OPTIKAI KARAKTEROLVASÓ RENDSZEREK KÖNYVTÁRI ALKALMAZÁSA

Telcs András
MTA Könyvtára

A könyvtári számítógépes rendszerek kialakításának eddigi akadályai közül egy elhárulni látszik: kisebb-nagyobb számítógépek jutnak mind több könyvtárunk birtokába. A helyi szellemi erők és az anyagi háttér kérdése most már az, hogy milyen programok fogják az egyes könyvtári munkafolyamatokat megkönnyíteni. Mint *Ungváry Rudolf* a Magyar Könyvtárosok Egyesülete rendezésében 1987-ben tartott előadásában kifejtette, semmit nem ér a számítógép program, mégpedig jó program nélkül. Tegyük azonban mindjárt hozzá, hogy adatok nélkül sem a számítógép, sem a program nem ér sokat. Magyarországon egyelőre még kevés a géppel kezelhető bibliográfiai adatok, adatbázisok száma. Ahhoz tehát, hogy számítógépeink, programjaink megkönnyítsék a könyvtári munkát és növeljék az olvasóknak nyújtott szolgáltatások színvonalát, a katalógusok adatait gépre kell vinni. Ez pedig nem könnyű feladat. Abban a szerencsés helyzetben vagyunk, hogy a technikai feltételek lassan beérnek a feladat megvalósítására, jó ötletek közül válogathatunk, hogyan is fogjunk hozzá. Talán az is a javunkra válik, hogy nagy könyvtáraink a könyvtárgépesítési program elején tartanak, s még nem kötelezték el magukat egyik vagy másik adatrögzítő eljárás mellett. A jelen írás segítséget kíván nyújtani a célravezető döntés meghozatalához. Biztatásul az *1. ábrán* néhány nehezen olvasható katalóguscédulát mutatunk be, a *2. ábrán* pedig azt az eredményt, amelyet az OPTIRAM/LIBPACK* rendszer (Anglia) segítségével készítettek róluk. A számítástechnika jelenlegi fejlettsége lehetővé teszi régi katalóguscéduláink adatainak gépi úton való elolvasását és a bibliográfiai adatbázisba való betöltését. Így katalógusaink visszamenőleges gépre vitele nem látszik reménytelen, távoli vállalkozásnak.

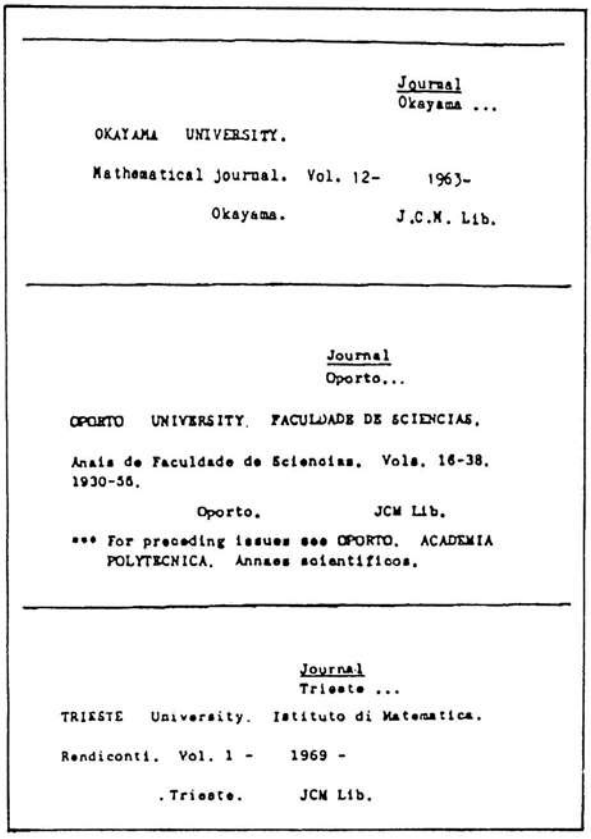
Az itt kifejtett javaslatunk az *Informatikai Infrastruktúra Fejlesztés (IIF)* keretében folyó könyvtárgépesítési munkálatok egyik alapvető, általánosan jelentkező feladatára, a meglévő katalógusok adatainak rögzítésére vonatkozik.

Tisztában vagyunk azzal, hogy a megvalósításra váró feladat igen nagy igényű még akkor is, ha csak egy máshol már működő rendszer adaptálására kerül sor. Nem térünk ki most a rendszerből kínálkozó egyéb felhasználási lehetőségekre (pl. CAD – computer aided design – céljaira tervrajzok beolvasása vagy egy-egy szótár anyagának feldolgozása), mert figyelmünket a könyvtári katalógusok kérdéseire összpontosítjuk.

A javaslat lényege

Az IIF program egyik fő célja a K+F tevékenység információs igényeinek magasabb szinten való kielégítése. A könyvtári állományokat feltáró eszközök egyelőre elhanyagolható része van számítógéppel kezelhető adathordozón. Ahhoz, hogy ezek a katalógusok és egyéb nyilvántartások megfelelhessenek az IIF követelményeinek, jelentős információs

* A szerző a tanulmány leadása után értesült arról a LIBPACK Computer Services vezetőjétől, hogy a program fejlesztése sikertelenül zárult. Az OPTIRAM-ot két egyetem könyvtárában is kipróbálták, de az eljárás nem bizonyult elég hatékonynak. Az utólagos javítások túl sok időt és költséget emésztettek föl. – Mindezek ismeretében adjuk közre javaslatunkat, felhívva a figyelmet az ismeretes és nem kis nehézségekre, másrészt arra, hogy 1985 óta jelentős fejlődés következett be (pl. a CD-ROM megjelenése), új megoldások is születtek, amelyek alkalmazása növelheti a javasolt rendszer várható teljesítményét.



Journal
Okayama ...

OKAYAMA UNIVERSITY.

Mathematical journal. Vol. 12- 1963-

Okayama. J.C.M. Lib.

Journal
Oporto...

OPORTO UNIVERSITY. FACULDADE DE SCIENCIAS.

Anais de Faculdade de Sciencias. Vols. 16-38. 1930-56.

Oporto. JCM Lib.

*** For preceding issues see OPORTO. ACADEMIA POLYTECNICA. Annaes scientificos.

Journal
Trieste ...

TRIESTE University. Istituto di Matematica.

Rendiconti. Vol. 1 - 1969 -

.Trieste. JCM Lib.

