Biszak Sándor
ARCANUM Databases Betéti Társaság

Az ARCTIS szöveges visszakereső rendszer

Az ARCANUM Databases Bt. munkatársai évek óta foglalkoznak szöveges információ-visszakereső rendszerek fejlesztésével, adatbázisok építésével, és ezek kompaktlemezen történő publikációjával. Elég talán a széles körben ismert PRESSDOK és NPA, vagy akár a teljes szövegű Biblia, illetve az éppen kiadás előtt álló MNB/Könyv (Magyar Nemzeti Bibliográfia) CD-ket említeni. A következőkben a fenti optikai lemezes adatbázisok alapjául szolgáló ARCTIS publikációs platformot, fejlesztésének főbb szempontjait, eredményeit ismertetjük.

Az ARCANUM 1990 közepén alakult az *Országos Találmányi Hivatal (OTH)* és magánszemélyek részvételével. Elsődleges célja az OTH számítástechnikai fejlesztésének támogatása, ezen belül is a CD-publikálás hazai meghonosítása volt. Már tevékenységünk korai szakaszában is nyilvánvalóvá vált, hogy a fejlesztések jól kamatoztathatók egyéb, nem iparjogvédelmi területen is.

A társaság megalakulásakor egyértelmű volt, hogy az eredményes működés, sikeres kiadványok csak saját fejlesztésű, teljesen kézben tartott, alakítható programrendszerrel képzelhetők el. Az azóta eltelt időszak igazolta az eredeti elképzeléseket.

Az alábbiakban részletesen ismertetjük a programrendszer lehetőségeit.

Indexkezelés (FXI)

A jó információ-visszakereső rendszer alapját egy hatékony indexkezelő rendszer alkothatja. Ezért kifejlesztettük az ARCTIS program részeként, de önmagában is használható *FXI* (FiX Index) modult. Ebben igyekeztünk minden általunk ismert jó megoldást hasznosítani. A cél (mint a név is mutatja) egy nem változó (tipikusan ilyen a CD-ROM) index hatékony kezelése volt.

Gyorsaság, nagy méret

Az első és legfontosabb cél, hogy az index gyorsan álljon elő. Egy CD-ROM-os adatbázis mérete több száz megabájt lehet. E nagy adatmennyiség indexelése, a keresőindexek előállítása nem kis feladat. Hosszas fejlesztés, optimalizálás után sikerült a sebességet olyan szintre hozni, amely állja a nemzetközi összehasonlítást is. Logikusnak tűnő ellenvetés, hogy a sebességnek nincs különösebb jelentősége, mert egyszer elő kell állítani valamilyen módon az indexet, aztán CD-ROM-ra kell tenni. Mégis a tapaszalatok azt mutatják, hogy 10–15 indexelést el kell végezni, amíg a végleges rendszer, a CD előáll, így a megvalósítás, üzemeltetés szempontjából a sebességnek hallatlan jelentősége van.

Az általunk kezelt legnagyobb adatbázis az American Petroleum Institute *APILIT* nevű, igazi nagy, a DIALOG Information Services szolgáltatón keresztül is elérhető adatbázisa. A mintegy 250–300 Mbájtos adatállomány minden mezőjét indexeljük (az igen hosszú tárgyszó és kivonat mezőket is). Az indexelés (486/33 MHz, 8 Mbájt RAM, SCSI winchesterkonfiguráción) csak 4 órát vesz igénybe. A nagynak tekinthető PRESSDOK adatbázis teljes indexelése, amely más eszközökkel nem volt elvégezhető, kevesebb mint egy óráig tart.

Bár az index nem módosítható (változás esetén teljes újraindexelés szükséges), állandóan növekvő adatbázisok esetén lehetőség van gyors indexelésre. Az APILIT adatbázis havonta bővül, a meglevő mintegy 130 ezer rekord havonta mintegy 1500 rekorddal egészül ki. Az index aktualizálásának ideje mintegy fél óra (a fenti konfiguráción).

Változó hosszúságú indexek, tömörített tárolás

A fejlesztés másik nagyon fontos pontja a változó hosszúságú keresőelemek (kulcsok) kezelése. Sok visszakereső rendszer tipikus hibája, hogy a keresőelemek végét adott hosszban levágja.

