy hisszük, nem ez az igazi megoldás. Az egyre terjedő, és egyre nagyobb háttértárral rendelkező mikroszámítógépek már most lehetővé teszik, hogy akár 100–150 ezer rekordot tartalmazó adatbázist építsünk. A FREETEXT rendszer jelenleg elvileg mintegy 1 millió rekordot képes kezelni, tehát esetében valóban elmondhatjuk, hogy korlátot csak a háttértár (illetve a rögzítési kapacitás) jelent.

f) A sor végére néhány apróság kívánkozik. Előny, hogy lehetőség van intervallumkeresésre, betekinthetünk az invertált fájlba a kereshető értékek közé. A felhasználó adhatja meg a szavas invertáláskor használatos szóelválasztó, az ismételhető mezőket elválasztó, valamint a kijelöléses invertálás kezdő és záró jelét. Ezek is könnyebbé teszik az adatok kezelését.

g) Rendszerünkkel szerettünk volna lépni egyet az intelligensebb információkeresés felé. Egyre sürgetőbb igény, hogy a ma már hagyományosnak mondható logikai operátorokkal történő keresés mellé, támogatására legyen valamilyen eszköz. A sok érdekes kísérlet közül az utóbbi két évben érezhetően előtérbe került, sőt több nagy adatbázis-szolgáltató központnál be is vezették azt a szolgáltatást, amit az ESA-IRS-nél a ZOOM, a Pergamon InfoLine-nál a GET parancs valósít meg. A módszer lényege, hogy adott találati halmaz elemeinek tetszőleges mezőjére statisztikai elemzés végezhető. Ennek eredményeképpen egy szó- vagy kifejezéslistát kapunk, azok előfordulási számával együtt, amely ez utóbbira rendezett, tehát a leggyakrabban előfordulók találhatók a lista elején. Az így kapott lista segítheti a további keresést, felhívhatja a figyelmet olyan szavakra, deszkriptorokra, amelyekre esetleg előre nem is gondoltunk. Kiválaszthatjuk vagy éppen kizárhatjuk a kifejezéseket a további kereséshez a lista alapján.

Míg a szokásos módszerrel végzett keresések esetén pontosan tudnunk kell, hogy milyen kifejezéseket akarunk összekapcsolni, e lehetőség segítséget ad a majdan logikai operátorokkal összekapcsolandó kifejezések, szavak kiválasztásához. A módszert annyiban fejlesztettük tovább, hogy a mezőértékek gyakoriságát még egy másik mező függvényében is vizsgálhatjuk, ti. hogy az adott érték évenként milyen számban fordul elő, azaz gyakorisága hogyan változik. A módszer természetesen emellett valódi statisztikák készítésére is lehetőséget ad.

Továbbfejlesztések

A jelenlegi rendszer ismertetése után fejlesztési elképzeléseinkről ejtünk néhány szót. Egyrészt szeretnénk minden, a Micro-ISIS által megvalósított lehetőséget beépíteni programunkba. Elsősorban az almezőket, a nem limitált mezőhosszakat és a keresőnyelvet tartjuk ezek közül fontosnak. Másrészt minden általunk ismert szöveges információkereső programnak megvan az a nagy hátránya, hogy egyetlen adatfájlt (egyetlen rekordtípust) képes csak kezelni. Ez minden rendszert nehézkessé tehet, hiszen egy adatbázisban tulajdonképpen csak egyetlen típusú dokumentum tárolható. Nyilvánvaló, hogy más adatokat kell tárolni egy könyvről, mint egy folyóiratcikkről, és az is, hogy a keresés során általában teljesen mindegy, hogy az milyen típusú. Ezt a problémát persze így-úgy át lehet hidalni, de úgy érezzük, hogy a valódi megoldás az, hogy egy adatbázis többféle rekordtípusból áll, amelyek invertált fájljai közösek és önállóak is lehetnek. Ha pl. egy adatbázis könyvekből, folyóiratcikkekből, szabadalmakból, kutatási jelentésekből áll, közös lehet a cím, ill. a kulcsszómező, de valószínűleg a szerző is. De önálló (pl. csak a folyóiratcikkekhez tartozik) a folyóirat neve: invertált fájl. Ekkor, ha a cím szerint keresünk, minden dokumentum találatként jelenik meg, amely tartalmazza az adott szót, tekintet nélkül a típusára. A több rekordtípusból álló adatbázis legnagyobb előnyét mégsem itt, hanem ott látjuk, ahol ezek a rekordtípusok kapcsolódnak egymáshoz. Erre példa egy kölcsönzési rendszer. Itt két egymással kapcsolatban álló adathalmaz van, a könyveké, valamint az olvasóké. A kölcsönzés pedig nem más, mint e kettő kapcsolata. Úgy érezzük, ha a fent vázolt program elkészül, segítségével már igen bonyolult alkalmazások, rendszerek építhetők.