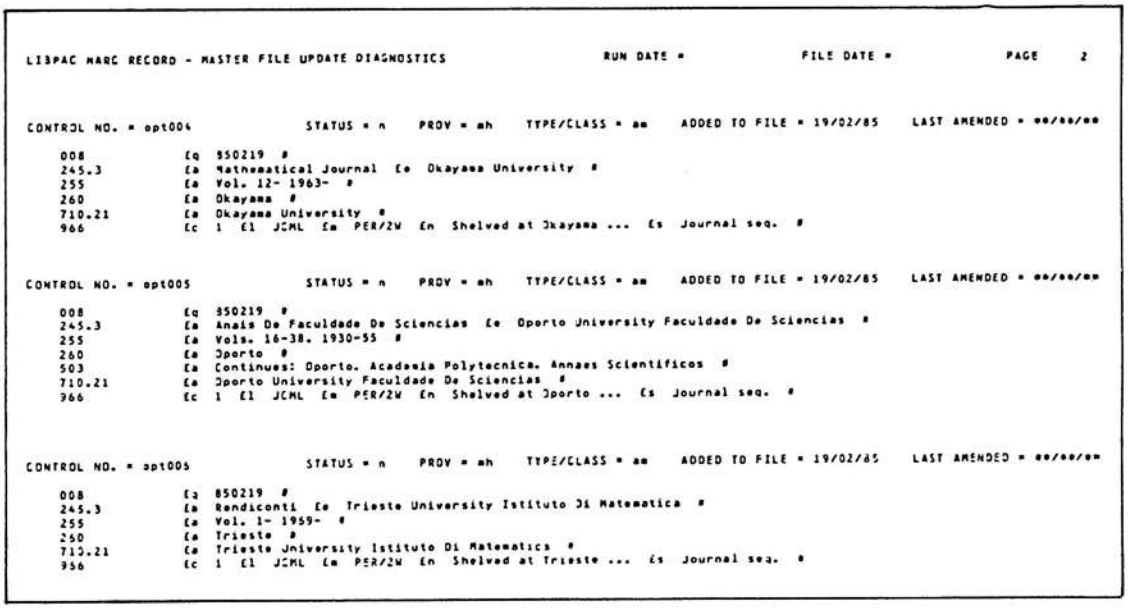
1. ábra

vagyonukat előbb vagy utóbb számítógépes adathordozóra kell vinni. Ez a feladat pedig csak egy adaptív, intelligens és interaktív adatfelviteli rendszer kifejlesztésével oldható meg. A rendszer a következő főbb elemekből állna:

- mechanikus adagoló,
- optikai karakterolvasó berendezés, amely kb. 2000 karakter/s, azaz 4 katalóguskártya/s teljesítményre képes,
- adaptív karakterfelismerő program,
- a bibliográfiai és a formai elemeket elhatároló adattárolási program,
- nyelvi és bibliográfiai háttér-adatbázisok és kezelőprogramjaik,
- helyi hálózat a többterminálos adatrögzítéshez,
- az adatjavítás és -rögzítés hálózati programja.

A fenti komponensekből álló rendszernek képesnek kell lennie arra, hogy

- kvázihomogén írásképű, rossz minőségű kártyákat is elolvasson;
- a bevitt adatokat konvertálható struktúrában tárolja;
- a feldolgozást 60—70%-ban teljesen automatikusan, 20—25%-ban ellenőrzéssel, ill. apró javítással végezze el, s csak a tételek 5—10%-ánál legyen szükség az adott tétel újbóli rögzítésére részben vagy egészben;
- olyan sebességet produkáljon, amely lehetővé teszi 1—2 milliónyi tétel egy évnél rövidebb idő alatt való feldolgozását (kb. 2000 katalóguskártyát egy óra alatt, vagy 2 másodpercenként kb. egy kártyát).



LIBPAC MARC RECORD - MASTER FILE UPDATE DIAGNOSTICS RUN DATE = FILE DATE = PAGE 2

CONTROL NO. = opt004 STATUS = n PROV = mh TYPE/CLASS = am ADDED TO FILE = 19/02/85 LAST AMENDED = 00/00/00

008 |a 850219 #
245.3 |a Mathematical Journal |e Okayama University #
255 |a Vol. 12- 1963- #
260 |a Okayama #
710.21 |a Okayama University #
966 |c 1 |l JCML |m PER/2W |n Shelved at Okayama ... |s Journal seq. #

CONTROL NO. = opt005 STATUS = n PROV = mh TYPE/CLASS = am ADDED TO FILE = 19/02/85 LAST AMENDED = 00/00/00

008 |a 850219 #
245.3 |a Anais De Faculdade De Sciencias |e Oporto University Faculdade De Sciencias #
255 |a Vols. 16-38. 1930-55 #
260 |a Oporto #
503 |a Continues: Oporto. Academia Polytecnica. Annaes Scientificos #
710.21 |a Oporto University Faculdade De Sciencias #
966 |c 1 |l JCML |m PER/2W |n Shelved at Oporto ... |s Journal seq. #

CONTROL NO. = opt006 STATUS = n PROV = mh TYPE/CLASS = am ADDED TO FILE = 19/02/85 LAST AMENDED = 00/00/00

008 |a 850219 #
245.3 |a Rendiconti |e Trieste University Istituto Di Matematica #
255 |a Vol. 1- 1969- #
260 |a Trieste #
710.21 |a Trieste University Istituto Di Matematica #
966 |c 1 |l JCML |m PER/2W |n Shelved at Trieste ... |s Journal seq. #

2. ábra

A feldolgozás módozatait meghatározzák:

- az input fizikai hordozója, jellemzői,
- az input minősége, mennyisége,
- a kívánt gépi output fizikai és logikai paraméterei,
- a megcélzott felhasználási mód,
- az anyagi lehetőségek,
- egyéb szervezeti és irányítási feltételek.

A fenti szempontok alapján lehet választani aközött, hogy számítógéppel segített kézi adatrögzítés vagy karakterbeolvasáson alapuló interaktív bevitel történjen.

A két lehetőség csak a feldolgozás első két fázisában, a "nyers" adatok bevitelében és az adatalemek elhatárolásában különbözik egymástól. A hibajavítás, az adatok kiegészítése már azonos módon zajlik. Nagy állományok (1–2 millió bibliográfiai tétel) esetén az első megoldás gyakorlatilag kivihetetlen.

