UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

SMOTRNOST UPORABE TEHNOLOGIJE PREPOZNAVANJA OBRAZCEV V DOKUMENTNIH SISTEMIH

MITJA NIKOLIĆ

IZJAVA:

Študent MITJA NIKOLIĆ izjavljam, da sem avtor tega magistrskega dela, ki sem ga napisal pod mentorstvom dr. MOJCE INDIHAR ŠTEMBERGER in skladno s 1. odstavkom 21. člena Zakona o avtorskih in sorodnih pravicah dovolim objavo magistrskega dela na fakultetnih spletnih straneh.

V Ljubljani, dne _____

Podpis: _____

KAZALO

KAZALO SLIK

1. UVOD

1.1 Opredelitev problema, namen in cilj dela

Vprašanje: Kdaj slika NI vredna tisoč besed?
Odgovor: Ko je SLIKA tisočih besed.

Zgornje vprašanje in odgovor v sebi skriva bistvo tehnologij optičnega prepoznavanja znakov. V informacijskem svetu, v katerem živimo, ima informacija na papirju ali na poskeniranem dokumentu le malo veljave, če je ne pretvorimo v tekstovno (ASCII), infomacijskim sistemom razumljivo obliko in omogočimo iskanje. Podatke lahko pretvorimo ročno ali pa uporabimo tehnologije računalniškega upodabljanja (skeniranja) in optičnega prepoznavanja znakov.

Izpolnjevanje obrazcev je del našega vsakdanjika in ročno vnašanje podatkov v podatkovno bazo, ki so rezultat neke ankete ali popisa, skratka izpolnjenega obrazca, je lahko časovno zelo potratno opravilo. Tehnologija avtomatičnega prepoznavanja in procesiranja obrazcev omogoča hitrejši, lažji in v veliko primerih tudi cenejši proces za podjetje, predvsem v primerih, ko se soočamo z velikim številom enakih obrazcev. Na ta način lahko v povprečju prihranimo 70 do 85 odstotkov dela (Spencer, 1996). Zaradi pretežno ročno popisanih obrazcev, v tovrstnih sistemih prevladuje ICR (*angl. Intelligent Character Recognition)* tehnologija. Procesiranje obrazcev se najbolj uporablja v bančništvu (npr. procesiranje čekov), zavarovalništvu (procesiranje zavarovalnih polic), zdravstvu in državnih ustanovah. Kot primer uporabe naj navedem tudi procesiranje vhodnih faks sporočil s pomočjo faks strežnikov, kar omogoča avtomatično elektronsko razporejanje prispelih faksov.

Iskanje algoritmov in metod, ki bi omogočile, da bi računalniški sistemi lahko “videli” in prepoznavali objekte tako kot človek (prepoznavanje oblike predmetov, gibanja, zvoka oziroma govora, pisave, ipd.), je bil izziv mnogim raziskovalcem in znanstvenikom v različnih disciplinah. Optično prepoznavanje znakov (OCR, *angl. Optical Character Recognition*) je bilo in je eno najbolj fascinantnih področij tovrstnih raziskav. Pravi razcvet so doživeli sistemi prepoznavanja predvsem zaradi napredka tehnologije nevronskih mrež in ker so postali dostopnejši uporabnikom osebnih računalnikov za relativno nizko ceno (McCarthy, 1994).

Še tako izpopolnjen algoritem prepoznavanja znakov pa ima napake. 100% natančnost je pač utopično pričakovati. Postopki za izboljšanje prepoznavanja so: čiščenje motenj pred prepoznavanjem (*angl. despeckle*), uporaba slovarja za

prepoznavanje besed, izločevanje nemogočega zaporedja znakov, ipd. Zato je potrebno sistem optičnega prepoznavanja znakov analizirati predvsem iz treh izhodišč: natančnosti, hitrosti in cene. **Uporaba je torej smiselna, v kolikor seštevek stroškov in časa porabljenega za kasnejše popravljanje prepoznanega besedila, ne presega časa in stroškov za ročen vnos besedila, oziroma podatkov**. Ob tem je potrebno upoštevati tudi stroške uvajanja ter morebitne posredne koristi na ostale delovne procese. Pravtako so v večini primerov pomembni tudi originalni dokumenti oziroma obrazci ter njih arhiviranje in dostop, zato je tehnologija prepoznavanja obrazcev del sistemov za elektronsko upravljanje dokumentov.

Avtomatično procesiranje obrazcev s pomočjo OCR in ICR sistemov omogoča še dodatne prihranke. Zaradi standardnih in enakih oblik obrazcev postane shranjevanje in arhiviranje poskeniranih izpolnjenih obrazcev nepotrebno, saj pride do redundance. Poleg podatkov iz obrazca naj se v bazo zapiše še unikatni podatek obrazca. Ob zahtevi pregleda obrazca s strani uporabnika naj se torej ne prikaže slikovna datoteka izpolnjenega obrazca, temveč le primerni vzorec (glede na unikatni podatek obrazca) s “prilepljenimi” podatki iz baze na pravilnih lokacijah v vzorcu. Tako lahko dosežemo velike performančne in prostorske prihranke. Poleg tega lahko podatke iz obrazca izkoristimo tudi pri pogojno definiranih delovnih postopkih. Glede na definirane podatke se po obdelavi dokumenti pošiljajo v različne poti kroženja dokumentov (*angl. dynamic workflow*), ki so del krmiljenja delovnih procesov.

Seveda je pri procesiranju obrazcev omembna natančnost sistema in v povezavi s tem tudi avtomatična in človeška verifikacija rezultatov. Podatke iz obrazcev je potrebno pregledati še preden so vpisani v sistem za upravljanje baz podatkov (v nadaljevanju SUBP). Velika večina sistemov omogoča verifikacijo podatka ob procesiranju tako, da na zaslonu prikaže tako prepoznani podatek, kot tudi del skeniranega obrazca, kjer se podatek nahaja. V večini primerov se verifikacija izvede pred izvozom v SUBP, saj je tam oblika in tip podatkov striktnejša kot v dokumentnih sistemih. Metod za avtomatično odkrivanje napak je veliko, temeljijo predvsem na osnovi vnaprej definiranih pravil za posamezna polja. Naj jih nekaj naštejemo:

- datumsko polje ne more imeti podatka 30. Februar,
- primerjava z obstoječimi podatki v ciljni bazi (t.i. look-up tabele) kot je npr. enakost poštne številke in kraja,
- preverjanje povezanih podatkov kot je npr. seštevek zneska, davka in napitnine na slipih bančnih kartic, itd.

Namen naloge je ugotoviti smotrnost tehnologije prepoznavanja obrazcev v dokumentnih sistemih. Opisane bodo osnove dokumentnih sistemov, optičnega prepoznavanja znakov, posebnosti prepoznavanja obrazcev in lastnosti sistema za prepoznavanje obrazcev IMiS/ICR. Predstavljeni bodo praktični primeri uporabe tehnologije prepoznavanja obrazcev ter idejni projekt avtomatskega prepoznavanja obrazcev za prijavo DDV (Davek na dodano vrednost) v okviru Davčne uprave Republike Slovenije (v nadaljevanju DURS). Le-ta je v letu 2000 je objavila javni razpis za idejni projekt avtomatskega zajema in prepoznavanja obrazcev za prijavo DDV. Podjetje SRC.SI d.o.o. se je prijavilo na razpis in bilo tudi izbrano, sam sem bil tehnični vodja projekta s strani izvajalca.

1.2 Metode preučevanja in zasnova dela

Namen dela je ugotoviti smotrnost uporabe tehnologije prepoznavanja obrazcev v dokumentnih sistemih. Tehnologijo bom analiziral predvsem iz treh izhodišč: natančnosti, hitrosti in cene. Uporaba je torej smiselna, v kolikor mi bo na praktičnih primerih uspelo pokazati, da seštevek stroškov in časa porabljenega za kasnejše popravljanje prepoznanih podatkov ne presega časa in stroškov za klasičen (ročen) vnos podatkov iz obrazcev. Pri tem je potrebno upoštevati višino naložbe v tehnologijo in uvajanje uporabnikov ter morebitne dodatne posredne koristi tovrstne tehnologije, saj velikokrat omogoča prenovo delovnih procesov, reorganizacijo, povečanje produktivnosti, ipd. Seveda pa smotrnost ne moremo opredeliti splošno za vse procese, temveč je potrebno posamezne primere oziroma procese, ki vsebujejo obrazce, ločeno analizirati na dovolj velikem vzorcu.

Metode dela, ki jih bom pri izdelavi magistrskega dela uporabil, temeljijo predvsem na teoretičnem proučevanju tehnologije optičnega prepoznavanja znakov, iskanju uspešnih in neuspešnih primerov uporabe v svetu, ter analizi smotrnosti uporabe tehnologije na primeru obrazcev za prijavo DDV v Sloveniji. Ob tehnološki analizi smotrnosti uporabe v smislu natančnosti prepoznavanja množice realnih vzorcev DDV obrazcev, bom podal tudi stroškovno analizo vpeljave tovrstnega sistema na DURS. Pri teoretični analizi tehnologije in njenem razvoju se bom naslonil na strokovno literaturo pretežno tujih avtorjev, saj v Sloveniji ni veliko napisanega o optičnem prepoznavanju znakov, predvsem ne o t.i. neindustrijskem prepoznavanju v dokumentnih informacijskih sistemih. Pri analizi praktičnih primerov uporabe sem se naslonil na lastno znanje in izkušnje pri implementaciji in integraciji tovrstnih tehnologij v informacijske sisteme podjetij.

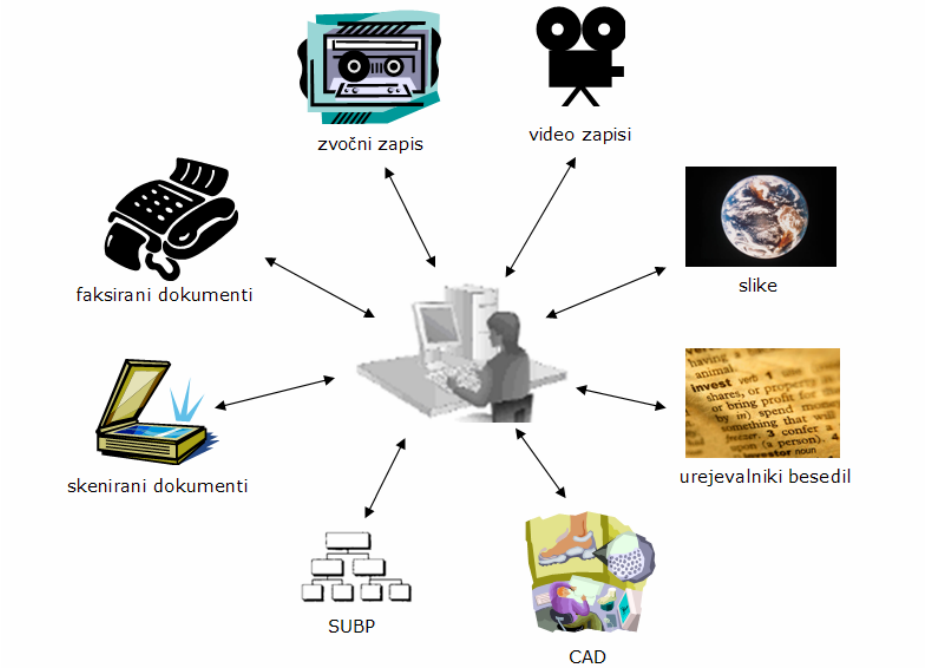
2. DOKUMENTNI SISTEMI

Sistemi za elektronsko upravljanje dokumentov ali dokumentni sistemi (EDMS – *Electronic Document Management Systems*) so danes ključni del pri prenovi poslovnih informacijskih sistemov. Hiter razvoj tako strojne kot programske opreme je omogočil ekspanzijo dokumentnih sistemov. Omogočajo zbiranje, sestavljanje, evidentiranje, elektronsko distribucijo, indeksiranje za potrebe iskanja, arhiviranje in hitro iskanje informacij ter s tem znanja. Osnova sistema je dokument v elektronski in ne v papirnati obliki.

2.1 Vloga dokumentnih sistemov

Razvoj koncepta informacijskih sistemov lahko zajamemo v treh obdobjih. Prvo je znano kot Von Neumann-ov model procesiranja. Gre za linearni proces vnosa podatkov, njihovo obdelave in izhodnih podatkov, kot rezultata določenega procesa. Nato je prišlo obdobje relacijskih podatkovnih baz in velikih gostiteljskih (*angl. host*) sistemov. Gre za tradicionalne sisteme za upravljanje baz podatkov. Danes izginja koncept centralno shranjenih in zgolj relacijsko povezanih podatkov, saj uporabnika ne zanima, od kod prihaja informacija. Pomembno je, da je na voljo, ne glede na obliko (glej sliko 1): podatkovna baza, faksirani dokument, skenirani dokument, video zapis, datoteka urejevalnika besedil ali preglednic, slika, zvočni zapis, itd (Mantelmann, 1995).