Rendszerünkben a kulcsok teljes hosszukban tárolódnak. Nem arról van szó azonban, hogy egy igen hosszú hely van lefoglalva minden keresőelemnek, hanem hogy minden keresőelem annyi helyet foglal el, amennyi a hossza. Egy egybetűs keresőelem 1, egy hatvanbetűs 60 karakter helyet foglal el. Jelenleg csak a képernyő szélessége, az azon történő megjelenítés miatt maximáltuk 75 karakterben a keresőelemek hosszát.

Rendszerünk másik fontos tulajdonsága a tömörített tárolás. Ez azt jelenti, hogy egy kulcs csak azt tartalmazza, hogy mennyiben azonos az előzővel, és mi az eltérés. Lássunk egy példát:

Kulcs	Tárolása
e	0e
előállít	1lőállít
előállítás	8ás
előállítására	10ára
előállítását	11t

A fenti tárolás különösen a ragozó magyar nyelvnél hatékony, hiszen igen gyakran fordul elő egy szavas indexben, hogy csak a ragokban térnek el az indextételek egymástól. A változó hosszúságú és tömörített indexkezelés alkalmazásával igen kis méretű indexeket hozhatunk létre. Ismét jöhet az ellenvetés, hogy egy szinte végtelennek tekinthető CD-ROM-on miért van fontos szerepe a helytakarékos tárolásnak. Mégis azt kell mondanunk, hogy éppen a nagyon lassú CD-n lényeges ez, hiszen egy bonyolultabb kérdés kielégítésekor az index végigolvasásának sebessége kizárólag annak méretétől függ.

Karakterkészlet

A magyar nyelv esetében különös jelentősége van annak, hogy az indexkezelő minél nagyobb szabadságot adjon a felhasználónak a karakterek kezelését, sorrendjét illetően. Bár egyre elterjedtebb az IBM 852 kódtáblázat (a DOS 5.0 már ezt támogatja kelet-európai szabványként, sőt ma már magyar szabványként is megjelent, és adatbázisainkon is ezt a karakterkészletet használjuk), a programunk lehetőséget ad tetszőleges kódkészlet használatára, és a felhasználó adhatja meg az indexben a karakterek sorrendjét. Adatbázisaink nagy részben az indexben csak az Á, É, Ö, Ü karaktereket különböztetjük meg (tehát hosszú magánhangzó helyett a megfelelő rövid párjára konvertálunk), ugyanakkor van ellenpélda is. A Magyar Szabványügyi Hivatal MSZHIR adatbázisánál teljes ékezetes index van (tehát a hosszú magánhangzók is benne vannak az indexben, pl. a BOR, BÓR szavakat megkülönböztetjük).

Betekintés az indexbe (EXPAND, BROWSE)

A létrejött index segítségével végezhetjük el kereséseinket. Legalább ilyen fontos azonban, hogy az indexnek támogatnia kell a keresőkérdés megfogalmazását is. Minden eszközt meg kell adni a keresőnek, hogy megfogalmazhassa kérdését. Az indexbe való betekintést (amit EXPAND, BROWSE kifejezésekkel illetnek) igen fontos műveletnek tartjuk, és a fejlesztésnél nagy hangsúlyt fektetünk rá. A program lehetőséget ad arra, hogy az indexen a kurzormozgató billentyűkkel navigáljunk, az általunk begépelt keresőelem környezetére ugorjunk stb. Emellett a szokásosnak mondható műveletek mellett különösen az összetett szavak keresését támogatjuk azzal, hogy lehetőség van minta (pattern) megadására. A minta egy csonkolási jeleket is tartalmazó karaktersorozat (*: tetszőleges, ?: 0 vagy 1, !: pontosan 1 karakter helyettesítését jelenti), a program pedig csak a mintának megfelelő kulcsokat mutatja meg.