El kell dönteni azt is, hogy mágneslemezen (módosítható-kiegészíthető formában) vagy véglegesen rögzítve, pl. CD-ROM-on fogjuk-e az adatokat tárolni. A nagy és lezárt könyvtári katalógusok számára az utóbbi megoldásnak számos előnye van (pl. az egy tételre jutó alacsonyabb költség, a hordozhatóság, nincs online igény). Még ha egy-három évenként az új gyarapodás integrálása érdekében új CD-ROM-ot készít is a könyvtár és a legutóbbi időszak gyarapodását mágneslemezen tartja, integrált rendszerrel egyszerre tudja kezelni a teljes állományt. (Itt ennek a részleteire nem térünk ki.) A mágneslemezes tárolás a háttéradatok jobb elérhetőségét nyújtja a feldolgozó rendszer számára, de ennek határt szab a lemezkapacitás.

Hazánkban a legnagyobb könyvtáraknak milliós állományuk van, a közepesnek 100–500 ezer egységük. Ennél kisebb állományok optikai lemezre vitele jelenleg még nem tűnik célszerűnek, sőt a mágneslemezes tárolási technika bővülésével a kisebb könyvtárak igényei gond nélkül kielégíthetők.

A kis és közepes könyvtárak számára a WORM (Write Once Read Many) lemezes megoldás is kínálkozik. A távolabbi jövőben a törölhető-javítható optikai lemez megjelenése e kérdések újragondolását teheti szükségessé.

Más megvilágításba helyezi a kérdést, ha valamelyik nagy könyvtár az IIF-hálózatba integrálva osztott katalógus építését vállalja fel. Ez a kisebb könyvtárak számára a nagyobb gépi kapacitás hozzáférésének előnyével kecsegtet.

A katalóguscédulák gépi olvasását rendkívül megnehezíti a cédulák kopottsága, az eltérő írógépek használata, színbeli eltérése, az íráskép színe. Ráadásul a rajtuk levő információk egyes elemei nem állandó sorrendben, esetleg felcserélve helyezkednek el; az egyes adatelemeket nem mindig ugyanaz a jel határolja el; szükség lehet az aláhúzásnak mint információnak a feldolgozására is; a karakterkészlet sokkal bővebb (csak latin betűk esetén is) a bővített ASCII-nél. Ugyanakkor megkönnyíti a beolvasást a betűközök és a betűméret viszonylagos állandósága.

A tétel logikai szerkezetét a különböző címleírási szabványok és a belső könyvtári konvenciók határozzák meg.

A *Magyar Nemzeti Bibliográfia* nyomtatott köteteinek gépi olvasását ugyan a sorkizárásos, kéthasábos szedés nehezíti, a tiszta íráskép és a jó papír viszont enyhíti a leolvasás nehézségeit. Ez a hatalmas, jól olvasható adatállomány biztos alapot szolgáltathatna a zömében magyar anyagot gyűjtő könyvtárak gépi katalógusának előállításához.

Elképzelhető, hogy igény lenne gépi feldolgozásra mikroformákból is, noha ezeken a könyvtári katalógusoknak csak kis hányada található. Az MTA Könyvtára esetleg mégis ezt a megoldást választja a mikrofilmlap-katalógus mechanikus kezelésének egyszerűbb volta miatt, bár olvashatóságuk esetenként még rosszabb, mint a papírkártyáké.

Az outputtal szemben támasztott követelményeink:

- a nyert CD-ROM, WORM vagy mágneslemezes adatbázis leképezhető legyen az ismert szabványos formákban (sőt esetleg rendszerfüggetlen formában);
- tegye lehetővé a kibővített ASCII vagy vele ekvivalens, egybájtos, 256 jelen alapuló kód-, illetve karakterkészlet alkalmazását IBM PC XT/AT, azaz 16 bites feldolgozásra; vagy
- más, kibővített, nemzetközi szabványon alapuló karakterkészlet alkalmazását, s az itthon várhatóan installált könyvtári rendszerekhez legalább interfészes illeszthetőségét;
- legyen indexelhető legalább a szokásos eljárások szerint (az utalásokat is beleértve), de lehetőleg más hasznos szempont szerint is.

A vázolt fejlesztés elképzelése ellen szól

- nagy bonyolultsága,
- a szűk hazai felvevőpiac,
- a külpiac szabványigényei, a távvállalkozás nehézségei,
- a fejlesztési erőforrások hiánya.

A fejlesztés melletti érvek:

- a feladat aktualitása,
- a hardver- és szoftverfeltételek nagyjából érettek a sikeres megvalósításhoz (az OSZK és az

MTAK várhatóan olyan számítógéphez jut, amely a multitaszk üzemmód és a helyi hálózat révén alkalmas lesz a nagytömegű gépi adatbevitelre),

- a növekvő bibliográfiai adatbázis növeli az újabb feldolgozás megbízhatóságát és hatékonyságát,
- a CD-ROM-on és más adathordozón lévő katalógusok már jelenleg is rendelkezésre állnak,
- a hagyományos katalógusok CD-ROM-ra vitele itthon esetleg a Videoton közreműködésével megvalósítható, de a nyugati bérnyomtatásnak is jól kialakult útja van.

A fejlesztésben érdekelt lehet az OTKA-pályázat felhasználására alakult korlátolt felelősségű társaság (MTAK, OSZK és SZTAKI), továbbá a Videoton mint gyártó, valamint más potenciális felhasználók. A fejlesztés melléktermékeként több, a dokumentációban, a számítógépek irodai felhasználásában alkalmazható eredmény, termék is létrejöhet, ami a vállalkozás biztonságát növelheti.

Karakterfelismerő eljárások

A gépi képfeldolgozás, alakfelismerés és ezen belül a karakter (betű, szám) felismerése igen régóta foglalkoztatja a matematikusokat, mérnököket. Ilyen tárgyú elméleti tanulmányok már a hatvanas évek elején is születtek [1], és még ugyanebben az évtizedben az első kísérletek is lezajlottak [2].

A karakterfelismerés feladata elvben a következőképpen fogalmazható meg. Adva van a jelek egy X halmaza, amelynek elemei egyértelműen az X osztályaiba vannak besorolva. Egy-egy osztály elemei azonosíthatók egy karakterrel, illetve annak formaváltozataival, amelyeket azonosnak "olvasunk". Keresendő egy olyan algoritmus, amely

1. egy X-beli elemről eldönti, hogy melyik osztályba tartozik,
2. felismeri, ha nem X-beli elemmel találkozik.