Slika 1: Princip dostopa na enem mestu



Uporabnik naj bi imel dostop do dokumentov na enem mestu, ne glede na izvor in obliko dokumenta. Tak koncept imenujemo koncept izgubljene škatle (*angl. the missing box*) in je osnova sodobnih dokumentnih sistemov. Dostop do informacije se poveča, struktura le-te pa upade. Znebiti se je torej potrebno razmišljanja o podatku kot polju v neki podatkovni bazi. Na podatke je potrebno gledati prek nove osnove – dokumenta. SUBP služijo kot spodnji nivo dokumentnega sistema, kot zbirka referenčnih podatkov za opis dokumenta, torej kot zbirka ključnih podatkov za iskanje dokumenta. Usmerjenost k sistemom, kjer je v središču uporabnik in kjer je osnova dokument, je pripeljala do nastanka in razvoja tehnologij kot so krmiljenje poslovnih procesov (*angl. workflow*) in podpora skupinskem delu (*angl. groupware, collaboration*). Pravtako je bistvena razlika, da se informacije filtrira ob poizvedbi in ne ob nastanku, kot je to pri sistemih s strukturirano informacijo. (Koulopoulos, 1995)

Če zadevo posplošimo, pridemo do spoznanja, da nam je na voljo velika ali celo preveč podatkov in informacij. Pridemo torej do problema, ko informacije so na voljo, potrebno jih je le primerno izluščiti in uporabiti. Iskanje informacij, oziroma destilacijo podatkov, opišemo preko treh različnih pristopov (Koulopoulos, 1995):

- ***iskanje po ključnih besedah in podatkih:*** Ključne besede se vnesejo ob nastanku oziroma vstopu dokumenta v sistem in se nahajajo v tradicionalnih podatkovnih bazah. Pod ključnimi podatki pa pojmujemo avtorja, datum kreiranja in morda še nekaj podobnih atributov v odvisnosti od profila dokumenta. Tu igra pomembno vlogo tehnologija prepoznavanja obrazcev.
- ***iskanje po polnem besedilu (angl. full text search)*:** Vsebina dokumenta se nahaja v vzporedni podatkovni strukturi in služi kot indeks originalnemu dokumentu. Tu pridejo do izraza sistemi optičnega prepoznavanja znakov znotraj sistemov za elektronsko upodabljanje oziroma skeniranje dokumentov.
- ***princip združevanja (angl. clustering)*:** Dokumenti se združujejo v skupine "podobnih." Poizvedbe vrnejo množico dokumentov, ki ustrezajo iskalnim pogojem, kar zahteva primerno organizacijo dokumentov.

Izbira primernega orodja za iskanje informacij v sistemu je ključnega pomena, saj če informacij ne moremo poiskati, potem je morda bolje, da je sploh ni.

2.2 Osnovni gradniki dokumentnih sistemov

Osnovne gradnike dokumentnega sistema najbolje ponazarja slika 2. Koncept dokumentnega sistema je sestavljen iz naslednjih komponent (Ponikvar, 2003):

- podsistema za upravljanje dokumentov,
- podsistema za krmiljenje delovnih procesov (*angl. workflow*)

- podsistema za računalniško uporabljanje oziroma skeniranje dokumentov (*angl. imaging*),
- podsistema za upravljanje zapisov,
- logičnega arhiva,
- fizičnega arhiva.

Slika 2: Koncept dokumentnega sistema

Vir: Ponikvar, 2003

Komponente medsebojno povezujemo s funkcijami, ki jih nudi sporočilni sistem, del sistema v smislu izmenjave podatkov je tudi transakcijski sistem. Seveda v posameznih primerih implementacije niso zajeti vsi gradniki. Predvsem upravljanje zapisov in povezava na transakcijske sisteme niso pogosti, brez sistema za računalniško upodabljanje dokumentov si pa denimo dokumentnega sistema z vhodnimi dokumenti v papirni obliki ni mogoče zamisliti, saj bi v tem primeru imeli le nek evidenčni sitem, torej hranjenje informacije o obstoju dokumenta, ne pa tudi o njegovi vsebini. Na tak način izgubimo osnovni namen iskanja informacije ne glede na izvor.

2.2.1 Podsistem za upravljanje dokumentov

V okviru podsistema za upravljanje dokumentov zagotavljamo naslednjo funkcionalnost:

- (logično) enoten vhod in izhod za dokumente (enotni postopki evidentiranja za vse vhodne in za vse izhodne dokumente),

- indeksiranje dokumentov,
- povezovanje dokumentov,
- pripravo izhodnih dokumentov (predloge, rezervacijo[1], vračanje, verzije),
- iskanje dokumentov.

V okviru tega podsistema zagotavljamo tudi varnost (avtentikacija in avtorizacija) ter podporo postopkom, povezanih z morebitnim digitalnim podpisovanjem dokumentov.

2.2.2 Podsistem za računalniško upodabljanje dokumentov

V okviru podsistema za računalniško upodabljanje (skeniranje) dokumentov (*angl. Imaging*) izvajamo zajem dokumentov v elektronski obliki ter tudi zajem elektronskih dokumentov (npr. faks). Več o sistemih za skeniranje dokumentov je opisano v posebnem podpoglavju 2.4.

2.2.3 Podsistem za upravljanje zapisov

Pri množičnih izpisih (računi, obračuni, itd) lahko vsebino, ki jo pošiljamo na tiskalnik, obenem pošljemo tudi v sistem za upravljanje zapisov (*COLD – angl. Computer Output to Laser Disc, novejši termin je ERM – angl. Enterprise Records management*). V COLD sistemih shranjujemo ASCII zapis dokumenta, indeksne podatke in podatke o obrazcu za izpis. Dokumenta ne hranimo v celoti, temveč ob zahtevi dokument sestavimo in prikažemo na zaslonu. COLD torej ni namenjen arhiviranju.

2.2.4 Podsistem za krmiljenje delovnih procesov

V okviru sistema za krmiljenje delovnih procesov lahko definiramo, pripravimo in izvajamo krmiljenje delovnih procesov ter nadziramo njih izvajanje. Več je opisano v poglavju 2.5.

---

[1] Rezervacija in vračanje dokumentov (angl. check-out in check-in) je pomembna lastnost dokumentnih sistemov, ki omogoča kontrolo nad konflikti spreminjanja in shranjevanja dokumentov. V primeru rezervacije se dokument v sistemu zaklene in je na voljo le za branje. Spreminjamo ga lahko šele, ko ga uporabnik, ki ga je rezerviral, vrne v sistem.

2.2.5 Logični arhiv in fizični arhiv

V logičnem arhivu hranimo evidenčne podatke o dokumentih, fizično se pa posamezni objekti (datoteke) nahajajo v fizičnem arhivu. V logičnem delu se nahajajo le kazalci na objekt v fizičnem arhivu. Fizični arhiv skrbi tudi za migracijo starejših objektov ter da so dokumenti varni pred nepooblaščenim dostopom. Fizični arhiv torej ni le arhiv dokumentov, temveč tudi aplikativna rešitev, ki zagotavlja varnost in sledljivost dokumentov. Več o arhiviranju dokumentov je opisano v poglavju 2.3.

2.2.6 Sporočilni sistem in podpora sodelovanju

S pomočjo sporočilnega sistema znotraj dokumentnih sistemov uporabnikom omogočamo pomembne funkcionalnosti, kot npr.:

- obveščanje o nastanku ali spremembi dokumenta v sistemu,
- obveščanje o spremembi faze v delovnem procesu oziroma delovnem toku in morebitni potrebni akciji,
- nastavitev alarmov za posamezna opravila,
- kreiranje opravil v elektronskih koledarjih,
- enostavna izmenjava sporočil, kjer je še posebej potrebno poudariti uporabnost tehnologij sodelovanja v realnem času (*angl. IM - Instant messaging*). Tu ne gre več le za preproste pogovore, temveč za multimedijsko sodelovanje v okviru virtualnih sestankov, video konferenc, prenosa tako zvočnega kot video signala ter delitvijo in delom na dokumentih v realnem času.

2.2.7 Povezovanje s transakcijskim sistemom

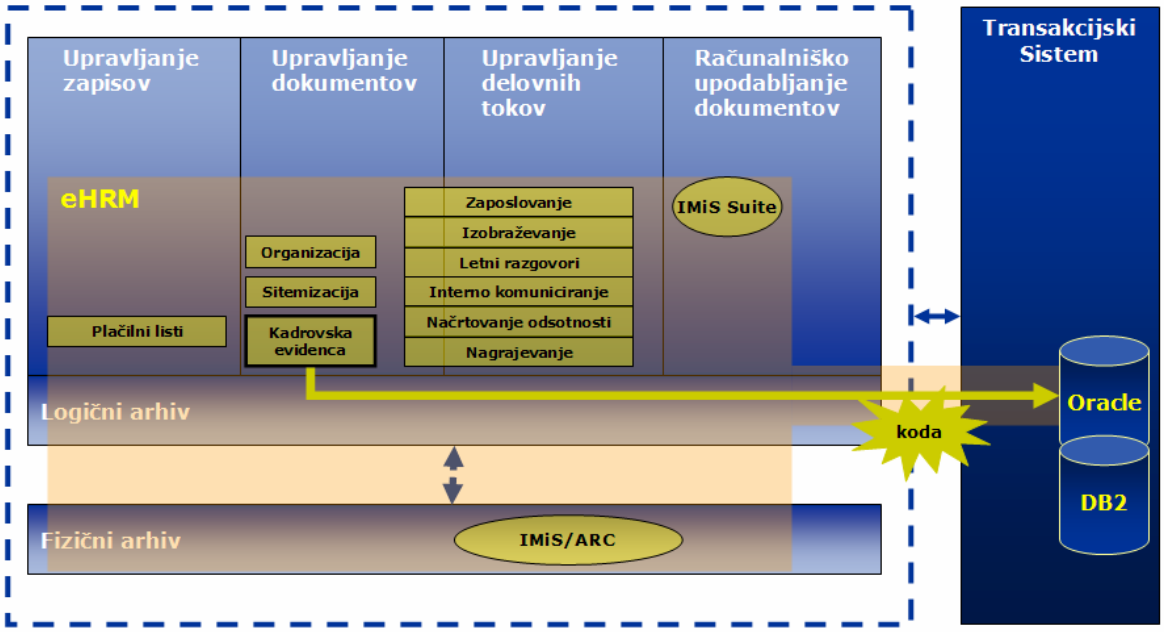
Vseh podatkov ne moremo hraniti v dokumentnem sistemu, saj je namenjen zgolj delu z dokumenti. Zato, da se izognemo dvojnemu vnosu ter dvojnim zapisom podatkov mora biti dokumentni sistem povezan s transakcijskim sistemom, pri čemer je povezava lahko zgolj enostavna (dostop do šifrantov) ali pa kompleksnejša (pisanje novih zapisov v transakcijske sisteme, spremembe podatkov v enem ali drugem sistemu, vpliv na delovni tok v dokumentnem sistemu, itd). Pomembno je, da ločimo namembnost sistemov. Znotraj transakcijskega sistema ne moremo reševati problema dokumentov, dokumentni sistem pa ni namenjen informatizaciji kompleksnih podatkovnih relacij in transakcij.

2.2.8 Primeri

2.2.8.1 eHRM podjetja SRC.SI

Kot prvi primer naj navedem arhitekturo aplikacije eHRM (HRM - Human Resource Management) za upravljanje s človeškimi viri podjetja SRC.SI (glej sliko 3). Podsistem upravljanja z zapisi vsebuje zajem izpisa plačilnih listov, znotraj sistema za upravljanje z dokumenti je zajeta organizacija in sistemizacija kadrov ter kadrovska evidenca podjetja, v sklopu upravljanja delovnih tokov so definirani procesi zaposlovanja, izobraževanja, letnih razgovorov, nagrajevanja, internega komuniciranja ter načrtovanja odsotnosti. Za računalniško upodabljanje dokumentov se uporablja IMiS družina produktov, platforma logičnega arhiva je IBM Lotus Domino in fizični arhiv predstavlja IMiS/ARC kot arhivski strežnik iz družine produktov IMiS. Logični arhiv izmenjuje podatke s podatki v transakcijskem sistemu. V tem primeru imamo zajete vse podsisteme EMDS sistema.