A *!ASSZONY* minta jelentse azokat a szavakat, amelyekben nem a szavak elején fordul elő az ASSZONY szó. A Biblia szövegéből véve a példát, az illeszkedő szavak: FEJEDELEMASSZONY, NAPAASSZONY, KIRÁLYASSZONY stb. E lehetőség nagy segítség nemcsak az összetett szavak keresése esetén, hanem pl. kifejezéses indexek esetén is, amikor a kifejezésnek csak egy részét ismerjük. A minta szerinti kiértékelés nem képzelhető el változó hosszúságú index nélkül, hiszen ha a keresőelemek vége levágásra kerülne, hamis eredményt kapnánk.

Adatkezelés (FXD)

Az ARCTIS rendszer másik tartópillére az adatok, rekordok, mezők, almezők kezelését végző, önmagában is használható *FXD* (FiX Data) modul. Mint az FXI, az FXD is nem változó adatbázisok kezelésére van optimalizálva. Kifejlesztésekor fő célunk az volt, hogy minél bonyolultabb adatszerkezet megvalósítása legyen lehetséges.

Egy szöveges visszakereső rendszernél az is alapvető követelmény, hogy az indexek mellett a mezők is változó hosszúságúak legyenek. Ugyanilyen elengedhetetlen az ismétlődő mező kezelése (nyilvánvaló példa az egy könyvhöz tartozó több szerző).

Véleményünk szerint ugyanilyen fontos, ugyanakkor még nemzetközileg igen jól jegyzett, jelentős piacokat kézben tartó CD-s szoftvereknél is elhanyagolt terület az almezők kezelése. Almezőről akkor beszélünk, amikor egy ismétlődő mező több, szerkezetileg elkülöníthető (és elkülönítendő) részt tartalmaz. Azt gondoljuk, hogy nehezen képzelhető el olyan bibliográfiai adatbázis, amely almezők nélkül épülne fel.

Vegyünk egy szabadalmi adatbázist, amelynek feltaláló mezője névből, foglalkozásból és lakcímből áll. Azon lehet vita, hogy a lakcím mezőt bontsuk-e további almezőkre (irányítószám, város, ország stb.), az azonban nyilvánvaló, hogy a fenti három adatelemet szerkezetileg el kell különíteni. Az elkülönítés indexelés és megjelenítés szempontjából is alapvető fontosságú, hiszen indexelni csak a névre akarunk.

Másik egyszerűnek látszó példa: a könyvek kiadási adatai, melyek helyből és kiadóból állnak (és ismételhetők, hiszen léteznek közös kiadások is). Ugyanakkor a hely almező ismétlődhet is, hiszen gyakran fordul elő, hogy több székhelye van kiadónak (gondoljunk csak az Elsevier kiadóra). Ez az eset tipikus példája az

ismétlődő almező szükségességének. Az FXD modulban mind az almezők, mind azok ismétlődése kielégítően megoldott.

Az adatkezelésben is van jelentősége a tömörítésnek. A legelterjedtebb, és általunk is alkalmazott Huffmann-féle tömörítési eljárással 40–50%-os tömörítési arány érhető el. Ezzel lehetővé válik, hogy egy CD-n 600 Mbájtnál nagyobb adathalmazt is tároljunk, illetve ha az adatbázist nem CD-n használják, kisebb az adatbázis winchesterigénye. A Huffmann-féle tömörítés lényege, hogy a karaktereket a szokásos 8 bites kódolással szemben egy változó hosszúságú bitméreten tárolja, a gyakori karaktereket 2–3 biten, majd az egyre csökkenő gyakoriságú karaktereket egyre nagyobb bithosszúságon. A kódolás kiindulási pontja tehát egy gyakoriságvizsgálat, majd ennek alapján egy ún. Huffmann-táblázat megalkotása.

Az FXD modul adatkezelése rekord, mező, almező típusú, tipikus alkalmazásai információ-visszakeresést szolgáló bibliográfiai adatbázisok. Emellett létezik egy teljes szövegű adatkezelő modul is, melynek tipikus példája a CD-n megjelent Biblia adatbázis. Itt nem beszélhetünk rekordokról, mezőkről, ez az alkalmazás egy teljesen más, ún. teljes szövegű adatkezelést igényel. Ezen adatkezelés részletes ismertetése nem tárgya jelen cikkünknek, annyit azért megemlítünk, hogy itt a bekezdések (versek) jelentik az alapegységet. Ugyanakkor jelentősége van azok egymásmellettiségének, illetve alá-, fölérendeltségi viszonyainak (könyvek, részek, versek, azaz hierarchikus szerkezet). Másik fontos eleme a bekezdések közötti keresthivatkozások kezelése (a Bibliában a versek mellett nagy számban fordulnak elő a Biblia más részeire való utalások).