Az algoritmus értékelése az alábbi hibás döntések valószínűsége alapján történik:

- X-beli elutasítása, mint ismeretlen (ún. elsőfajú hiba),
- nem X-beli X-be sorolása,
- X-beli hibás osztályba sorolása, azaz A-t lát, de B-t olvas (ez utóbbi két eset ún. másodfajú hiba).

Világos, hogy elsőfajú hiba esetén lehetőség van a hiba javítására, ami persze lehet időigényes, igényelhet emberi beavatkozást, de nem rontja az output hűségét. A másodfajú hiba viszont csak ellenőrzéssel tárható fel, automatikus kezelése nem lehetséges. Ezért az algoritmus optimalizálása elsősorban ennek a csökkentését és a lépésszám mérsékelt növekedését kell, hogy egyensúlyban tartsa. (Az irodalomban számos részben vagy elviekben más célfüggvényt is találunk.)

Az algoritmus, illetve a feladat kibővítésével három minőségileg igényesebb megközelítés is lehetséges.

- Az első egy *adaptív algoritmust* javasol: az algoritmus nagyszámú minta olvasása után maga alkotja meg az X halmazt és annak osztályait, illetve a teljes munkafolyamat során lehetséges az X, illetve az osztályok változtatása.
- A második eljárás lényege, hogy *interaktív feldolgozást* biztosít, azaz a felismerési algoritmus során (esetleg az adaptív eljárás egyes lépéseiben) emberi beavatkozást kér vagy tesz lehetővé.
- A harmadik a betűk teljes felismerése helyett *szavak felismerését* valósítja meg az alkotóelemek részleges ismeretében. Az eljárás tudatbázisra épít, amely nyelvészeti elveket és szótárat tartalmaz. Ez utóbbi automatikus vagy interaktív módon bővíthető, az előbbi pedig esetleg nyelvtől függően változtatható.

Ezeket az eljárásokat részben vagy egészben máris alkalmazzák működő rendszerekben, de együttes alkalmazásukra eddig csak egyetlen példát ismerünk. (Lásd a konkrét rendszereket ismertető fejezetet.)

Az alakfelismerési — így a *karakterfelismerési* — módszerek három nagy csoportba sorolhatók:

- összehasonlító,
- statisztikus,
- strukturális módszerek.

A gyakorlati algoritmusok ezek kombinációját is alkalmazzák, de valamelyik eljárás dominál.

Az *összehasonlító* módszer a nyers vagy előfeldolgozott karakter és a tárolt "etalonok" összehasonlításán alapszik. (Ilyen pl. a postai irányítószám olvasási eljárása.)

A *statisztikus* módszer két fázisból áll. A "tanítás" során a tömörített alakjellemzőket leíró kód készül valamennyi jelről. A felismerés során a jelek kódja közt definiált valamilyen távolságfüggvényt felhasználva lehet a jelet a legkedvezőbb osztályba sorolni. A tömörített alakjellemző kód létrehozására ma az ún. gyors Fourier-transzformációt használják a legszívesebben.

A *strukturális* módszer egyrészt felhasználja a jelkészlet a priori jellegzetességeit, másrészt az előfeldolgozás és az esetleges szegmentálás (azaz részalakzatokra bontás) után az (elemi) alakzatot topológiai jellegzetességeivel írja le. Ilyen a bejárási útvonal kódja, a relatív helyzet kódja stb. Ezek a kódok az eddigi eljárásokkal ellentétben nem egyszerűen jelsorozatokat reprezentálnak, hanem a jelkészlet jellegzetességének megfelelő nyelvtani szabályoknak is eleget tesznek. A felismerés során az alakza-

tot leíró mondat nyelvtani elemzése vezet el az osztályba soroláshoz, vagyis az azonosításhoz. Ez a megközelítés az alakleíró, illetve a formális nyelvek elméletére épít; ezek szabályait, szintaxisát többek közt *Chomsky* vizsgálta [3].

Napjainkban a gyakorlatban alkalmazott eljárások vagy az etalonnal való összehasonlítást valósítják meg a viszonylag nagy tárolt jelkészletre építve, vagy a strukturális módszeren alapulnak. Ezek esetenként ad hoc egyszerűsítő megoldásokat alkalmaznak és a végső osztályba sorolásnál valamilyen statisztikus eljárásra kerül sor.

Karakterolvasó berendezések

Az első valódi karakterolvasást végző berendezést 1912-ben *Goldberg* szabadalmaztatta: ez a készülék távírókóddá konvertálta az elolvasott jeleket. Az első ismert, valóban működő optikai karakterolvasó (1914) *Fourier d'Albe* nevéhez fűződik. Készülékét vakoknak tervezte, melynek segítségével a kézzel vezérelt olvasófej "érthető hangot" generált.

Az első munkára fogott készülék 1954-ben született az Intelligent Machine Corporationnél a "Reader's Digest" számára; ez írógéppel írt szöveget tudott olvasni. E cég készített 1959-ben az USA légiereje részére karakterolvasót, amely képes volt kis- és nagybetűs alfanumerikus jelkészletet elolvasni. A szakirodalom 1960-ban a Szovjetunióban számol be ilyen eszközökről. Ezek és a következő időszakban született berendezések két csoportba sorolhatók: fényképészeti maszkkal tárolt, illetve elektromos (ellenállás-mátrix stb.) etalonok. Ezek a berendezések már a hatvanas évek elején is képesek voltak 1000–2600 betű/s sebességű feldolgozásra. Később a technikai fejlődéshez képest lassan nőtt a karakterolvasók teljesítménye.

Az elmúlt harminc évben a feladatspecifikus berendezéseket lassan felváltották az általánosabb, sok célra felhasználható készülékek. Megszűnt a technikai megoldások széles skálája, kialakult egy viszonylag általános eljárás, amely a beolvasandó információ fizikai fogadásához analóg-digitális átalakítót alkalmaz és tőle elkülönítve a digitális jelek értelmezésére szolgáló számítógépet, illetve célberendezést. Az olvasóegységek legfontosabb fajtái:

- tv-kamera bemenete,
- CCD töltéscsatolt képbontó,
- lézeres letapogatással működő berendezés.

A jelentős változást a töltéscsatolt berendezések (CCD) megjelenése hozta a hetvenes évek elején. Ezek napjainkra a felbontás jelentős mérvű növelését tették lehetővé, így az input információ sokkal alaposabb elemzésére adnak módot. Csökkent az írásképpel és a papírminőséggel szemben támasztott követelmény, gyorsult és finomabbá vált az információfeldolgozás. Napjainkban szabványként emlegetik a 300 pont/inch (kb. 120 pont/cm) felbontást, de természetesen ennél jóval nagyobb felbontású készülékek is találhatók a piacon. (Egy A4-es oldalról 1 Mbájt információegységet olvasnak le a szokásos készülékek, tehát kb. egymillió betűnek megfelelő információ szükséges egy A4-es oldalon lévő kb. 1800 jel megbízható azonosításához.)