Slika 3: eHRM podjetja SRC.SI



Vir: Ponikvar, 2003

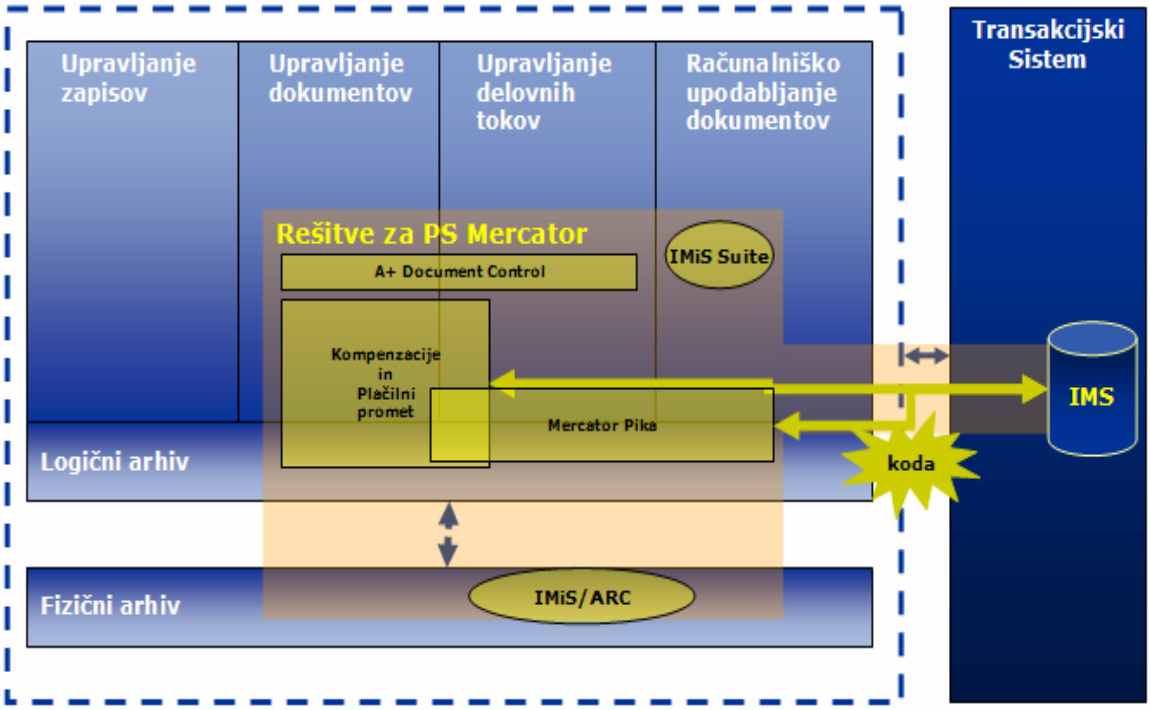
2.2.8.2 Mercator d.d.

Kot naslednji primer naj predstavim konkretno zasnovo dela dokumentnega sistema v podjetju Mercator d.d. (glej sliko 4). Vpeljava dokumentnega sistema se je začela ob težavah s papirji v sistemu kompenzacij. Tako se je zasnovala arhitektura prenosa predlogov in odobritev medsebojnih in verižnih kompenzacij

prek strežnika za pošiljanje faks sporočil. Delo s papirjem se je tako zmanjšalo na minimum, saj po prenovi procesa oddelek kompenzacij dobi dokument v elektronski obliki prek faks strežnika, ki le-tega avtomatično ob sprejemu arhivira na fizični in logični arhiv, pri čemer se za usmerjanje uporabi podatek o številki pošiljatelja. Tudi izhodni dokumenti se ne tiskajo na papir, temveč se potrdijo z žigom elektronsko ter prek sporočilnega sistema in faks strežnika pošljejo osebam izven sistema. Morebitne redke dokumente, ki pridejo po navadni pošti, pa zajamejo s pomočjo skenerja in uvrstijo v elektronski arhiv. Podatki o kompenzacijah se izmenjujejo s podatki iz transakcijskega sistema.

Na ta način se je proces izdelave kompenzacij prenovil in zelo pohitril, poleg tega so vsi dokumenti na voljo v elektronski obliki v kateremkoli trenutku. Prednosti so se kmalu dokazale, zgrajena je bila primerna platforma dokumentnih sistemov, ki je omogočila podobno prenovo in informatizacijo ostalih oddelkov in procesov v podjetju. Ob vpeljavi kartice Pika se je lahko takoj začelo s skeniranjem vseh vrst dokumentov komintentov in s tem z gradnjo elektronskega arhiva. Tudi tu se logični arhiv delno polni s podatki iz transakcijskega sistema, nekateri podatki se pa na nivoju logičnega arhiva prenašajo tudi med aplikacijo kompenzacij in Pika kartice.

Slika 4: Primer dokumentnega sistema v podjetju Mercator d.d.



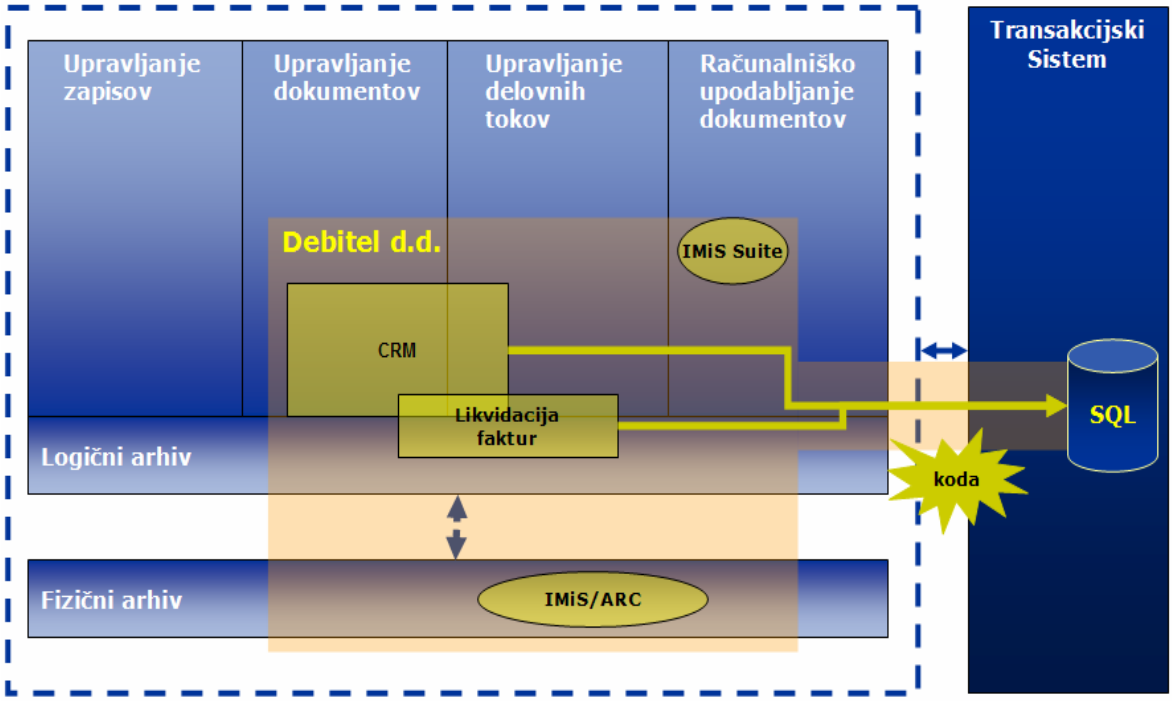
Vir: Ponikvar, 2003

2.2.8.3 Debitel d.d.

Na spletni strani podjetja Debitel d.d. lahko preberemo, da Debitel namenja veliko skrb vzgoji svojih zaposlenih, predvsem v naročniški službi in centru za odnose z naročniki, kjer so stiki z Debitelovimi uporabniki najbolj neposredni. In ravno v naročniški službi so nastale prve težave s papirjem ob širitvi poslovanja in pridobivanju vse večjega števila naročnikov. Naročniška služba je namreč ugotovila, da ima veliko dela s papirno dokumentacijo, ki ni na voljo dovolj hitro, predvsem ko gre za direkten klic naročnika v naročniško službo. Tako je dogajalo, da so tudi trije zaposleni iskali dokument naročnika, medtem ko je ta čakal na telefonski liniji. Nemalokrat se dokument celo ni našel, saj še ni bil v arhivu, ampak v obdelavi v drugih oddelkih. Pokazala se je potreba po hitrejšem pretoku dokumentov in po dostopnosti dokumentov pooblaščenim osebam v vsakem trenutku. Izvedena je bila analiza stanja in ugotovljena nujnost vpeljave dokumentnega sistema za upravljanje odnosov s strankami - CRM (*angl. Customer Relationship Management*). Zanimivo je, da so v začetku delovali brez dokumentnega informacijskega sistema, saj so bili zagonski stroški s strani matičnega podjetja zelo omejeni, tako da so ga začeli postavljati šele, ko so pridobili več kot 5000 naročnikov. Takrat je bilo najrazličnejših pogodb in podobne dokumentacije že toliko, da je nastala prostorska stiska, poleg tega pa so ugotovili, da do določenih dokumentov potrebujejo trenutni dostop, kar je bilo v papirni obliki praktično nemogoče doseči. Matično podjetje je Debitelu predpisovalo nekatere standarde, vendar jim je pri aplikativnih rešitvah pustilo nekoliko bolj proste roke in odločili so se za vpeljavo produktov IMiS. Izdelana je bila aplikativna rešitev, ki je omogočala iskanje podatkov in dokumentov po ključnih podatkih in po vsebini. Poenostavilo se je tudi vnašanje podatkov, saj podatkov ni potrebno vpisovati na več mestih, temveč se prenašajo v dokumentni sistem iz transakcijskega sistema za vodenje baze naročnikov in obratno. Zajem vse papirne dokumentacije se izvaja na enem mestu. Ko pride dokument v podjetje, se s pomočjo procesa skeniranja pretvori v elektronsko obliko in je takoj na voljo vsem pooblaščenim osebam v različnih oddelkih podjetja in na različnih aplikativnih nivojih znotraj dokumentnega sistema. Dokumentacijo, ki je nastala pred implementacijo dokumentnega sistema, so v celoti poskenirali za nazaj s pomočjo zunanjega izvajalca ter jo uvrstili v CRM sistem. Po vpeljavi sistema se je zmanjšalo število ljudi v službi za pomoč uporabnikov, saj delavec vpraša naročnika le za njegovo telefonsko številko in že ima pred sabo vso dokumentno komunikacijo s stranko v elektronski obliki. Pravtako je v trenutku na voljo večina znanja za reševanja problema naročnika in problem rešijo generalisti. Tako večino primerov lahko rešijo takoj v naročniški službi, brez odvečne porabe časa za iskanje dokumenta.

In kakšne so bile še ostale prednosti, ki jih je prinesla postavitev sistema za upravljanje z dokumenti? Začnimo s povsem osnovno – prostorom za zlaganje dokumentov in zanj potrebnimi omarami. Na sedežu podjetja se sedaj hrani le aktivna dokumentacija tekočega tedna, nakar se papirna dokumentacija prenese na cenovno mnogo ugodnejšo lokacijo izven Ljubljane. Pravtako je zelo poenostavljeno sortiranje dokumentacije, saj se vse dnevne dokumente, ne glede na namembnost in pomembnost, po skeniranju zlaga v isto škatlo z označenim datumom. V primeru pravne zahteve (do 5 primerov letno) je v dokumentnem sistemu zelo enostavno ugotoviti dan nastanka dokumenta in tako najti dokument v papirni, pravno veljavni obliki. Analiza podjetja Debitel d.d. je pokazala, da so samo s prostorskimi prihranki in zmanjšanjem števila omar za hranjenje papirne dokumentacije investicijo povrnili v prej kot pol leta!

Slika 5: Primer dokumentnega sistema v podjetju Debitel d.d.



Elektronski arhiv je omogočil iskanje dokumentov po poljubno mnogo ključih, medtem ko je v klasičnem, papirnatem arhivu, mogoče iskati zgolj po enem ključu. Hitrost iskanja se zato občutno poveča. Ravnotako je elektronski dokument mogoče poljubno mnogokrat reproducirati, denimo natisniti ali poslati po elektronski pošti. Seveda se vseskozi delajo tudi rezervne kopije podatkov, ki so shranjene na drugem prostoru, tako da dokumenti v primeru kakšne nesreče niso izgubljeni, kot se lahko zgodi pri papirnatem arhivu. Sicer lahko tudi pri njem naredimo rezervno kopijo, a to zahteva veliko časa, veliko prostora in s tem povezane dodatne stroške.

In nenazadnje, po obdobju uvajanja so se začeli spreminjati tudi delovni procesi v drugih oddelkih podjetja. Finančna služba ima trenuten vpogled v vse prejete račune podjetja in vpeljala je elektronsko likvidiranje računov na aplikativnem nivoju. Proces likvidiranja računov se je skrajšal, dokumentov ni potrebno kopirati, v primeru težav je zagotovljena sledljivost likvidacije, kar zagotavlja bistveno večjo natančnost in nadzor procesa. Tako je v večini delovnih procesov vpeljan vzporedni tok dela in toka dokumentov v elektronski obliki, s čimer je skrajšan čas obdelave dokumenta. Pravtako lahko v vsakem trenutku ugotovimo, v katerem stadiju procesa se dokument nahaja in opozorilni moduli na aplikativnem nivoju omogočajo opozarjanje na zaostanke v obdelavi dokumenta.

Iz zgornjih na kratko opisanih primerov lahko ugotovim, da dokumentni sistemi omogočajo neposredne in posredne pozitivne stroškovne učinke, pravtako omogočajo prenovo delovnih procesov v smeri večje učinkovitosti in produktivnosti. Ker dokumentni sistem omogoča arhiviranje vseh dokumentov v elektronski obliki, se tako v sistem lahko prenesejo tudi dokumenti, ki so sicer na lokalnih diskih uporabnikov. S tem se centralizira znanje podjetja in iskanje vsebine in znanja, ki nastaja v podjetju.

2.3 Arhiviranje dokumentov

Pri velikem številu dokumentov, s katerimi se srečujemo iz dneva v dan, igra najpomembnejšo vlogo ustrezen sistem za shranjevanje, oziroma arhiviranje dokumentov, saj je dokumentom nemogoče slediti ročno, oziroma preko datotečnega sistema. Tovrstno početje je vprašljivo tudi iz vidika varnosti podatkov. Tu ne mislim le na shranjevanje skeniranih dokumentov, temveč vseh dokumentov v sistemu za elektronsko upravljanje z dokumenti. Zato je pred načrtovanjem sistema za arhiviranje potrebna podrobna analiza predvidene količine dokumentov v določenem časovnem intervalu, količina “on-line” dokumentov, ki naj se nahajajo na medijih in morajo biti na voljo ob minimalnih dostopnih časih. To imenujemo profiliranje sistema in je bistvenega pomena ob snovanju arhitekture arhivskega dela.