A teljes szövegű adatkezelő rendszer segítségével készült ez az IPC:CLASS CD-ROM, a Szellemi Tulajdon Világszervezete (WIPO), valamint a Német, Spanyol és Magyar Szabadalmi Hivatal közreműködésével. Az adatbázis a Nemzetközi Szabadalmi Osztályozás ötnyelvű (angol, francia, német, spanyol és magyar), több kiadást tartalmazó CD-ROM-ja. Ennek részletes ismertetése a *Szabadalmi Közlöny és Védjegyértesítő* mellékleteként megjelenő *Iparjogvédelmi Szemle* 1992. októberi számában jelent meg. Ugyanez a teljes szövegű adatkezelő rendszer az alapja a WIPO iparjogvédelmi jogszabály CD-jének, amely a világ országainak iparjogvédelmi jogszabályait tartalmazza.

Adat- és indexdefiníció

A fenti alapeszközök (FXI, FXD) a programozót segítik az alkalmazás létrehozásában. Az ARCTIS programrendszer definíciós része segítségével hozhatjuk létre adatbázisunkat. Ennek részei:

- adatdefiníció: adatbázis-szerkezet létrehozása (mezők, almezők),
- indexek definiálása,
- formátumok létrehozása.

Adatdefiníció

Az adatbázis szerkezetének létrehozását alapos elemzőmunka előzi meg, melynek célja a szerkezet feltárása. Még ha az ARCTIS segítségével már meglevő, valamilyen adatbázis-kezelővel korábban felépített adatbázis publikálása a feladat, akkor sem hanyagolható el ez a lépés. Tipikus eset sok adatbázisnál, hogy a használt eszköz nem képes almezőket kezelni. Ezt a problémát vagy fix hosszúságú adatelemekkel vagy elválasztójelekkel kezelik. Egy példa erre a Magyar Szabványok adatbázisa, amelyben a KÜLFÖLDI SZABVÁNY mező első három karaktere az átvétel módjára utal, és ezután következik a magyar szabvány alapjául szolgáló külföldi szabványazonosító. Például: EQU ISO 1234, amelynek jelentése, hogy a magyar szabvány megegyezik az ISO 1234-gyel, fordítása annak. Az ARCTIS rendszerben ez a mező két almezőből, TÍPUS (EQU) és SZABVÁNYSZÁM (ISO 1234) almezőből áll.

Gyakori eset a kódok alkalmazása az adatbázisokban, mint pl. a PRESSDOK-ban a tárgykörök azonosítására szolgáló jelzetek. Itt is segít az almezős szerkezet, szokásosan a jelzet mellett a magyar és angol (esetleg más) nyelvű feloldások szerepelnek az adott mező almezőiként. Így lehetővé válik, hogy a nehezen értelmezhető kód helyett (mellett) a feloldását is használjuk, megjelenítsük, indexeljük, méghozzá több nyelven.

Az alapos elemzőmunka után az adatszerkezet leírása következik, amelyben a mezők, almezők nevét, azonosítóját (tag) és jellemzőit (pl. ismétlődés, indexelés típusa) kell megadnunk.

Indexdefiníció

A megalkotott adatszerkezetre építkezve, de attól bizonyos mértékig elválasztva, hozhatjuk létre az adatbázis indexdefinícióját. Ebben igyekeztünk az általunk ismert indexelési típusokat és módszereket megvalósítani, összegezni.