Az alábbiakban a teljesség igénye nélkül ismertetjük az 1980 után forgalomba került néhány berendezés nevét és gyártóját, valamint főbb paramétereit [4]. (Általában a többféle betűtípust feldolgozni képes készülékek esetében is szabványos jelkészletekről, pl. a leggyakoribb írógépek betűtípusairól – Courier 10, 12 vagy Prestige Elite 12, Pica stb. – van szó.)

Alphaword 3+ (1985 előtt), *Formscan* Compuscan. 120 jel/s, azaz 145 lap/óra, online javítás, 300 000 jelenként egy hiba, tízféle (szabványos) jelkészlettel bővíthető, de alapul egy jelkészlet szolgál. Ára: 22 850 font.

Series 80 A, B, C (1985 előtt), *Formscan*. 145 lap/óra, 1, max. 3 betűtípus. Ára: 11 500 font.

Dest Workless Station Model 211, 212, 213, 222, 223 (1982), *Lexisystem*. Képes sorkizárt szöveg olvasására, jobbra és balra zárt bekezdések értelmezésére, max. 12 betűtípus olvasására automatikus betűtípus-kiválasztással. Többféle betűtípust is olvas egy lapon belül. 145 lap/óra. Nincs saját billentyűzete. Ára: 8000–11 500 font.

Omni-Reader (1985. jan.), *Oberon International's*. 25 betű/perc, kézi mozgatású olvasófej, négy jelkészlet. Sajátos tagja az OCR-technikának, olcsó, egyszerű, mégis számos igényt kielégít. Ára: 500 USD.

Palantir Compound Document Processor (1986), *Palantir Corp.* Etalonos összeállítás, 0,25–2 oldal/perc a betűkészlettől függően. Ára: 40 000 USD.

OCR 6001, *Sepsi* (francia cég). 9 nyomtatott betűtípus, min. 7,5x12,5 cm, max. 21,5x35,5 cm. Lapméret: 250 lap/óra.

A továbbiakban bemutatott olvasókészülékek az ún. ICR- (**I**ntelligent **C**haracter **R**eader) családhoz tartoznak. Közös jellegzetességük a taníthatóság, illetve az automatikus adaptivitás, a nagy háttérinformáción alapuló intelligencia. Tulajdonképpen olvasásra, érzékelésre kialakított mesterséges intelligenciát képviselnek. Ennek a családnak az első és talán leghíresebb tagja a *Raymond Kurzweil* által kifejlesztett

Kurzweil 4000 (1983), *Kurzweil Computer Products*, Cambridge Mass (USA). Strukturális, topológiai eljárással képes érintkező karakterek szétválasztására, illetve hiányos vonalú karakterek felismerésére, szerkesztési forma visszaadására. Ára: 34 500 USD.

KDEM 1100, KDEM 1200, Compact Computer GmbH (NSZK). Tanítható, 250 000, illetve 160 000 jel/óra, 6500 DEM havi bérleti díj.

Meg kell különböztetni az olvasókészülékeket és a karakter felismeréséhez szükséges programokat. Egyre szélesebb körben terjednek el az ún. "desktop", azaz asztali olvasókészülékek, amelyek mini- vagy mikroszámítógéphez csatlakoztathatók aszinkron kimenettel. Bár a gyártók természetesen karakterfelismerő programot is kínálnak hozzájuk, lehetőség van más programok alkalmazására is a szabványos csatlakozással. Ilyen készüléket majd valamennyi hardvergyártó kínál az IBM-től kezdve az AGFA-ig. Töltéscsatolt kamerát a Híradástechnika Szövetkezet is gyárt, amely ugyan nem irodai alkalmazásra való, de felbontási és egyéb paraméterei alapján alkalmas karakterolvasásra.

Az 1987-es őszi BNV-n mutatták be az SZKI nemzetközi érdeklődést is kiváltó RECOGNITA karakterfelismerő programját. Az algoritmus strukturális módszert használ, és képes új karakterek megtanulására is. A rendszer irodai célokra készült, ezért a benne felhalmozott jelentős intelligencia ellenére tömegfeldolgozásra, gyengén olvasható input fogadására nem célszerű felhasználni.

A katalógusok visszamenőleges beviteléhez szükséges berendezések ma elérhető áron beszerezhetők, a feldolgozáshoz szükséges algoritmusok, illetve programok megvannak, megvásárolhatók, illetve kifejleszthetők.

A nyugati informatikai piacon megjelentek azok a vállalkozók, akik olvasóberendezés és számítógép birtokában széles skálán vállalnak feldolgozást, minden elképzelhető kapcsolódó szolgáltatással együtt.

Hagyományos katalógusból géppel olvasható katalógus

A hagyományos katalógusnak géppel olvasható katalógussá való átépítését nagyban segíti, ha valahol már hozzáférhető valamilyen, az átépítést tervező könyvtár állományát legalább részben lefedő gépi katalógus. Ennek online vagy offline felhasználása jelentősen megkönnyíti a feladatot. Igénybevételénél esetleg az okozhat nehézséget, hogy a fogadó fél házi szabványa eltér a felhasznált adatbázisétól. Ezek a nehézségek nem mindig hidalhatók át az átvételkor.

A tapasztalatok szerint a kézi bevitel igen sok hiba forrása, ami a gépi visszakeresést lehetetlenné teszi; ugyanakkor adatok bizonyítják, hogy a gépi olvasás lényegesen olcsóbb, mint az emberi munka [5].

A OPTIRAM/LIBPACK rendszer

Harrison cikkében [6] egy igen fejlett OCR-, illetve ICR-technológiára épülő, retrospektív katalógusfeldolgozó rendszert ismertet. Az OPTIRAM/LIBPACK rendszer hardverkonfigurációja a következő elemekből áll:

- CANON FAX 320E (Group 3),
- MC VESCA számítógép 1 Mbájt memóriával,
- 80 Mbájtos winchester-lemezes tároló,
- ellenőrző terminálok,
- Data Scope (a feldolgozott információ ellenőrzésére),
- PERTEC, 9 csatornás mágnesszalagegység (800 és 1600 bpi-s interfésszel).