Zaradi že omenjenih večjih potreb po diskovnih kapacitetah, so že davno pred pojavom osebnih računalnikov razvili in uporabljali posebne programske produkte, ki so omogočali hierarhično arhiviranje dokumentov. To so HSM (*angl. Hierarchical Storage Management*) sistemi, ki skrbijo za optimalno zasedenost pomnilnih medijev in za čim ustreznejšo razpoložljivost datotek. Večnivojski sistem arhiviranja lahko organiziramo tako, da določimo hierarhični vrstni red pomnilniških medijev za posamezne skupine dokumentov (izbrane glede na profil), na osnovi

katerega potem HSM sistem razvršča, arhivira in migrira dokumente glede na stopnjo uporabe, oziroma glede na frekvenco dostopa. Pri tem je najpomebneje, da uporabnikom, oziroma nadzornikom sistema ni potrebno skrbeti, na katerem mediju se posamezni dokument nahaja, temveč jih HSM sam prestavlja po nivojih (trdi disk, izmenjalnik optičnih diskov, zunanje enote, itd). Na tak način je onemogočeno tudi sicer pogosto izgubljanje in nepotrebno kopiranje papirnih dokumentov za primer, ko isti dokument hkrati potrebuje več uporabnikov. Z uporabo različnih medijev (dražjih in cenejših) zgradimo elektronski arhiv, v katerem dokumenti kontrolirano migrirajo iz medija na medij, na osnovi predhodno definiranih pravil oziroma postopkov znotraj profilov dokumentov. (Žerko, 2003)

2.4 Računalniško upodabljanje (skeniranje) dokumentov

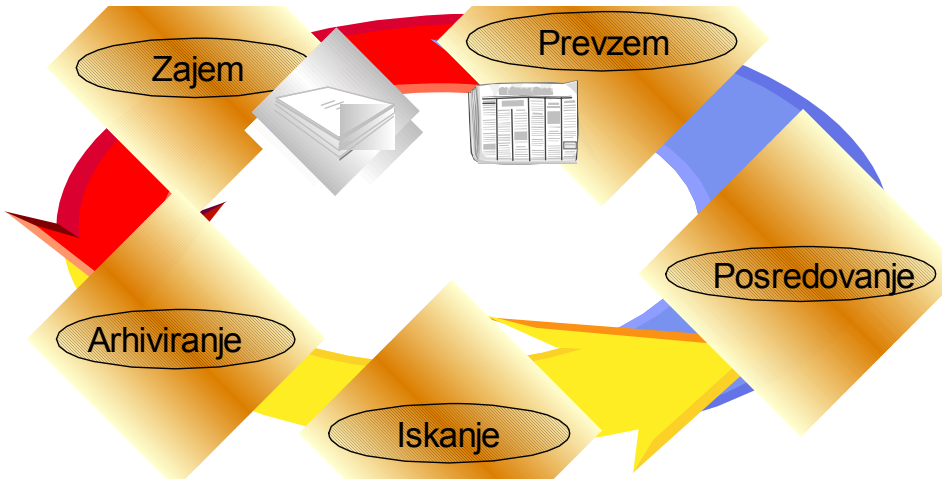
Sistemi za računalniško upodabljanje dokumentov predstavljajo vhodni del sistemov za upravljanje z dokumenti, kjer je izvorni dokument v papirni obliki. Brez učinkovitega sistema zajema in arhiviranja papirnih dokumentov, imamo opravka zgolj z nekakšnim evidenčnim sistemom, ki omogoča evidenco papirnih dokumentov, dokumente same pa še vedno fotokopiramo, da počasi krožijo in se izgubljajo. Skeniranje dokumentov je torej nadgradnja za dokumentne (zgolj evidenčne) sisteme, ki slonijo na papirni obliki dokumenta.

Življenjski ciklus (skeniranega) dokumenta je proces ponavljajočih se faz in aktivnosti (glej sliko 6) (Žerko, 2003):
- zajem in vnos podatkov v sistem za upravljanje dokumentov,
- indeksiranje podatkov,
- arhiviranje,
- procesiranje podatkov iz dokumentov,
- distribuiranje,
- iskanje, pregledovanje in/ali popravljanje.

Na osnovi teh ugotovitev lahko predvidimo temeljne funkcije, ki jih mora imeti ustrezna programska oprema (Avedon, 1995):
- kontrola, nadzor delovanja čitalca (skenerja) in skeniranje samo,
- kontrola kakovosti, oziroma vidnosti dokumenta (osvetljenost, kontrast, povečava),
- izboljšanje vidljivosti (čiščenje ozadja) in vertikalna poravnava strani,
- vpis dodatnih opomb in komentarjev, kot sestavni del skeniranega dokumenta,
- optično prepoznavanje znakov in priprava podatkov za iskanje po polnem besedilu (*angl. full text search*),

- pregledovanje in arhiviranje dokumentov.

Slika 6: Življenski ciklus dokumenta



Vir: Žerko, 2003

Pri skeniranju znotraj dokumentnih sistemov ločimo različne načine skeniranja. Ti so odvisni predvsem od izvora, količine in oblike odkumentov:

- ***posamično skeniranje*** ob vpisu evidenčnih podatkov in določanju profila dokumenta.
- ***skeniranje v paketu* (*angl. batch scanning*):** več dokumentov naložimo na podajalec in sistem vsakič, ko poskenira dokument, ponudi uporabniku pogovorno okno za vpis indeksnih podatkov. Pri tem je obstajajo različni načini določanja separatorja, kdaj se konča en in kdaj začne drug dokument (*angl. document separator*). Separator je lahko prazna stran (zahteva ustrezno pripravo), črtna koda (zelo priporočljiv in najzaneslivejši pristop) ali pa enostavno skeniramo vse dokumente z vnaprej določenim številom strani. V določenih primerih okno za vpis indeksnih podatkov niti ni potrebno, če s pomočjo črtne kode lahko pridobimo unikatni podatek, ki določa ostale indeksne podatke s poizvedbo v podatkovni bazi.
- ***skeniranje v paketu s tehnologijo prepoznavanja obrazcev*** je v bistvu skeniranje v paketu, vendar uporabnik ne vpisuje podatkov, pač pa jih skuša sistem prepoznati sam. Ta način je uporaben zgolj tam, kjer skeniramo enake vnaprej pripravljene obrazce. ICR je smiselen zgolj, če lahko prepoznano vsebino tudi preverimo (logične kontrole, povezave na transkacijski sistem, itd).
- ***skeniranje kot zunanja storitev* (*angl. outsourcing*):** skeniranje izvaja specializirano podjetje, ki prevzame dokumente ter vrne medij s skeniranimi objekti ter opisi atributov. Cena storitve je določena na dokument in je odvisna predvsem od števila dokumentov (s količino cena pada) ter zahtevanega števila indeksnih podatkov (s številom cena raste). Outsourcing lahko izvajalec izvaja tudi pri stranki. Velikokrat je tovrstna storitev primerna ob vpeljavi dokumentnega

sistema, ko zagotovimo skeniranje dnevnih dokumentov, medtem ko skeniranje dokumentov iz preteklega obdobja (*angl. backfile conversion*) prepustimo zunanjemu izvajalcu.

S skeniranjem pridobimo dokument v elektronskem grafičnem formatu in ga lahko opremimo z dodatnimi podatki, predvsem v obliki anotacij in obarvanjem različnih delov dokumenta. Tu se že pojavlja kot dodatek modul optičnega prepoznavanja znakov, ki lahko razpozna cel dokument ali le posamezen del. Poleg funkcij skeniranja in prepoznavanja, omenimo še podporo različnim skenerjem, možnost shranjevanja v različnih slikovnih formatih in podpora primernim algoritmom zgoščevanja (*angl. compression)* slikovne datoteke ter različne nastavitve skeniranja, kot je resolucija, osvetljenost, kontrast, itd.

2.4.1 Zgoščevanje

Arhiviranje skeniranih dokumentov zahteva veliko pomnilniškega prostora, zato je uporaba primernih algoritmov zgoščevanja nujno potrebno. V preteklosti se je zgoščevanje izvajalo s pomočjo posebnih strojnih vmesnikov, zdaj so na voljo učinkovite programske rešitve. Tehnike in algoritmi zmanjšajo velikost datoteke z odstranitvijo bitov, ki predstavljajo neuporabljeni bel prostor na dokumentu, ali večje črne površine. Teoretično bi A4 stran poskenirana v črno beli tehniki, z ločljivostjo 200 dpi (*angl. dots per inch*) zavzela 200x200x8,5x11 = 3,74 milijona bitov. To pomeni 467,500 zlogov (456 KB) prostora. Enak račun pri 300 dpi (resolucija priporočljiva za optično prepoznavanje znakov) da 1,027 mega zlogov prostora. Pri veliki količini poskeniranih dokumentov seveda hitro pridemo do giga ali celo tera zlogov potrebnega prostora. ITU-TSS *(angl. International Telecomunication Union-Telecommunication Standardization Sector)*, bolj znan s prejšnjim imenom CCITT (*angl. Committee for International Telephony & Telegraphy)* je opredelil dva standardna algoritma zgoščevanja za prenos faksiranih sporočil: Group 3 (G3) in Group 4 (G4). Ta dva algoritma sta bila kasneje prirejena za zgoščevanje slikovnih črno belih datotek – standardno formata TIFF (*angl. tagged image file format*). Stopnja je seveda odvisna od zasedenosti površine skeniranega dokumenta, v povprečju se giblje v razmerju med 7:1 in 30:1 (Koulopoulos, 1995; Avedon, 1996). Na tržišču je že nekaj časa standard JBIG (*angl. Joint Bi-level Image Experts Group*), patent podjetja IBM, ki ima povprečno dvakrat večjo stopnjo zgoščevanja v primerjavi s standardom G4 in je še posebno primeren za zgoščevanje t.i. sivinskih (*angl. grayscale)* slik z večbitno globino (Mantelman, 1995). Seveda ne smemo pozabiti na zgoščevanje dokumentov poskeniranih v barvnem načinu, kjer prevladujejo JPEG algoritnmi,

npr. sekvenčni ali progresivni JPEG, vendar je v poslovnih dokumentnih sistemih količina barvnih dokumentov v primerjavi s črnobelimi zanemarljiva.

2.4.2 Vpliv kvalitete skeniranja na optično prepoznavanje znakov

Kvaliteto prepoznavanja lahko poveča skeniranje v več sivinskih odtenkih, kar pomeni zapolnjevanje praznin (lukenj) v znakih, ki so posledica slabega tiska ali nekvalitetnega papirja. Vendar se zaradi tega (več bitov za barve in globino) poveča velikost slikovne datoteke, kar je potrebno upoštevati pri arhiviranju. Potreben je neke vrste kompromis, čeprav so novi algoritmi na področju OCR sistemov vse manj občutljivi na tovrstne parametre skeniranja. Enako velja za določanje ločljivosti. Ločljivost (*angl. resolution*), ki je potrebna za prepoznavanje se giblje med 240 in 400 dpi, ponavadi je primerna vrednost 300 dpi. Zanimivo je, da v primeru motenj (*angl. noise, garbage*) oziroma majhnih pack, kot posledic nekvalitetnega tiska v slikovnem formatu, dosežemo boljši rezultat pri skeniranju z nižjo ločljivostjo. Pomembno vlogo igra tudi osvetljenost in kontrast pri skeniranju, vendar tu ni nekih posebnih pravil nastavljanja, tako da je potrebno v primeru slabega rezultata prepoznavanja določiti te parametre s poizkušanjem. Paziti je potrebno tudi na vertikalno poravnavo papirja, čeprav današnji komercialno dosegljivi OCR sistemi avtomatično poravnavajo (*angl. deskew*) nagib do 15° v obe smeri.

2.4.3 Pravna veljavnost skeniranih dokumentov

Gre za lastno laično mnenje glede pravne veljavnosti skeniranih dokumentov, saj nikjer nisem zasledil uradnega pravnega mnenja na to temo. Pravna veljavnost dokumenta je pomembna le tedaj, ko je dokument ali podatek iz dokumenta sporen. V tem primeru se glede na trenutno zakonodajo po mojem mnenju nikoli ne bo upošteval poskeniran dokument, ki je originalno oziroma izvorno v papirni obliki. Četudi je elektronsko podpisan, to nikoli ne bo pomenilo, da je enak originalu, saj je nastal v drugi obliki in je bil lahko pred, med ali po elektronski pretvorbi spremenjen. Postaviti bi bilo potrebno sistem oziroma standard, ki bi zagotavljal, da se skenirani dokument od nastanka naprej ne more spremeniti (npr. uporaba WORM enot - Write Once Read Many), pa še takrat se poraja vprašanje avtentičnosti zaradi dvoma o morebitni potvorjenosti dokumenta še preden je ta nastal v elektronskem arhivu. Zato se kot dokazilo v sporu ne more uporabiti. Papirna dokumentacija se torej v Sloveniji mora hraniti v papirni obliki, če želi služiti kot dokaz v sporu. Eno od podjetij, s katerimi sem sodeloval ob uvedbi dokumentnega sistema, je zahtevalo pravno mnenje glede pravne

veljavnosti poskeniranih dokumentov, saj so po uspešni implementaciji sistema želeli poskenirati celotno poslovno dokumentacijo za nazaj in papirni arhiv uničiti. Izkazalo se je, da morajo zakonsko zahtevane dokumente za primere sporov hraniti v papirni obliki.