ERDŐS Iván – BISZAK Sándor: A FREETEXT szöveges információkereső rendszer

A FREETEXT szöveges információkereső rendszer MS-DOS operációs rendszerű mikroszámítógépekre készült TURBO PASCAL programnyelven. A programrendszer legfontosabb tulajdonságai: változó rekord és mezőhosszúság; az adatbázist a felhasználó a saját igényei szerint hozhatja létre, ehhez a rugalmas formátumnyelv segítségével változatos adatbeviteli és megjelenítési formátumokat rendelhet; lehetőség van minden szóra, teljes mezőtartalomra, ill. a mező kijelölt részeire invertálni; rugalmas, könnyen kezelhető keresőnyelv biztosítja a releváns tételek visszakeresését; tetszőleges mezőre statisztikai elemzést végezhetünk. Külön említést érdemel, hogy képes maradéktalanul kezelni a teljes magyar (ill. igény esetén más) karakterkészletet, valamint hogy tetszőleges méretű adatbázisok működtetésére alkalmas.

* * *

ERDŐS, I. – BISZAK, S.: The FREETEXT textual information system

The FREETEXT information retrieval system was written in TURBO PASCAL language for its implementation on microcomputers under MS-DOS operating system. The features of the program system are variable record and field lengths, database definition by the user with a variety of input and display formats, inverting possiblities of all words, of complete field contents or of selected field parts, flexible and simple search language for retrieval, statistical analysis capability for any field. The program handles total Hungarian or any other character set, and it is suitable for the management of databases with any size.

* * *

ЭРДЁШ, И. — БИСАК, Ш.: Текстовая информационно-поисковая система FREETEXT

Статья знакомит с текстовой информационно-поисковой системой FREETEXT, которая разработана на программном языке TURBO PASCAL для микро-ЭВМ с операционной системой MS-DOS. Важнейшие особенности этой системы: переменная длина записей и полей. Базу данных потребитель может создать по своим требованиям, с использованием гибкого языка описания формата можно получть разные форматы для ввода и вывода данных. Для создания инвертированных файлов можно использовать каждое слово, полное поле, а также определенные части поля. (Инвертированный файл можно построить на основе каждого слова.) Гибкий командный язык обеспечивает поиск релевантных документов. Возможен также статистический анализ отдельных полей и удобно, что система может работать с полным набором венгерских (или других) букв, и применима для ведения баз данных любого размера.

ERDŐS, I. – BISZAK, S.: Das FREETEXT Informationssuchungssystem mit Text

Das FREETEXT Informationssuchungssystem ist zu den Mikrorechnern von MS-DOS Operationssystem auf TURBO-PASCAL Programmsprache hergestellt worden. Die wichtigsten Eigenschaften des Programmsystems sind die folgenden: veränderlicher Rekord und veränderliche Feldlänge; der Benutzer kann die Datenbasis nach seinen eigenen Ansprüchen zustande bringen, dazu kann er mit Hilfe der flexiblen Formatensprache abwechslungsreiche Dateneingabe- und Darstellungsformate verordnen; es ist möglich, auf jedes Wort, auf den vollen Feldgehalt bzw. auf die festgesetzten Teile des Feldes zu invertieren; die flexible, leicht behandelnbare Suchungssprache sichert die Recherche der relevanten Titel; auf das beliebige Feld kann eine statistische Analyse durchgeführt werden. Es ist besonders zu erwähnen, dass es fähig ist, den vollen ungarischen (bzw. je nach Bedarf auch anderen) Charakterbestand restlos zu behandeln, sowie dass es geeignet ist, Datenbasen von beliebigem Masse zu betätigen.

Nemzetközi Számítástechnikai Szakkiállítás először Magyarországon

Korunk meghatározó technikája a számítástechnika. Az ezen a szakterületen mutatkozó eredmények jelentős fokmérői az adott ország fejlettségének.

Hol állunk mi magyarok a világban? Milyen eredményeket értünk el? Hogyan álljuk az összehasonlítást a keleti és a nyugati versenytársakkal? Beszélhetünk-e valóságos számítástechnikai piacról Magyarországon? Menynyire segítik az ágazat fejlődését a kooperációk, az együttműködések, a keleti – nyugati találkozók?

Ma már Magyarországon is a gazdasági, a kulturális élet szinte minden területén a tervezést, a termelést, az ügyvitelt segítő korszerű gépek és programok állnak rendelkezésre. A fejlődés nálunk is rohamléptekkel halad.

Mindenképpen megérett tehát az idő a megméretésre, az első magyarországi *Nemzetközi Számítástechnikai Szakkiállítás* megrendezésére, amelyhez a Budapest Kongresszusi Központ nyújt méltó keretet *1988. október 17. és 21.* között.

A kiállítás megrendezésének gondolatával egyetért az Ipari Minisztérium, az OKISZ, a Központi Statisztikai Hivatal, a Magyar Tudományos Akadémia, a KISZ KB és a Neumann János Számítógéptudományi Társaság is. Ott lesz a kiállítók között a magyar és a nemzetközi számítástechnika számos jelentős gyártó és forgalmazó vállalata.

Bár a jelentkezési határidő lezárásáig még van idő, már eddig 72 kiállító jelentette be részvételét, 1432 négyzetméteres területre a kiállítást szervező *COMPEXPO*-nál.

COMPFAIR IRODA, 1022 Budapest, Bég u. 3–5.