A beolvasást egy közismert telefakszimile berendezés végzi, a nyers információt pedig egy nagy bonyolultságú intelligens program bontja egyedi karakterekké. Első lépésként a mágneslemezen található mintákkal hasonlítja össze a jeleket és ASCII kódot generál belőle. A továbbiakban egy, a karakterfelismerési eljárásokról szóló fejezetben ismertetettektől drasztikusan különböző módszert alkalmaznak.

A szóképolvasás

Az előfeldolgozás után minden egyes szót megvizsgál a rendszer, hogy szerepel-e a szótárában; ha igen, azonnal outputra kerül. Ha nem, akkor a nyers inputot behatóan újra elemzi. Ezután a szót végül is vagy megtalálja a szótárban, vagy elfogadja annak ellenére, hogy a szótár nem ismeri, mert a benne szereplő jelek pontosan felismerhetők, vagy pedig nem fogadja el, mert a jelek nem ismerhetők fel. Egy új szó akkor kerülhet be a szótárba, ha

- minden karaktere nehézség nélkül felismerhető, vagy
- a benne szereplő hangzók egymásutánja eleget tesz a nyelv ezekre vonatkozó általános szabályainak.

A szótárat a felhasználó kívánsága szerint bővítheti. A rendszer kiterjedt hibajavító eljárásokat tartalmaz. A nem elfogadott szavak is outputra kerülnek, de ezeket a rendszer kívánság szerint megjelöli.

Az *OPTIRAM leolvasó* képes kézzel írt alfanumerikus szöveg feldolgozására. A feladat bonyolult-

sága lassítja a feldolgozást, de amint az 1. és 2. ábra mutatja, az eredmény rendkívül jó. A kézírás felismerése vegyes, strukturális és etalonos módszeren alapszik. Nehéz esetekben a statisztikus döntési eljárás "legjobb tipp"-jét helyettesíti be a rendszer, amelyet azután a szótári ellenőrzés hagy helyben. Speciális eljárások gondoskodnak

- a megszakadt vonalak kitöltéséről;
- a papírhibák, foltok, összemosódások zavaró hatásának kiküszöböléséről;
- az áttetsző papír, a színes nyomás, a szakadások, eltűnt jelek okozta nehézségek leküzdéséről.

A rendszer nagy ereje tanulási képességében rejlik, amely a korábbi tapasztalatok felhasználásával növeli a felismerés sebességét, biztonságát.

A jelkészletre semmilyen korlátozás nincs [6], a rendszer minden latin betű, minden európai nyelv ékezetei és diakritikus jelei, minden szokásos speciális karakter, az alsó és felső index, a matematikai jelek, a görög jelek feldolgozására képes. Tervezik a cirill betűk és a héber jelek bevonását is.

A rendszer UKMARC és LCMARC feldolgozására és előállítására alkalmas, de szó van az ISO-karakterszabvány minden lényeges elemének átvételéről is. A nem szabványos jelek transzliterálását a felhasználó szabhatja meg.

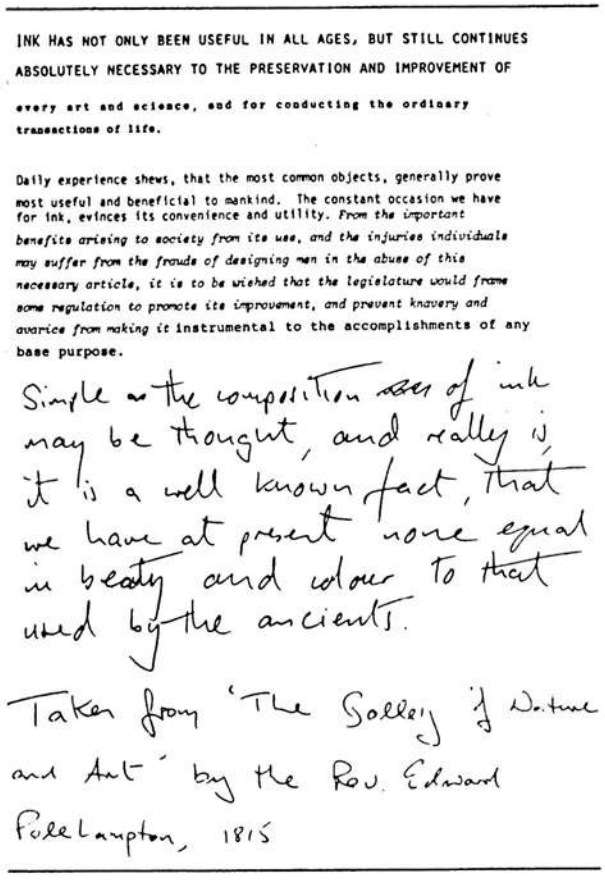
A *LIBPACK formaelemző* a szöveg kezelése mellett lehetővé teszi a kártyakép feldolgozását, pozicionálását, azaz az egyes adatelemek funkcionális elhatárolását, felismerését és kapcsolataik rögzítését is. Képes követni a katalóguskártya formaváltozatait akár különböző könyvtárak katalógusairól, akár egy könyvtár katalógusán belüli változatokról van szó. A felhasználó kívánsága szerint el lehet hagyni bizonyos adatelemeket, felcserélni mások sorrendjét a rekordban.

Harrison szerint a katalóguskonverziónál az alábbi tényezőket kell figyelembe venni:

- költségek,
- a programok irányítása,
- hozzáértő szakembergárda megléte, illetve alkalmazása,
- a minőség ellenőrzése,
- megfelelő szabványok.

1985-ben e rendszer segítségével a gépre való átvitel katalóguscédulánként a bonyolultságtól függően 30 penny körül mozgott, míg a hagyományos módon az USA-ban 1,55–3 dollár, Angliában 1–5 font volt az átlag.

Harrison példáit az 1–6. ábrákon közöljük. Ezek a rendszer hallatlan erejét bizonyítják. Az 1. ábra három katalóguscédula másolata, a 2. az ezekről


INK HAS NOT ONLY BEEN USEFUL IN ALL AGES, BUT STILL CONTINUES ABSOLUTELY NECESSARY TO THE PRESERVATION AND IMPROVEMENT OF

every art and science, and for conducting the ordinary transactions of life.

Daily experience shews, that the most common objects, generally prove most useful and beneficial to mankind. The constant occasion we have for ink, evinces its convenience and utility. *From the important benefits arising to society from its use, and the injuries individuals may suffer from the frauds of designing men in the abuse of this necessary article, it is to be wished that the legislature would frame some regulation to promote its improvement, and prevent knavery and avarice from making it* instrumental to the accomplishments of any base purpose.