Zakon o elektronskem poslovanju in elektronskem podpisu (ZEPEP) je pomemben za t.i. e-dokumente, kateri so že izvorno v elektronski obliki. Projekt eSLOG Gospodarske zbornice Slovenije, ki omogoča elektronsko izmenjavo podatkov bo torej omogočal pravno veljavne elektronske dokumente, saj bo veljaven elektronski podpis dokazoval pravno veljavnost dokumenta. Enako velja za npr. sistem edavki Davčne uprave Republike Slovenije. Podobno v svojem članku razmišlja tudi Žarko Štrumbl, ko govori o pravnih in ekonomskih vidikih hrambe dokumetnega gradiva, tako v papirnati kot v elektronski obliki. Štrumbl ugotavlja, da sicer obstoječa zakonodaja načelno predpisuje hrambo in roke hranjenja v elektronskem arhivu, vendar je nedorečena. Pri tem omenja Zakon o arhivskem gradivu (ZAGA)[2] in Zakon o elektronskem poslovanju in elektronskem podpisu (ZEPEP)[3]. Problem je v zagotavljanju avtentičnosti elektronskega dokumenta, saj v Republiki Sloveniji nimamo pravno formalne zagotovitve avtentičnosti in niti ZAGA niti ZEPEP tega (še) ne zagotavljata. Elektronski arhiv je torej potrebno pravno urediti s sprejemom ustreznega standarda, vendar žal o tem trenutno opozarjajo le posamezniki (Štrumbl, 2004). Neredko se zgodi, da kateri od ponudnikov dokumentnih sistemov v Sloveniji trdi, da je poskeniran dokument pravno veljaven. Po mojem mnenju gre zgolj za marketinško prodajne motive. Bistveno je, da pravna veljavnost ne sme biti glavni motiv vpeljave sistema za skeniranje dokumentov. Elektronski arhiv naj nastane zaradi podpore dokumentov v delovnih in poslovnih procesih, trenutnega dostopa do dokumentov v elektronski obliki ter neposrednih in posrednih stroškovnih koristi elektronskega arhiva in preoblikovanih procesov. Če bo pa poskeniran dokument kdaj uradno pravno veljaven dokument, bodo pa podjetja z elektronskim arhivom v prednosti, saj bodo pravno veljavne poskenirane dokumente že imeli.

2.5 Krmiljenje delovnih procesov

Krmiljenje delovnih procesov je eden ključnih delov sistema za elektronsko upravljanje dokumentov. Primerjamo ga lahko s konceptom avtomatizacije

[2] Ur. L. RS 20/97
[3] Ur. L. RS 57/00

proizvodnje s tekočim trakom. Omogoča avtomatizacijo informacijskih nalog in aktivnosti z definicijo pravil, vlog in usmerjanj (*angl. rules, roles & routing*). Je proces, ki povezuje ljudi in postopke v poslovnem procesu, da bolj učinkovito izvršujejo naloge in si medseboj lažje izmenjujejo informacije. Osnovna predpostavka je sprejeti pisarniško okolje kot proizvajalca dokumentov skozi več povezljivih faz, oziroma nalog. V tej arhitekturi nastopa aplikacija kot povezovalec teh nalog in aktivnosti, ki nepotrebne izloči, ostale pa avtomatizira. Za tovrstne sisteme sem našel v literaturi več izrazov, kot na primer kroženje dokumentov, bolj primerno bi ga lahko poimenovali (informacijsko) definiranje delovnih postopkov ali pa krmiljenje delovnih procesov. Za definiranje delovnih postopkov je potrebna predhodna analiza sistema in temu primerno kasnejše načrtovanje (Koulopoulos, 1995).

Dejavnosti se znotraj postopka lahko izvajajo zaporedno in/ali vzporedno ter tudi pogojno (dinamično). Vsak del postopka ima predvidenega izvajalca, ki je lahko človek, skupina ljudi ali pa celo sam informacijski sistem. Vključen naj bo tudi sistem obveščanja, ki se aktivira na prehodu iz ene v drugo fazo, ob spremembi dokumenta, ali pa ob poteku časa predvidenega za določeno dejavnost, zato je nujna povezava s sporočilnimi sistemi.

Sami dokumenti se med delovnim postopkom lahko spreminjajo, zato je prav tako pomembna sledljivost dokumentom (*angl. access logging*) in kontrola različic (*angl. version control*). Aplikacija mora omogočati varnost dostopa do dokumenta (*angl. access control*), ki naj bo odvisna tudi od trenutne faze procesa. V povezavi z nadzorom dokumenta je poleg sledljivosti in kontrole različic potrebno omeniti še mehanizme kot so: vrnitev dokumenta v sistem (*angl. check-in*) ter rezervacija dokumenta v sistemu (*angl. check-out*).

V sistemu za upravljanje z dokumenti ima krmiljenje delovnih procesov pomembno vlogo, saj omogoča kontrolo pretoka dokumentov, ki podpirajo poslovne procese.

2.6 Optično prepoznavanje znakov znotraj dokumentnih sistemov

Sistemi optičnega prepoznavanja znakov (v nadaljevanju OCR sistemi) so zaradi avtomatičnega in hitrejšega zajemanja podatkov povzdignili popularnost in uporabnost sistemov za upodabljanje dokumentov. Omogočili so nove prijeme v avtomatizaciji kroženja dokumentov in iskanje po polnem besedilu oziroma polni vsebini dokumenta.

OCR sisteme lahko uporabljamo na različne načine znotraj sistemov za skeniranje dokumentov. Prepoznavanje lahko izvedemo že ob samem skeniranju dokumenta, za celoten dokument ali pa le posamezen del. Enako velja v modulu za pregledovanje skeniranih dokumentov. Rezultat lahko uvozimo v določeno polje (npr. polje vsebine) v sistemu za upravljanje dokumentov, lahko kreiramo nov dokument v urejevalniku besedil in s tem nov samostojen dokument v sistemu, ali pa ga uporabimo le za potrebe indeksiranja in kasnejšega iskanja po polnem besedilu. Način uporabe je seveda odvisen od namena (in zmogljivosti) posameznih aplikacij, oziroma od tega, zakaj potrebujemo poleg računalniško upodobljenega dokumenta še zapis v tekstovni obliki. Današnji sistemi za računalniško upodabljanje dokumentov imajo v večini primerov OCR že integriran, druga možnost je integracija OCR funkcij v aplikacije prek razvojnih orodij proizvajalcev OCR sistemov.

Poleg opisanega interaktivnega OCR-a, se v sistemih z veliko skeniranimi dokumenti zaradi relativno velike procesorske zahtevnosti optičnega prepoznavanja, uporablja tudi paketno prepoznavanje. Gre za optično prepoznavanje skupin dokumentov, ki jih npr. poskeniramo čez dan, aplikacijski OCR strežnik pa jih obdela skozi proces optičnega prepoznavanja ponoči, ker je takrat procesorska in komunikacijska obremenitev manjša.

2.7 Problematika vnosa evidenčnih podatkov

Ob skeniranju in vnosu dokumenta v dokumentni sistem, je potrebno dokument do neke mere opisati, oziroma mu pripeti evidenčne podatke in ključne besede, predvsem zaradi kasnejšega iskanja dokumenta. V večini primerov to opravijo operaterji in delo je zamudno ter velikokrat predstavlja ozko grlo v delovnem toku dokumenta. Dejansko je pomembno optimizirati proces zajema dokumenta na način, da delovno mesto s skenerjem vnaša čim manj evidenčnih podatkov, ki naj se vstavijo v zaporednem procesu na drugih delovnih mestih. V sklopu pisarniškega poslovanja se praviloma dokumenti knjižijo v vložišču oziroma glavni pisarni.

Včasih sama narava delovnega procesa omogoča avtomatsko vstavljanje podatkov, kot smo videli na primeru podjetja Mercator d.d., ko se dokument iz faks strežnika vmesti v dokumentni sistem glede na številko pošiljatelja. Pravtako današnje programske rešitve skeniranja s pomočjo dodatnih orodij za obdelavo dokumenta (*angl. Image Processing – IP*) lahko iščejo in preberejo vrednosti črtnih kod na dokumentu v fazi skeniranja in vrednost lahko predstavlja ključ za umestitev dokumenta v sistem in pridobitev nekaterih evidenčnih podatkov.

Posebna kategorija vnašanja podatkov iz dokumentov so sistemi za prepoznavanje obrazcev.

2.8 Vloga sistemov za prepoznavanje obrazcev

Velikokrat imamo opravka z obrazci, ki jih uporabljamo za prenos informacij med posamezniki in podjetji. Pomislimo samo, s koliko vrstami obrazcev se srečujemo vsakodnevno: bančni, zavarovalniški, ankete, popisi, itd. V poslovnem svetu so obrazci pomemben medij za prenos informacij med oddelki, med podjetjem in kupcem ter med podjetji. Najbolj se jih uporablja v zavarovalništvu, bančništvu, zdravstvu ter v transportni industriji (Rapaport, 1997). Pri obrazcih ne gre več za klasično pojmovanje vsebine dokumenta, temveč so bistvenega pomena vnešeni podatki na obrazcu. Splošno procesiranje obrazcev je stroškovno zelo potratno opravilo. Obrazce je potrebno oblikovati, natisniti, distribuirati, izpolniti in nato podatke prenesti v informacijski sistem. Po podatkih organizacije Association for Information and Image Management (AIIM) so obrazci dominatna pot informacije v poslovnem svetu: kar 80% vseh poslovnih dokumentov so obrazci, povprečno 60 milijard ameriških dolarjev se porabi za tiskanje obrazcev, kar 360 milijard dolarjev pa za prenos podatkov iz obrazcev v informacijske sisteme (AIIM, 2004). Ta velik odstotek izhaja iz pojmovanja obrazca kot dokumenta, iz katerega potrebujemo določene podatke. Tako o obrazcih lahko govorimo tudi v primeru prejetih računov, dobavnic, letalskih kart, ipd. Današnji algoritmi definiranja in prepoznavanja dinamičnih obrazcev omogočajo avtomatsko prepoznavanje prispelih računov ne glede nato, da ni stalne strukture.

Seveda takoj pomislimo na rešitev za zmanjšanje stroškov v elektronskem izpolnjevanju obrazcev (Internet), vendar smo na žalost v veliko primerih še vedno vezani na papir in izpolnjevanje na roko. Vzroki so predvsem v informacijski kulturi izpolnjevalcev ter pravni veljavnosti, v veliko primerih pa sama narava procesa izpolnjevanja obrazcev ne omogoča elektronskega vpisovanja.

V primerih izpolnjenih obrazcev na papirju imamo torej le dve možnosti: vključiti veliko ljudi v ročno vnašanje podatkov ali implementirati sistem avtomatskega prepoznavanja obrazcev. Ročno vnašanje podatkov je tradicionalno dolgočasen, težaven in zastarel proces. Na ta način se lahko soočimo z različnimi problemi in težavami, kot so stroški dela, napake ob vnašanju, stroški prostora, itd. V primeru pravilnega načrtovanja in implementacije sistema za avtomatsko prepoznavanje obrazcev, se je mogoče teh težav rešiti. Kakor je pravilo v dokumentnih sistemih, se tudi pri sistemih prepoznavanja obrazcev rado zgodi, da zrastejo mnogo prek

meja prvotne namestitve. Ko se namreč izkažejo za uspešne pri eni vrsti obrazcev, se jih da z relativno majhnimi stroški uporabiti tudi za druge vrste obrazcev oziroma za druge procese v podjetju. S tem se pa seveda izkažejo mnogo večji prihranki in hitrejši čas povrnitve investicije, kot je sprva kazalo. Tehnologije skeniranja in prepoznavanja dokumentov lahko torej stroške in čas vnosa podatkov občutno znižajo.

3. TEHNOLOGIJA OPTIČNEGA PREPOZNAVANJA ZNAKOV

3.1 Zgodovinski razvoj

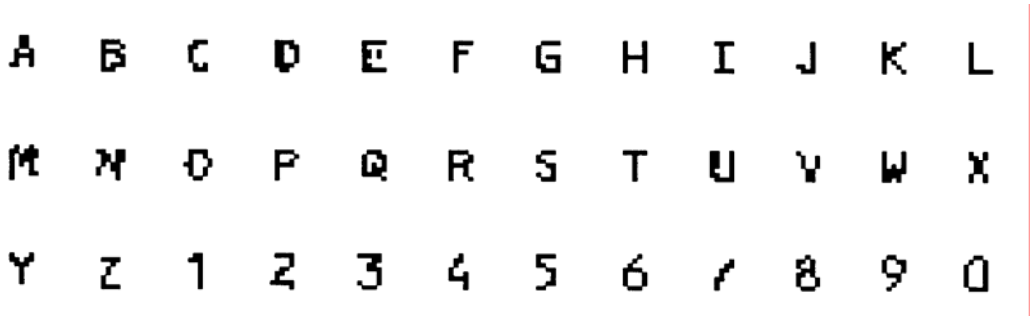
Zgodovina sistemov optičnega prepoznavanja znakov je relativno stara na področju prepoznavanja vzorcev. Kljub začetnim uspehom in prepričanjem, da gre za relativno lahko rešljiv problem, se je kmalu izkazalo, da ovira ni tako nizka. Na srečo so bile zahteve trga dovolj močne, da je prišlo do investiranj v razvoj tovrstnih sistemov (Rapaport, 1997).