Fontos tulajdonság, hogy egy indexbe több mező, almező építkezhet, illetve egy mező, almező többféleképpen indexelhető. Nincs akadálya annak, hogy egy indexbe egy mező, almező szavasan is és kifejezésre is indexelve legyen. (Ez tipikus pl. testületek esetén, ahol mindkét indexelési típusnak létjogosultsága van.) A közös indexek lehetőséget adnak arra, hogy eltérő mezőkben tárolt adatok közösen legyenek kereshetők. Az indexekhez tiltott szavakat adhatunk meg (stopword), amelynek nem kerülnek be az indexbe. Ezt a lehetőséget nemcsak szavas, hanem kifejezéses indexnél is használhatjuk.

Az indexképzés egyik alaptípusa a szavas indexelés, ahol azonban megadhatjuk a szóelválasztó karaktereket is. Ennek segítségével megoldható pl. az elég bonyolult ETO-mező indexelése. A : és + jeleket szóhatároló jeleknek definiálhatjuk (de pl. a szóközt nem), így az egyébként egy egységenként szereplő összetett ETO-jelzetek minden alapelemüknél kereshetők lesznek.

A másik alaptípus a kifejezéses index, amelyen egy mező, almező teljes, változatlan tartalmának indexelését értjük. Ennek specialitása, hogy megadható tiltott rész (tipikusan ilyenek lehetnek a névelők), amelyek az indexeléskor kimaradnak.

A fentieken kívül képezhetünk indexet mezők összegére is. Ez általában egy adott mezőben szereplő almezők összeadását jelenti. (A szabadalmi példánál maradva, indexeljük a felataláló mezőt a név és foglalkozás összegére.)

Tapasztalataink azt mutatják, hogy az alkalmazások gyakran igényelnek speciális (általánosan talán meg sem fogalmazható, csak az adott alkalmazásra jellemző) indexelési eljárásokat. Ezek beillesztése, beprogramozása az alkalmazás létrehozásának a része.

Adatreferencia

Az indexdefiníció részeként érdemes megismerkedni az adatreferencia fogalmával is. Egy keresőindex (invervált fájl) alapvetően két logikai részből áll. Egyrészt az indexben található keresőelemek (szavak, kifejezések), valamint azok előfordulásait (rekordazonosító) tartalmazó adatreferencia. Az ARCTIS rendszerben az adatreferencia tartalmazza, hogy az adott keresőelem mely rekordban, mezőben, almezőben, ismétlődésben, szavas indexelés esetén pedig azt is, hogy hányadik szóként fordul elő. Sok rendszer csak a rekordsorszámot tartalmazza. Az adatreferenciában található adatok egyértelműen meghatározzák, hogy milyen finomságú kereséseket lehet elvégezni. Ha csak rekordsorszám van, nem képzelhető el (legalábbis gyorsan) a szomszédos szavak keresése (proximity search).

Az adatreferenciák a posting fájlban találhatók, amely teljes egészében, fizikailag is elválik az indextől, annak méretét nem növeli feleslegesen. Az adatreferencia mérete optimalizált, mindig az adott alkalmazáshoz van igazítva. Például, ha nincs almezős szerkezet, nem tartalmaz almezőazonosítót, illetve a rekord-, mezősorszámokra felhasznált bitek száma a kész adatbázis statisztikai elemzése alapján minimalizált. Az adatreferencia mérete különösen fontos kereséskor, mivel ennek sebessége nagymértékben attól függ, hogy mennyi találat (adatreferencia) fér el a számítógép memóriájában.

Keresés

Miután megismerkedtünk az ARCTIS adatbázisok szerkezetével, térjünk át a keresésre. A keresés során lehetőségünk van betekinteni az indexbe (EXPAND), onnan keresőelemeket választhatunk ki. A keresés során az EXPAND műveletnél ismertetett csonkolási, maszkolási karaktereket használhatjuk. Lehetőségünk van tehát nemcsak jobbról csonkolni a keresőelemet, hanem balról is, vagy akár annak belsejében. Kereshetünk tehát szóvégződésekre, vagy összetett szavak második tagja alapján is.

A keresésünket megfogalmazhatjuk kereső űrlapon (támogatott keresés), vagy keresőkifejezés formájában (szakértői, parancsmódú keresés). Lehetőségünk van a kérdés lemezre mentésére, valamint egy elmentett kérdés betöltésére is.