Simple as the composition & use of ink may be thought, and really is it is a well known fact, that we have at present none equal in beauty and colour To that used by the ancients.

Taken from 'The Gallery of Nature and Art' by the Rev. Edward Polelampton, 1815

3. ábra

készült MARC rekord. A *3. ábrán* különböző betűtípusok és kézírás látható, a *4. ábrán* ezek gépi olvasata, a megfelelő tipográfiai adatokkal. Az *5. ábra* könyvbeli szöveg, a *6. ábra* ennek olvasata. Érdemes megfigyelni, hogy a szedett szöveg sorkizárásos, többféle címet és bekezdésmódot tartalmaz, amelyeket a rendszer szintén visszaad. Az output szövegben >.. < jelek közt található a tipográfiai utasítás. Ezek a következők lehetnek:

as: kisbetű
fc: betűtípusváltás
h1: cím
h2: cím (antikva, középre)
it: kurzív
10: bekezdés folytatása
11: bekezdés kezdete
gc: bekezdés középre zárva
ql: bekezdés balra zárva
ro: antikva
t1: szöveg kezdődik (cím után)
xs: kisbetű vége

Az elválasztott szavak jelölésére (ez katalóguscédulán nem fordulhat elő) a + jel használatos.

```
>h1<INKS>ql<
>l1<>t1<INK>as< HAS NOT ONLY BEEN USEFUL IN ALL AGES, BUT STILL CONTINUES
ABSOLUTELY NECESSARY TO THE PRESERVATION AND IMPROVEMENT OF>xs<>ql<
>l0<>fc<every art and science, and for conducting the ordinary transactions
of life.>ql<
>l1<>fc<Daily experience shews, that the most common objects, generally
prove most useful and beneficial to mankind. The constant occasion we have
for ink, evinces its convenience and utility. >it<>fc<From the important
benefits arising to society from its use, and the injuries individuals may
suffer from the frauds of designing men in the abuse of this necessary
article, it is to be wished that the legislature would frame some
regulation to promote its improvement, and prevent knavery and avarice from
making it >ro<>fc<instrumental to the accomplishments of any base
purpose.>ql<
>l1<>t1<Simple as the composition of ink may be thought, and really is, it
is a well known fact, that we have at present none equal in beauty and
colour to that used by the ancients.>ql<
>l1<Taken from 'The Gallery of Nature and Art' by the Rev. Edward
Polehampton, 1815>ql<
```

4. ábra

28

BLACK OR GARDEN NIGHTSHADE, *Solanum nigrum*, PLATE XIV.

has proved fatal in several cases, its effects being more powerful than those produced by the Bitter-sweet, though similar in nature. The plant is annual, rising to the height of one or two feet, and bearing small white flowers resembling those of the last mentioned in form. The berries, when mature, are black. It is a common weed in gardens, by waysides, and about manure heaps. Like the Woody Nightshade, it has sometimes been employed in medicine.

HENBANE, *Hyoscyamus niger*. PLATE XV.

An annual plant, not uncommon in some parts of the country in waste ground near towns and villages, generally in dry places or about rubbish heaps. It grows from a few inches to a foot or more in height, with downy stems and leaves, having a strongly fœtid odour. The flowers are greenish-yellow, generally, with dark veins forming a network over the corolla, but sometimes of a pale yellow and without veins; they are produced in July and August. The whole herb is poisonous to man, though apparently having little effect on domestic cattle and horses. It occasions delirium and stupor, accompanied by great dilatation of the pupil of the eye. The seeds have been used for smoking, as a remedy for toothache, but should never be employed, having caused convulsions, and even insanity, in some instances. The leaves are the most powerful portion of the plant; even the odour of these, when fresh, will produce giddiness and stupor. Two fatal cases at least of poisoning by this plant are re-

5. ábra

```
)h2(B)as(LACK OR)xs( G)as(ARDEN)xs( N)as(IGHTSHADE,)xs( )it(Solanum
nigrum,)ro( P)as(LATE)xs( XIV.)qc(
)l1()t1(has proved fatal in several cases, its effects being more powerful
than those produced by the Bitter-sweet, though similar in nature. The
plant is annual, rising to the height of one or two feet, and bearing small
white flowers resembling those of the last mentioned in form. The berries,
when mature, are black. It is a common weed in gardens, by waysides, and
about manure heaps. Like the Woody Nightshade, it has sometimes been
employed in medicine.)ql(
)h2(H)as(ENBANE,)xs( )it(Hyoscyamus niger.)ro( P)as(LATE)xs( XV.)qc(
)l1()t1()M(An annual plant, not uncommon in some parts of the country in
waste ground near towns and villages, generally in dry places or about
rubbish heaps. It grows from a few inches to a foot or more in height, with
downy stems and leaves, having a strongly f)s1(tid odour. The flowers are
greenish-yellow, generally, with dark veins forming a network over the
corolla, but sometimes of a pale yellow and without veins; they are
produced in July and August. The whole herb is poisonous to man, though
apparently having little effect on domestic cattle and horses. It occasions
delirium and stupor, accompanied by great dilatation of the pupil of the
eye. The seeds have been used for smoking, as a remedy for toothache, but
should never be employed, having caused convulsions, and even insanity, in
some instances. The leaves are the most powerful portion of the plant; even
the odour of these, when fresh, will produce giddiness and stupor. Two
fatal cases at least of poisoning by this plant are re+)qj(
```

6. ábra

A katalógusok visszamenőleges konverziója CD-ROM segítségével

Az első CD-ROM kiadványok közt jelent meg a *Library of Congress* (LC) angol, idegen nyelvű és Any-book (szerzeményezési) katalógusa a Library Systems and Services és ettől függetlenül a Library Corporation gondozásában. Már a kettős kiadás is nagy piaci sikert sejtet. Azóta további CD-ROM kiadók is megjelentek az LC katalógusának valamilyen formájával. A CD-ROM könyvtári alkalmazása eddig két területen kezd komoly méreteket ölteni [7]. Az egyik a katalógusok retrospektív konverziója, a másik a szerzeményezés. Az utóbbi a kiadói listákat tartalmazó CD-ROM-okra, az előbbi pedig valamely könyvtár katalógusát tartalmazó CD-ROM-okra támaszkodik. Mindkét területen számos intelligens szoftvermegoldás könnyíti a könyvtárak munkáját, egyes CD-ROM-ok online hozzáféréssel, illetve online rendelési lehetőséggel vannak kiegészítve. (E helyütt nem térünk ki a szerzeményezést segítő CD-ROM termékekre.)