O nekakšnih zametkih OCR sistemov lahko govorimo že v 19. stoletju, o modernih sistemih pa od 1950. leta dalje. Prvi uspešni poskus na tem področju je izvedel ruski znanstvenik Tyurin leta 1900. Goldberger je leta 1912 patentiral stroj, ki je bral tipkane črke in jih pretvoril v telegrafsko kodo. Leta 1929 je nemški znanstvenik Tausheck prvi patentiral OCR v Nemčiji, štiri leta kasneje pa še Handel v ZDA. To je vzbudilo sanje o stroju, ki bi znal brati črke in številke. Sanje so se začele uresničevati s pričetkom infomracijske dobe v 50-ih letih. Osnovna ideja je vredna omembe, ker je še danes živa. Pri Tausheckovem patentu je šlo za princip podobnosti, oziroma primerjalnih metod z obstoječimi vzorci (*angl. template matching*) (Mori, 1992).

Kasnejši najbolj znani poskusi na področju OCR sistemov so Sheppardov bralno pisalni robot imenovan GISMO, Kelner in Glaubermanov magnetni pomikalni register iz leta 1956, pa Hannanova metoda primerjave iz leta 1962, vendar tedanji sistemi niso doživeli komercialnih uspehov. Do zgodnjih sedemdesetih let so OCR sistemi podpirali le omejeno število točno definiranih pisav (kot nabori OCR-A in OCR-B, glej sliki 7 in 8) in bili prisotni le v nekaj ozkih, specializiranih področjih vladnih agencij in velikih korporacij (McCarthy, 1994; Schantz, 1982).

Slika 7: Nabor znakov OCR-A

Slika 8: Nabor znakov OCR-B



Kasnejša predstavitev sistemov, ki so se lahko naučili različnih pisav, je razširila trg in povzročila priljubljenost ter komercialno uspešnost OCR sistemov. Predvsem je vredno omeniti Hitachijev H 8959 laserski skener z OCR procesorjem v sredini 70-ih, ki ga imajo za enega mejnikov v razvoju OCR-a, saj je imel za tisti čas izredne sposobnosti in relativno nizko ceno. Z naglim razvojem strojne opreme so se OCR sistemi preselili v območje programskih in ne več strojnih rešitev. Tehnologija se je radikalno spremenila ob koncu 80-ih let z razvojem sofisticiranih sistemov, ki so se dvignili nad preprosto primerjavo z obstoječimi vzorci. Topološke oziroma strukturne analize so izpodrinile prepoznavanja s pomočjo primerjave in so teoretično omogočile prepoznavanje različnih znakov (pisave) glede na unikatne lastnosti njihove oblike. Danes pa vse bolj prevladujejo algoritmi na osnovi tehnologije nevronskih mrež. Vsaka metoda ima seveda svoje prednosti in slabosti, implementacija pa je odvisna predvsem od narave zapisa in problema, ki zahteva prepoznavanje (Mori, 1992; Schantz, 1982).

3.2 Delitev sistemov prepoznavanja znakov

V literaturi o OCR sistemih sem zasledil več načinov delitve. Opisal bom delitev glede na namen in glede na vrsto zapisa. Same metode in tehnologije (kot eden možnih načinov delitve) so opisane v nadaljevanju tega poglavja.

3.2.1 Glede na namen

Delitev sistemov na industrijski in neindustrijski OCR lahko obravnavamo kot delitev glede na namen. Razlike so predvsem v robustnosti in zanesljivosti sistemov. Pri industrijskih sistemih so napake nedopustne in v primeru, da verjetnost prepoznanega ni zelo velika je potrebno sporočiti napako, medtem ko je pri neindustrijskem OCR-u rezultat najbolj verjeten znak in redkeje pride do zavrnitve znaka.

Prav tako se v industrijskih OCR sistemih ponavadi zgradijo algoritmi in metode prepoznavanja za vsak problem posebej, predvsem zaradi prilagoditve naboru znakov in pogojem delovanja.

3.2.2 Glede na vrsto zapisa

Optično prepoznavanje znakov delimo glede na vrsto zapisa na:

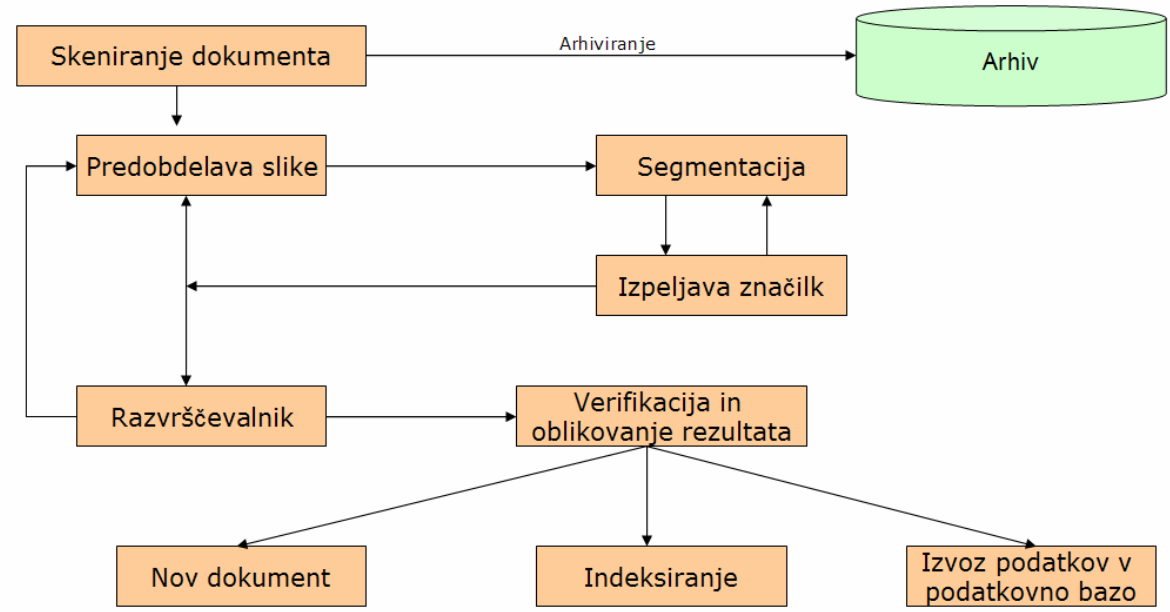
- prepoznavanje ene vrste pisave (*angl.* ***Fixed-font*** *character recognition*), ki je vnaprej znana, kar je primer predvsem pri industrijskih sistemih. Prevladujejo primerjalne metode.
- prepoznavanje poljubne pisave (*angl.* ***Omnifont*** *character recognition)*, ki zajema prepoznavanje natipkanega teksta in naj bi bilo čim bolj neodvisno od oblike pisave. Omogočalo naj bi tudi interakcije s samim uporabnikom, kar pomeni učenje znakov in stilov pisav. Sodobni sistemi imajo povezavo z raznimi slovarji, ki omogočajo avtomatično odpravljanje napak. Prevladujejo metode strukturne oziroma topološke analize.
- **ICR** (*angl. Intelligent character recognition*) – inteligentno prepoznavanje znakov. Ta oblika se navezuje predvsem na prepoznavanje ročne pisave (*angl. handwritting recognition).* Prevladujejo metode na osnovi nevronskih mrež.
- **OMR** (*angl. optical mark recognition*) – prepoznavanje označevanj, kjer gre za prepoznavanje označevanj na obrazcih in izhodni rezultat je označeno/neoznačeno. Tovrstno prepoznavanje je predvsem pomembno v procesih avtomatskega prepoznavanja obrazcev in zapisa rezultata v transakcijski sistem.
- prepoznavanje črtnih kod (*angl.* ***Barcode*** *recognition*).

3.3 Postopek prepoznavanja znaka

Pri obravnavanju zgodovinskega razvoja smo že omenili, da poznamo dve veliki skupini metod za prepoznavanje: **matrično primerjalno** (*angl. pattern matching,*

matrix matching, template matching) in **strukturno razvrščevalno metodo** (*angl. structural classification, topological analysis*). Znane so še metode na osnovi verjetnostne teorije in pa seveda metode na osnovi **nevronskih mrež**. Velikokrat se v sistemih metode med seboj prepletajo in tako dobimo hibridne strukture. Najprej si oglejmo osnovne korake prepoznavanja, ki so več ali manj skupni vsem metodam (glej sliko 9).

Slika 9: Osnovni koraki prepoznavanja



3.3.1 Predobdelava slike

V želji po doseganju čim boljšega rezultata prepoznavanja, je potrebno slikovni format nekoliko transformirati pred samim procesom prepoznavanja. Govorimo o prvem delu prepoznavanja, kar imenujemo predobdelava (*angl. preprocessing*) slike. Poskušal bom zajeti vse osnovne potrebne postopke, katere vsebuje večina OCR sistemov.

OCR orodja vsako sliko najprej pretvorijo v interni format, ki je odvisen od samega orodja in uporabljenega algoritma. Čiščenje slike (*angl. despeckle*) izloči motnje (npr. posamezni izolirani slikovni elementi), ki so nastale zaradi nekvalitetnega tiska ali zaradi napak pri skeniranju. Nato se izvede izločitev nepotrebnih delov, kot so črte, okvirji, ozadja in podobno in tovrstni elementi se po potrebi uporabijo pri oblikovanju (formatiranju) rezultata. Izločitev podčrtanih delov besedila denimo pripomore k temu, da se znaki med seboj ne dotikajo. Po potrebi se poveča kontrast. Sestavni del predprocesiranja je tudi avtomatična poravnava manjših nagibov, kar dosežemo z izračunom povprečnih kotov nagiba po vseh vrsticah in

po potrebi z nevtralizacijo tega kota. V primeru rotirane slike orodja omogočajo avtomatsko odkrivanje orientacije in s tem poravnavo.

Naslednja faza je analiza oziroma dekompozicija slike. Gre za razčlenitev vsebine slike na bloke, pri čemer je predvsem pomembna določitev tipa bloka, saj imajo različni tipi prirejene različne algoritme prepoznavanja:

- tekst,
- tabela,
- grafični element,
- črtna koda,
- itd.

Ta proces nekateri avtorji uvrščajo v proces segmentacije in sicer kot začetno segmentacijo na nivoju dokumenta. OCR sistemi omogočajo uporabniku, da celo sam popravi tipe blokov, če je to potrebno. Nato se ti bloki obdelajo v procesu prepoznavanja, eden za drugim. Vsi ti procesi so v bistvu del procesa izločitve teksta iz grafične podobe, kar je predpogoj za kasnejšo segmentacijo in prepoznavanje znaka. V primeru obdelave obrazca s predhodno definiranimi lokacijami podatkov in pravili, kar je shranjeno v predlogi obrazca, se izvede tudi proces identifikacije obrazca.

3.3.2 Segmentacija

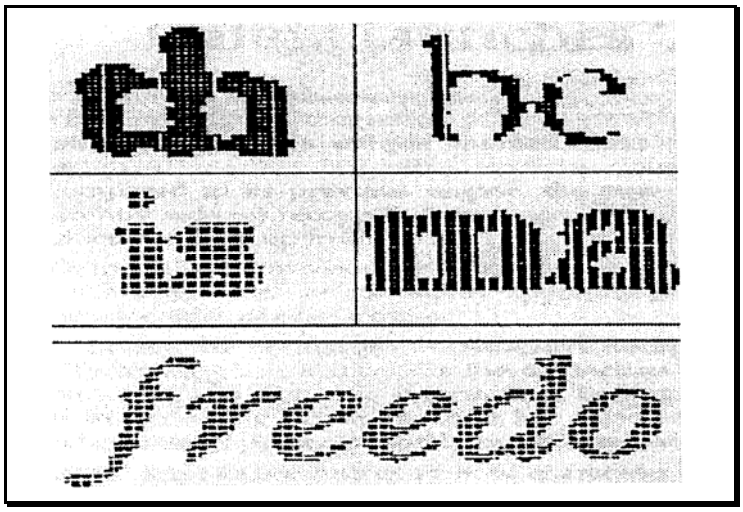
S segmentacijo razgradimo sliko posameznega bloka na slike posameznih znakov, ki jih kasneje ločeno prepoznamo. Segmentacija poteka v več korakih. V bistvu je prvi korak segmentacije že omenjena dekompozicija dokumenta. Pri prepoznavanju blokov je postopek običajno razdeljen na segmentacijo stolpcev, vrstic, besed in končno znakov. Segmentacija je otežkočena predvsem v primeru (glej sliko 10) zlepljenih (*angl. touching characters*) ali prelomljenih znakov ter delno prekrivajočih se območij znakov (*angl. kerning*), kar se dogaja predvsem pri ročni pisavi. Dodatni nivoji in algoritmi segmentacije poskušajo te napake odpraviti, saj v nasprotnem primeru pride do hipersegmentacije (več znakov združenih v enega) ali podsegmentacije (znak razdeljen) znakov (Nadler, 1993).