Logikai és helyzeti operátorok

A keresés során operátorokat használhatunk. A megszokottnak (és kötelezőnek) mondható AND, OR, NOT operátorok mellett helyzeti operátorok segítik a pontosabb keresést. Ezek egyrészt szavak egymásmellettiségét biztosítják, másrészt azonos mezőn, illetve azonos ismétlődésen belüli feltételt jelentenek. Az (nW), illetve (nN) operátorok esetén az operátor két oldalán álló szavak között maximum n számú egyéb szó állhat, nN esetén a két szó sorrendje tetszőleges. Az (F) operátor esetén a két szónak azonos mezőben kell előfordulnia, (S) esetén azonos ismétlődésben.

Az (F) operátor nagyon hasznos lehet, ha az adatbázisunk több szöveges mezőt tartalmaz (cím, kivonat stb.). Találatot csak akkor kapunk, ha a két szó ugyanazon mezőben fordul elő (nem lesz találat, ha az egyik szó a címben, a másik a kivonatban fordul elő). Az (S) operátor almezők megléte esetén nélkülözhetetlen, ezzel tudjuk az adatok összetartozását biztosítani. Maradva a szabadalmi példánál (feltaláló mező, név, város almező), ha a budapesti illetőségű Nagy nevű feltalálókat keressük, az (S) operátor biztosítja, hogy az a rekord, ahol egy budapesti Kis, és egy pécsi Nagy a feltaláló, ne jelentkezzék találatként.

Találati halmazok kezelése és megjelenítése

A keresés eredményeként találati halmaz jön létre, a képernyőn a találatok számát kapjuk meg. A létrejött találati halmazok a keresés ideje alatt megőrződnek, későbbi keresőkérdésben felhasználhatók. A találati halmaz képzésekor az FXI modulban kifejlesztett gyors indexelési technikát használjuk fel, így a keresés sebessége nagyon jónak mondható. Még igen nagy (8–10 ezer) találati halmaz képzése is megtörténik 5–6 másodpercen belül, bár a keresés sebessége nagymértékben függ a rendelkezésre álló szabad memóriától, illetve természetesen a számítógép sebességétől.

A meglevő találati halmazok közül bármelyik megjeleníthető, szükség esetén későbbi felhasználásra elmenthető. A megjelenítés első lépéseként egy rövid találati listát kapunk (minden rekord egy sor). E listán navigálhatunk, egyes tételeket kijelölhetünk, illetve kérhetjük a teljes rekordot. A teljes rekord megjelenítésekor a találatot okozó szó, kifejezés kivilágításra kerül a könnyebb eligazodás érdekében (highlighting).

A formátumnyelv segítségével viszonylag egyszerűen hozhatunk létre formátumokat. A mezők, almezők listáján kijelölhetjük, hogy mely mezőket kívánjuk megjeleníteni, és melyeket nem. Megadhatjuk, hogy az adott mező előtt, illetve után milyen karaktersorozat jelenjék meg, mi legyen az ismétlődő mezők között. Pozicionálhatunk adott oszlopra (tabulátorfunkció), illetve beállíthatjuk a jobb és bal margót. A bonyolultabb formátumok létrehozásához feltételeket definiálhatunk, melyek mező meglétére, hiányára, vagy akár mezőtartalomra is vonatkozhatnak.

Bemenő adatok

Az ARCTIS rendszer kiindulópontjaként azt feltételeztük, hogy egyre nagyobb számban állnak elő aktualizált, publikálásra kész adatbázisok. Az évek folyamán minden adatbázis-építőnél kialakult az általuk preferált, az igényeket legjobban kielégítő aktualizálási rendszer. Feladatunknak azt tekintettük, hogy a nemzetközi adatcserében használatos ISO 2709 mellett a legelterjedtebb adatbázis-kezelők által támogatott export (csere-) formátumokat is támogassuk. A tapasztalat azt mutatja, hogy Magyarországon az ISO-formátum mellett a TEXTAR csereformátuma a legelterjedtebb, így e kettő fogadására készültünk fel. Ritkán előforduló eset az ún. „tag" (a mezőt azonosító címke) output, amely a TEXTAR csereformátumához nagyon hasonló szerkezetű.