A visszamenőleges katalógusépítést CD-ROM szolgáltatásra alapozó vevő bízvást számíthat arra, hogy a kezelőprogram segítségével a házi szabvány szerinti formában kapja meg a bibliográfiai rekordokat, s kényelmes szövegszerkesztő program segíti a rekordban esetleg szükséges módosítások végrehajtását, a CD-ROM-on nem található tételek kézi bevitelét. Az alkalmazott rendszerek könnyen áttekinthető, egyszerű, funkciói között menürendszer segítségével navigálhatunk. A bibliográfiai tételek sok közvetlen elérési ponton keresztül visszakereshetők az adatbázisban. A keresést logikai operátorok és csonkítási lehetőség könnyíti. A rendszer segítségével katalóguscédula és címke nyomtatható. Felismerve a CD-ROM állandósága és a katalógus bővülése közti ellentmondást, egyes kezelőprogramok úgy vannak kialakítva, hogy a winchesteren és a CD-ROM-on tárolt rekordok közt úgy keres, mintha egyetlen adatbázist alkotnának. 1987-ben a következő rendszerek álltak rendelkezésre:

Bibliofile, Library Corporation. Az első ilyen típusú termék, amelyet napjainkig a legtöbb, ezernél több példányban adtak el. Az adatbázis 2,2 millió

angol nyelvű rekordot tartalmaz négy lemezen, egy továbbin pedig egymillió idegen nyelvű rekordot. A bővített szolgáltatású rendszer a következőket kínálja: max. nyolc munkahely helyi hálózatba kapcsolása és egy közös adatbázishoz való kapcsolódás; vonalkódolvasó; 340 Mbájtnyi (láncba köthető) winchester-lemez a helyi adatbázis tárolására. A rendszer egyetlen gyöngéje, hogy nagykönyvtárak számára nem igazán magas a találatok száma, azaz a fogadó könyvtár katalógusa és a CD-ROM állomány közös tételeinek száma nem teszi ki az állomány túlnyomó többségét. Ezért szükséges más gépi katalógusforrások bevonása is.

LaserQuest, General Research Corporation. 1986 júliusában jelent meg, jelenleg 100-nál több installált példánya van. Az állomány 4,5 millió rekordot (2,2 LC, 2,3 kiegészítés) tartalmaz három lemezen; a negyedik lemez 260 ezer kanadai könyv CANMARK adatait tartalmazza.

Spectrum 400, 800, 1000, Library Systems & Services Inc. A Laserfile 2,2 millió rekordjára épül, s várhatóan a NICEM adatbázis adataival is bővül. Igen drága rendszer, viszont a személyi számítógépektől kiépíthető egészen a sok munkahelyes MICROVAX rendszerig.

CD-CATALOG-OCLC. Az OCLC 1981-től felvett 1,4 millió tételét, továbbá az LC 1,8 millió tételét tartalmazza. Az OCLC tagkönyvtárai számára készül.

DISCON-UTLAS:DISCON. Az UTLAS tagjai számára készült katalogizálási segédeszköz. Az összes 1984 előtti LC rekordot és a REMARC adatbázist tartalmazza. A négy lemez és a terminál havi bérleti díja 800 USD. Az automatikus feldolgozást sokkal megbízhatóbbá teheti e háttérkatalógusok alkalmazása. Ezzel az olvasást a szótári összehasonlításhoz hasonló támogatással lehet ellátni, azaz e célfeladat esetén betű-, szó- és kártyaképolvasásról együtt lehet beszélni. A feldolgozás megfelelő fázisában néhány azonosított szó (név) alapján kereshetjük a háttérkatalógus azonos tételét; megtalálása, ill. pontos azonosítása után a további olvasást mellőzve átemelhető a tétel készülő katalógusunkba.

Irodalom

[1] FANSTEIN, A.: Osnovi teorij informacij. Moskva, 1960.

[2] KOVALESKIJ, V. A.–SLEZINGER, M. Y.–RIBAK, V. I.: Čitaûšie avtomati i raspoznavanie obrazov. Naukava Dumka, Kiev, 1965.

[3] HOPCROFT, J. E.–ULLMAN, J. D.: Formal languages and their relation to automata. Massachusetts, 1969.

[4] SCHANTZ, H. F.: OCR data entry systems: current and future uses. = The Office, 91. köt. 2. sz. 1980. p. 58, 60, 63, 65.

[5] MORTEN, H.: Optical scanning for retrospective conversion of information. = The Electronic Library, 4. köt. 6. sz. 1986. p. 328–331.

[6] HARRISON, M.: Retrospective conversion of card catalogues into full MARC format using sophisticated computer-controlled visual imaging techniques. = Program, 19. köt. 3. sz. 1985. p. 213–230.

[7] JONES, R.: Compact disc technology; Developments and standards. = Data Processing, 28. köt. 6. sz. 1986. p. 295–298.

TELCS András: Optikai karakterolvasó rendszerek könyvtári alkalmazása

Az optikai karakterolvasó berendezések és algoritmusok fejlődése aktuálissá teszi a könyvtári adatbevitel automatizálását. A tanulmány ennek a feladatnak a főbb feltételeit igyekszik áttekinteni, bemutatva a nemzetközi szakirodalomban található eredményeket.

* * *

ТЕЛЧ, А.: Применение оптических читающих устройств в библиотеках

Введение и развитие оптических читающих устройств и алгоритмов создали актуальность автоматизации ввода данных в библиотеках. Автор пытается рассмотреть основные условия этого задания и продемонстрировать результаты, изложенные в научно-технической литературе.

TELCS, A.: Library applications of Optical Character Recognition (OCR) technology

The development results of optical scanners and character recognition algorithms made the automation of library data recording by OCR systems feasible. The state-of-the-art and main trends of the application of OCR technology in libraries are overviewed, based on the international literature.

* * *

TELCS, A.: Die Anwendung der optischen Zeichenleser-Systeme in den Bibliotheken

Die Entwicklung der optischen Zeichenleser-Systeme und Algorithmen macht die Automatisierung der Dateneingabe in den Bibliotheken aktuell. Die Studie bestrebt die wichtigsten Bedingungen dieser Aufgabe zu überblicken, in dem sie die Ergebnisse in der internationalen Fachliteratur darstellt.