Metode segmentacije delimo v tri skupine:

- od zgoraj navzdol *(angl. top-down)*
- od spodaj navzgor *(angl. bottom-up)*
- hibridne

Ob segmentaciji se ponavadi izvajajo procesi normalizacije znaka na standardno velikost in včasih tudi stanjšanje na debelino enega slikovnega elementa (*angl. thinning algorithm*). Poudariti velja, da so avtorji precej neenotni pri opisih in delitvi procesov predobdelave in segmentacije slike. V nalogi se zato dosledno ne zgledujem po nobenem od njih in sem procese razdelil, kot se mi je zdelo najbolj primerno.

Slika 10: Primer zlepljenih (zgoraj) in prekrivajočih se znakov (spodaj)



3.3.3 Iskanje značilk znaka

Po postopku segmentacije imamo pripravljen znak za vhod v naslednji korak, ki znak zdefinira in določi lastnosti, značilke znaka (*angl. feature extractor*). Ravno definicija značilk je poleg kvalitete slike znaka in vrste pisave bistvenega pomena za uspeh metode. Rezultat izpeljave značilk je vektor značilk, ki vsebuje informacije o znaku, ki so vhodni del razvrščevalnika.

3.3.4 Razvrščanje v razrede

Razvrščanje je izvedeno s klasifikacijo vrednosti značilk. Rezultat izpeljave vrednosti značilk primerjamo z referenčnimi vzorci posameznih razredov in ga razvrstimo v tisti razred, kamor spada referenčni vzorec z najmanjšo razdaljo, oziroma z največjo mero ujemanja. Tu je potrebno opozoriti, da v tem primeru čas prepoznavanja linearno narašča s številom referenčnih vzorcev. Druga možnost je klasifikacija z odločitvenimi drevesi. Uporablja se v primerih, ko imamo opravka z diskretnimi značilkami, ki lahko zavzamejo le malo različnih vrednosti, tipično dve ali tri. Dobra stran teh metod je, da čas razvrščanja počasneje narašča s številom

razredov. V primeru nevronskih mrež je vektor značilk kar vhod nevronske mreže. Končni rezultat prepoznavanja je oznaka razreda (znak), v katerega je vzorec razvrščen (Nadler, 1993).

3.3.5 Končna obdelava

Poleg uporabnikovega preverjanja rezultata prepoznavanja, sistemi vsebujejo različne metode avtomatične verifikacije in popravljanja rezultata. Prva možnost je preverjanje obstoja besede v slovarju (*angl. spell-checking*), ki je v nekaterih sistemih vključena v odkrivanje hiper ali podsegmentacije znaka. Nadalje lahko verifikacija poteka na osnovi sistema pravil. Ta pravila definirajo bolj in manj verjetna zaporedja znakov in ostale lastnosti jezika, kot je na primer v angleščini velika verjetnost črke “a”, če se znak pojavi samostojno. Seveda se tovrstne tehnike uporabljajo le v primeru nezanesljivo prepoznanega znaka. Zelo uporabno je barvno označevanje nezanesljivo prepoznanih znakov, kjer lahko z barvo označimo celo razred verjetnosti napačnega prepoznavanja. Po verifikaciji rezultata je potrebno rezultat še primerno oblikovati, glede na strukturo originala in nastavitev uporabnika. Nekatera orodja poleg rezultata vsebujejo tudi podatek o zanesljivosti prepoznavanja in omogočajo uporabniku, da s pragom zanesljivosti filtrira rezultat.

3.3.6 Uporaba rezultata

Na sliki 9 so označene tri možne uporabe rezultata prepoznavanja. Rezultat je lahko nov dokument v dokumentnem sistemu, kar je pomembno v primerih, ko prepoznavanje uporabimo za urejanje dokumenta. Izvoz rezultata prepoznavanja v podatkovno bazo uporabljamo predvsem v sistemih prepoznavanja obrazcev. Najpogosteje pa znotraj dokumentnih sistemov rezultat prepoznavanja uporabimo za indeksiranje vsebine dokumenta. Tako ustvarimo indeks, ki s pomočjo iskalnih orodij dokumentnega sistema omogoča iskanje po vsebini skeniranih dokumentov.

3.4 Pregled metod

3.4.1 Primerjalne metode

Primerjalne metode so znane kot tradicionalne metode prepoznavanja in so najbolj preproste. Razvrščanje se izvaja s primerjanjem grafične podobe znaka, oziroma

lastnosti (vrednosti značilk) z zbirko predlog oziroma vzorcev (*angl. template*). Rezultat, razvrstitev v razred in prireditev pomena znaku, določa mera ujemanja (*angl. similarity measure*) med znakom in vsemi razpoložljivimi vzorci. Tovrstne metode se obnesejo, kadar imamo opravka z omejenim naborom znakov oziroma pisav. Predvsem so uporabne za prepoznavanje določenega tipa dokumentov, kot je primer procesiranje in prepoznavanje številke čeka ali podatkov iz položnic, ki so zapisani s standardnim naborom znakov (Schantz, 1992).

Najbolj značilne metode v tej skupini so (Rogelj, 1998):

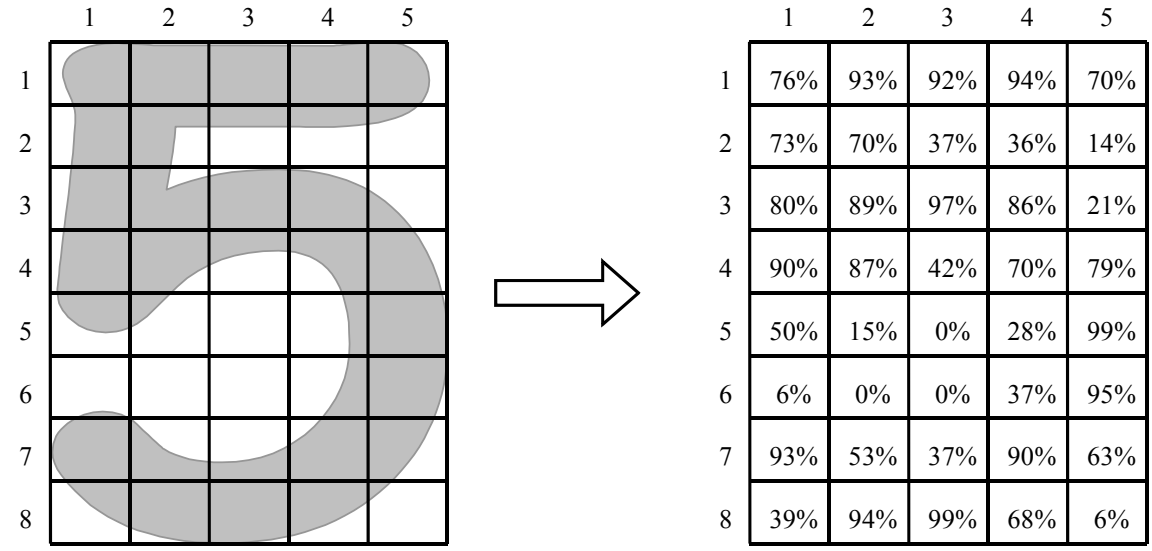
- Ujemanje s šablono (*angl. template matching*). Znake primerjamo v vseh slikovnih elementih (*pixel*) slike znakov (glej sliko 11).

Slika 11: Primer šablone črke "T" v primeru binarne primerjave

```
00000000000000000000000000000000
00000111111111111111111111100000
00000111111111111111111111100000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000001111110000000000000
00000000000000000000000000000000
```

- Projekcijski histogrami (*angl. projection histograms*). Primerjana lastnost je število točk, ki pripada znaku v posameznih horizontalnih in vertikalnih linijah slike znaka.
- Fourierovi deskriptorji (*angl. Fourier descriptors*). Znake ločujemo glede na velikost posameznih Fourierovih koeficientov funkcije, ki predstavlja rob znaka.
- Momenti. Znak se razpoznava po statističnih momentih njegove slike.
- Prilagojeni filtri (*matched filters*). Sliko znaka filtriramo z dvodimenzionalnimi prilagojenimi filtri. Vsak filter ustreza enemu razredu. Sliki znaka dodelimo tisto oznako, kot jo ima filter, ki da največji izhod.
- Ujemanje področij (*angl. zoning*). Sliko znaka razdelimo na več področij. Znak razpoznavamo glede na delež točk, ki na vsakem področju pripadajo znaku (glej sliko 12).

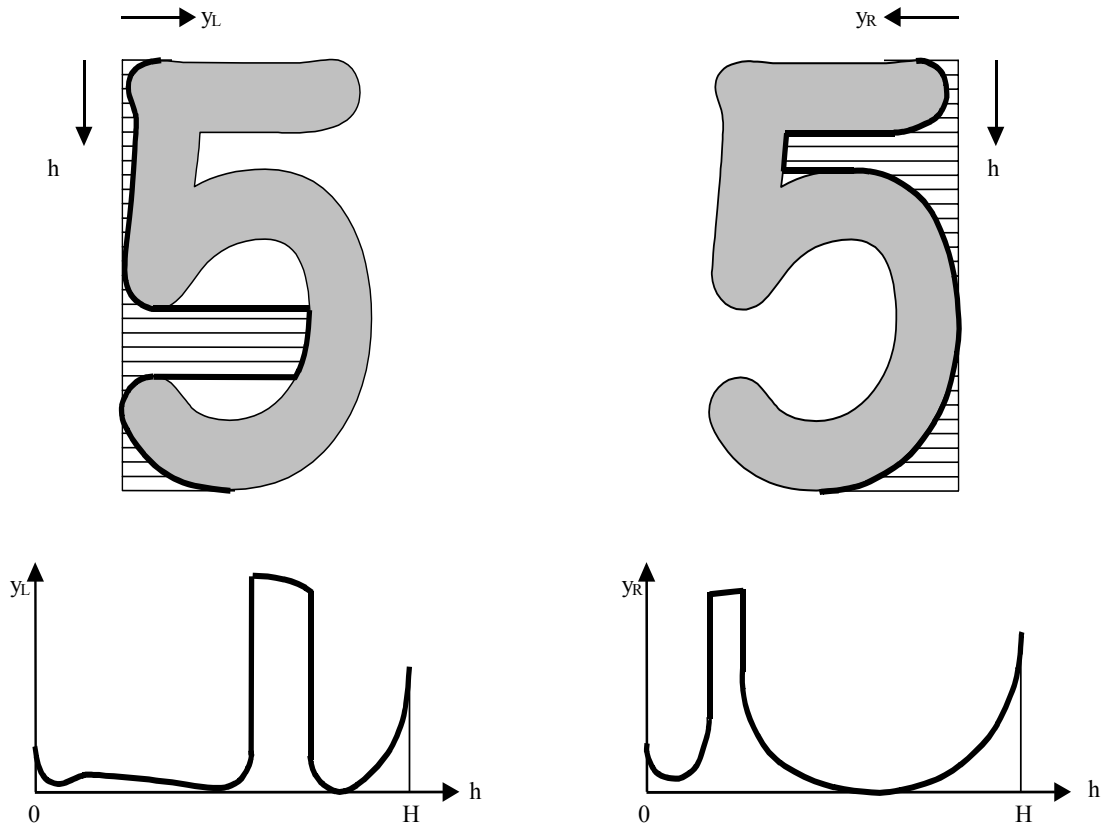
Slika 12: Metoda ujemanja področij



	1	2	3	4	5
1	76%	93%	92%	94%	70%
2	73%	70%	37%	36%	14%
3	80%	89%	97%	86%	21%
4	90%	87%	42%	70%	79%
5	50%	15%	0%	28%	99%
6	6%	0%	0%	37%	95%
7	93%	53%	37%	90%	63%
8	39%	94%	99%	68%	6%

- Robni profili (*angl. contour profiles*). Znake ločujemo glede na njihove ovojnice.

Slika 13: Metoda robnih profilov



3.4.2 Metode strukturne analize

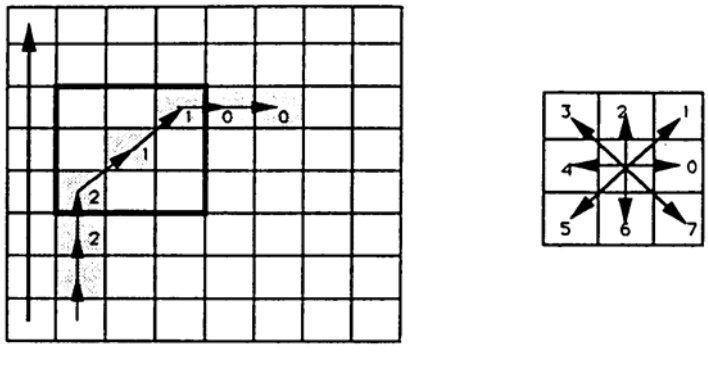
Metode strukturne analize uporabljajo strukturne opise in odločitvena pravila za razvrščanje znakov. Strukturni opisi so definirani kot navpične ali prečne črte, diagonalne črte, loki, konci črt, zamejena področja, konkavnost, konveksnost in podobno. Glavna prednost teh sistemov v primerjavi s primerjalnimi je mnogo večja neodvisnosti od nabora in velikosti znakov. Lahko ugotovimo, da so bližje človekovemu načinu prepoznavanja. Slaba lastnost je, da jih je težje realizirati. Bistveno je dobro izbrati značilnosti strukture znaka in te značilnosti tudi primerno ovrednotiti. V strukturni analizi je intuicija raziskovalca najbolj zanesljivo orožje pri reševanju problema (Mori, 1992).