A különböző input formátumok fogadásánál a gyakran előforduló eseteket igyekeztünk általánosan megoldani. Tipikusan ilyen a karakterek kérdése. Az inputokon jellegzetes az adatok, az ékezetes karakterek gizmo kódolása, míg adatbázisainkban az egyre terjedő, végre igazi szabvánnyá váló IBM 852-es karakterkészletet használjuk. Több adatbázisban használatos kódolt mező, amely egy CD-s adatbázisban általában nem megengedhető. Ilyenkor általában a kódok magyar és angol feloldásait is megadjuk. Mindkét feladat az inputkonverziókor végzendő el.

Számtalan példa akad annak illusztrálására is, hogy milyen sokféle egyedi konverziós igény vetődik fel egy adatbázis felépítésénél. Így minden alkalmazás – az általánosan használatos konverziók mellett – egyedi elemeket is tartalmaz.

Egyéb funkciók

A fent ismertetett általános eszköztár képezi az alapját egy CD kiadásának. Az általános, minden alkalmazásban fellelhető elemek mellett mindig vannak csak az adott alkalmazásban előforduló funkciók. Ha a speciális funkció elkészülte után bebizonyosodik, hogy az általánosan is használható, megtörténik a programrész beépítése az alaprendszerbe, tovább gazdagítva annak lehetőségeit. Érdemes a rendelkezésre álló, de nem minden alkalmazásban jelen levő funkciókról külön is szólni.

Képek kezelése

A CD-ROM nagynak mondható kapacitása lehetővé teszi, hogy a bibliográfiai adatbázis mellett az eredeti dokumentum képmása is tárolódjék. Ugyanakkor érdemes óvatosan kezelni ezt a kérdést, mert a rendelkezésre álló több mint 600 Mbájt csak mintegy 10 ezer fekete-fehér (nem tónusos) A4-es oldal tárolására elegendő, 300 dot per inch felbontásban. Ez a felbontás a jelenleg használt lézernyomtatók felbontása. Ez a tízezres szám is csak igen jó tömörítőeljárás esetén igaz, ugyanis egy, a fenti paraméterekkel rendelkező kép tömörítetlen mérete több, mint 1 Mbájt. A jelenleg széles körben elterjedt legjobb tömörítési eljárás (CCITT G4) segítségével a méret az eredeti huszadrészére csökkenthető (egy átlagos képoldal 50 kbájt helyet foglal el). A jelenleg használatban levő faxok a G3-as tömörítési eljárást használják, ennek egy továbbfejlesztett (a méretet felére csökkentő) változata a G4-es tömörítés. Azt lehet mondani, hogy az ún. képmás (fakszimile) CD-k esetén a G4-es tömörítés vált szabvánnyá, így természetesen mi is ezt használjuk. Hangsúlyozzuk, hogy a fenti adatok kizárólag fekete-fehér képekre vonatkoznak, árnyalatos (szürkefokozatú) vagy színes képek esetén egészen mások a jellemző számok.

Adatbázisunkban a képek külön fájlban találhatók, az oldalak az alaprekordhoz egy azonosítóval kapcsolódnak. A bibliográfiai rekord megjelenítése után lehetőségünk van arra, hogy a dokumentumot a képernyőn megjeleníthessük. A képet nagyíthatjuk, navigálhatunk rajta, lapozhatunk az oldalak között, szükség esetén pedig kinyomtathatjuk a dokumentumot.

A képkezeléssel kiegészített ARCTIS segítségével adtuk ki a Magyar Szabványügyi Hivatallal közösen az MSZHIR adatbázist, amely mintegy 3000 oldal szabványt is tartalmaz. Kísérletképpen a PRESSDOK adatbázis 93/II-es száma néhány HVG-cikk képmását is tartalmazza.

A tervek között szerepel a teljes szabványállomány megjelenítése, amely mintegy 20 CD-t jelentene.

Rekordkapcsolatok

Bibliográfiai adatbázisokban viszonylag elhanyagolt terület a rekordok közötti kapcsolatok kezelése. Jó példa erre az NPA adatbázisban található „előzménye/folytatása" kapcsolat. További jellegzetes példa az MNB adatbázisban a többkötetes könyvek esete. Kezelése ugyanis úgy történik, hogy külön rekordtípust alkotnak a többkötetes könyvek közös adatai, illetve a kötetadatok.

Rendszerünkben van arra lehetőség, hogy az aktuális rekord kapcsolataiból új találati halmazt hozzunk létre. Felfogásunkban a rekordkapcsolatok indexen keresztül valósulnak meg. Egy kapcsolattípus nem más, mint egy vagy több mezőből képződő keresőkérdés, majd egy adott indexen történő visszakeresés.

Tekintsük példaként a többkötetes könyvek esetét. A kötetadatok rekordtípusban létezik egy mező, amely a közös adat rekordazonosítóját (KOZOS nevű mező) tartalmazza, ezenkívül minden rekord (kötetrekord is) tartalmazza a rekordazonosítót (AZON mező). Az ISBN-indexet a rekordazonosító (AZON) és a közös adatra mutató mező (KOZOS) alapján képeztük. A kapcsolatot leíró keresőkérdést ezután AZON or KOZOS formában definiáljuk, a keresést pedig az ISBN-indexen kell végrehajtani. Az algoritmus a következő: „vedd ki" az aktuális rekordból az AZON és a KOZOS mezőből is az adatokat, kapcsold őket össze OR operátorral, majd tedd elé az ISBN „tag"-et, és az így létrejött keresőkérdést futtasd le. Könnyen belátható, hogy ezzel a módszerrel mind közös adatból, mind kötetadatból kiindulva megkapjuk az adott többkötetes mű összes kötetét a közös adataival együtt.

Nyelvi verziók

A fejlesztés során a többnyelvű felhasználói felület alapkövetelmény volt, ami nem csoda, hiszen a program egyik legelső alkalmazása egy adataiban és felhasználói felületében is ötnyelvű adatbázis, a már fent említett IPC:CLASS nevű WIPO-kiadvány volt. A többnyelvűség azt jelenti, hogy a „helpek" programüzenetek, és „helpjei" szerkeszthetők, ami lehetővé teszi, hogy lefordítsuk tetszőleges (angol, francia, német, spanyol) nyelvre a programüzeneteket, helpeket. Ha létrehozzuk a nyelvfüggő fájlokat (ilyenek lehetnek még a formátumok is), a felhasználónak lehetősége nyílik arra, hogy az elkészített menüből maga válassza ki a neki tetsző nyelvet.

Továbblépés, fejlesztési irányok

A létrehozott programrendszer alkalmas arra, hogy segítségével színvonalas, nemzetközileg is versenyképes CD-kiadványok készüljenek. Ugyanakkor jól láthatók a továbblépés lehetőségei, sőt kényszerei. Nem kerülhető meg a Windows-környezet, amely egyre inkább szabvánnyá válik a PC-k világában. Azt gondoljuk, hogy lassan a magyar könyvtároskörnyezet is felkészült erre az új kihívást jelentő világra. De nem szabad elfelejteni, hogy komoly gépi háttér szükséges a Windows futtatásához, ajánlott a 4 Mbájtos memóriával ellátott 386-os számítógép VGA-monitorral.

A Windows-környezet a barátságos és szabványos kezelőfelület mellett sokoldalú eszközt is ad a programozó kezébe az adatok megjelenítéséhez. A grafikus környezetnek köszönhetően szebbé, életszerűbbé válik a program felhasználói felülete. A grafikus környezetből következik a jelenleginél sokkal színvonalasabb képkezelés, de ne feledkezzünk meg a multimédia lehetőségéről sem.

Bízunk abban, hogy az ARCTIS jó tulajdonságait megtartva, és a Windows-környezet előnyeit kihasználva, színvonalas publikációs platformmal állhatunk az adat-előállítók rendelkezésére.

Beérkezett: 1994. május 2-án.

Új elnevezés

A földművelésügyi miniszter határozata értelmében az *Országos Mezőgazdasági Könyvtár* új elnevezése 1994. augusztus 1-jétől:
Országos Mezőgazdasági Könyvtár és Dokumentációs Központ

Álláshirdetés