Nekateri avtorji ločijo metode strukturne in topološke analize. Metode strukturne analize naj bi podajale rezultat na osnovi zbirke logičnih pravil, ki iščejo najmanjšo razliko med opisom lastnosti znaka in znanimi opisi. Metode topološke analize združujejo strukturno analizo in metode primerjanja in uporabljajo heuristične procedure in statistična primerjanja za razporeditev znakov v najboljši razred (Nadler, 1993).

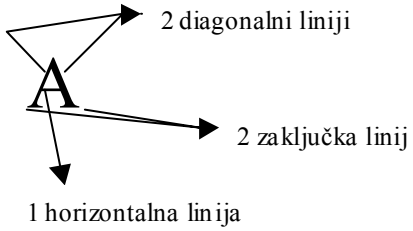
Možni so naslednji pristopi strukturne analize (Rogelj, 1998; Mori, 1992):

- Tokovno sledenje (*angl. stream following analysis*). Znak pregledujemo od enega do drugega roba in opazujemo kako se področje znaka razdružuje in združuje v posamezne tokove. Opis tokov (združevanja in razdruževanja) služi za prepoznavanje znakov.
- Sledenje robovom (*angl. contour following analysis*). Rob je krožno zaporedje točk v dvorazsežnem prostoru. Razdelimo ga na dele konveksnosti, konkavnosti in na notranje robove. Znak razpoznamo po relacijah med temi deli robov.
- Analiza ozadja (*angl. background analysis*). Ozadje znaka razdelimo po določenih kriterijih (na primer na konveksno lupino, konkavna področja in notranja področja), relacije med posameznimi področji pa nam dajo potrebno informacijo za prepoznavanje.
- Stanjšanje debeline (*angl. thinning line analysis*). Znake najprej stanjšamo na debelino enega slikovnega elementa, oziroma enega bita. Nato poiščemo značilne točke. To so točke, ki pripadajo vozliščem, križiščem, koncem segmentov in podobno. Povezave med značilnimi točkami opišemo s standardnimi segmenti, kot so ravna črta in različne krivine. Značilne točke in njihove medsebojne povezave tvorijo verižno kodo (*angl. chain encoding*), kot je prikazano na primeru Freemanove kode (glej sliko 14).
- Razčlemba celote (*angl. bulk decomposition*). Znak najprej razčlenimo na posamezne segmente, ki jih ločeno opišemo. Naveden je primer za znak “A” (glej sliko 15).

Slika 14: Primer verižnega kodiranja (levo) in pravil kodiranja (desno)



Slika 15: Primer opisa z metodo razčlembe celote



3.4.3 Metode prepoznavanja s tehnologijo nevronskih mrež

Razvoj sistemov za prepoznavanje znakov je v veliki meri pospešila uporaba tehnologije nevronskih mrež (NN – *neural netwroks*). Le-te namreč omogočajo akumuliranje znanja o strukturi dokumentov in že prepoznanih tipih pisav. Omogočajo torej primerjavo trenutnega dokumenta z znanjem, ki je bilo pridobljeno s preteklim prepoznavanjem. Tudi sama identifikacija dokumenta se lahko izvede s pomočjo karakterističnih podatkov, kot so fizična velikost, glava in podobno. Tovrstne metode so torej alternativa sistemom na osnovi pravil, kamor prištevamo prej omenjene metode primerjanja in strukturne analize. Predvsem so uporabne pri procesiraju “podobnih” dokumentov, kjer se ponavlja nabor pisav in segmentacija ter modeliranje dokumenta. Lahko ugotovimo, da gre za razumevanje dokumenta. Sistemi z metodami nevronskih mrež dosegajo dosti boljše rezultate pri obdelavi dokumentov, kjer je kvaliteta podobe znakov slaba. Ironično pa analize kažejo, da so pri visoki kvaliteti natančnejše in predvsem hitrejše metode strukturne analize. Metode nevronskih mrež so prevladujoče v sistemih prepoznavanja ročnih pisav, torej v ICR sistemih, ki so osnova sistemov za prepoznavanje obrazcev (Wang, 1993).

Nevronske mreže izkoriščajo izkušnje iz preteklih prepoznavanj. Pomemben je proces učenja, ki omogoča izdelavo klasifikacijskih procedur. Zmožnost učenja je osnovna značilnost inteligence. Učenje v sklopu nevronskih mrež pomeni posodabljanje mrežne arhitekture v smislu odstranitve ali dodajanja povezav in spreminjanje povezovalnih uteži v smeri reševanja določene naloge. Gre za neke vrste avtomatično učenje iz primerov, namesto da bi se postavila zbirka določenih pravil s strani človeških ekspertov. To je ena glavnih prednosti tovrstnega pristopa pri reševanju problemov prepoznavanja vzorcev. Načeloma se uspešnost klasifikacije povečuje s številom znakov, ki gredo skozi proces učenja (Jain, 1996).

OCR sistemi so ena najuspešnejših implementacij nevronskih mrež. Predvsem odkritje algoritma z vzvratnim učenjem je povzdignilo interes za uporabo tovrstnih tehnik v procesih prepoznavanja znakov. Zakaj? Prepoznavanje ogromnega števila variacij ročnih pisav predstavlja nerešljiv problem za sekvenčno programiranje. Analize kažejo, da se ročna pisava določenega posameznika spreminja z leti, utrujenostjo, odvisna je celo od tega ali je tekst napisan zjutraj ali popoldan. Sekvenčni algoritem prepoznavanja tako velikega števila vhodnih možnosti bi bil nefleksibilen in s tem omejen. Paralelnost in operatibilnost z nedoslednimi in nenatančno določenimi podatki je v tem primeru velika prednost nevronskih mrež. Trenutno so metode strukturne analize do neke mere še vedno prevladujoče metode omnifont prepoznavanja, predvsem zaradi hitrosti. Toda tudi na tem področju trend razvoja kaže na vse večjo vlogo nevronskih sistemov. Za potrebe prepoznavanja znakov je učenje nevronske mreže proces treninga, v katerem za učne vzorce uporabimo veliko število znakov. S tem si nevronski sistem s pomočjo učenja “zgradi” interno prezentacijo znaka, ki jo uporabi v procesu razvrščanja. To je kontrasten pristop v primerjavi s konvencionalnim strukturnim pristopom, kjer morajo biti pravila opisa strukture in na osnovi tega klasifikacije, natančno definirana s strani razvijalca sistema. Ko je proces učenja končan, imamo zgrajen sistem, ki ob primerni arhitekturi ne razpozna le znake, na katerih je bil treniran, ampak tudi prvič videne znake (Jain, 1996; Wang 1993).

Na splošno lahko pozicioniramo nevronske sisteme znotraj OCR sistemov na dva načina, oziroma v dveh shemah (Jain, 1996):

- kot razvrščevalnik značilk dobljenih s strani ekstraktorja oblike
- kot enotni sistem izbire značilk in razvrščanja le teh

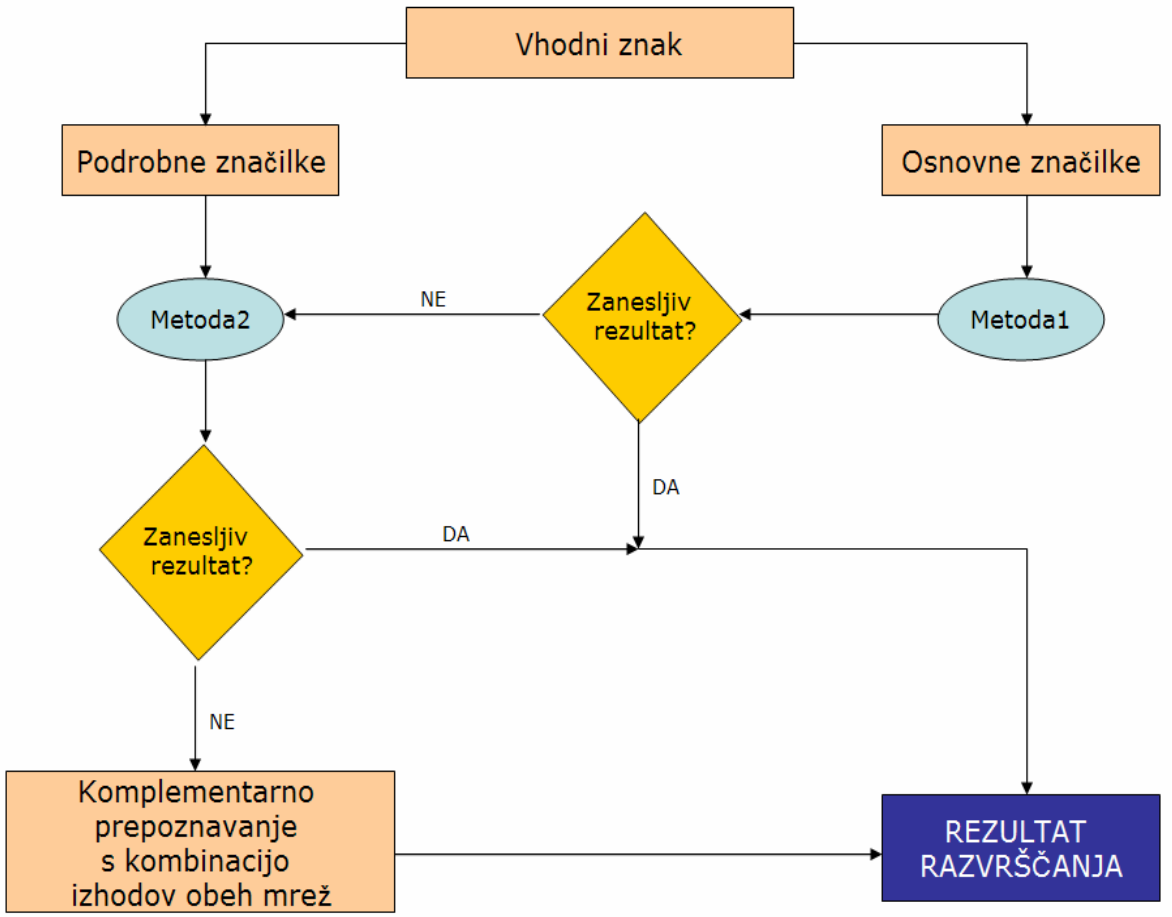
V drugem primeru sta torej oba procesa prepoznavanja integrirana in učenje poteka simultano v smeri optimalnih rezultatov razvrščanja.

3.4.4 Večnivojsko procesiranje

Večnivojsko procesiranje je sprožila potreba po večji hitrosti prepoznavanja in analiza načina branja pri človeku. Raziskave so pokazale, da naj bi človek ob branju preklapljal med načini hitrega in počasnega branja. Če je tekst čist in lepo berljiv, potem je proces branja hiter. Ko pa pride do slabe kvalitete tiska ali ob nekaterih dvomljivih in hitro zamenljivih znakih (kot so naprimer “rn” in “m”, “G” in “6”,itd.), je potrebna večja pazljivost in branje se upočasni.

Ta tehnika se lahko vpelje v sistem prepoznavanja znakov. Arhitektura sistema naj vsebuje dve metodi prepoznavanja. Prva ima manj vhodov (značilk znaka) in je s tem enostavnejša, s čimer pridobimo na preprostosti in hitrosti, vendar pa zajame manj podrobnosti znaka. Je dovolj dobra za razvrščevanje večine čistih znakov. Druga je bolj sofisticirana in počasnejša, vendar ima zaradi upoštevanja več značilnosti znaka višjo natančnost prepoznavanja v primeru dvoumnih znakov (glej sliko 16).

Slika 16: Arhitektura večnivojskega prepoznavanja



Osnovno prepoznavanje torej izvaja METODA1, ki prepozna nedvoumne znake. V primeru dvoumnosti METODA1 znak zavrne in ga preda v prepoznavanje METODI2. Zato tudi postavimo prag zavrnitve znaka na METODI1 nekoliko višje. Če tudi METODA2 ne da dovolj zanesljivega rezultata, lahko s kombinacijo rezultatov obeh mrež še vedno izračunamo verjetnostno najbolj pravilni znak.

Tovrstne arhitekture prepoznavanja znakov predvidevajo zadovoljive rezultate (tako v smislu natančnosti kot hitrosti), če 80 do 85 odstotkov znakov prepozna METODA1. Na ta način si lahko privoščimo računsko potratno arhitekturo METODE2, kar pa zagotavlja natančnejše rezutate. Obe metodi naj bi bili komplementarni pri napakah. Izdelava večnivojskih sistemov je kompliciran proces v smislu izbire arhitekture metod in njih kombinacije ter zahteva veliko eksperimentalnega dela (Nadler, 1993).

3.4.5 Opis metod orodja, ki ga uporablja IMiS/ICR modul

V nadaljevanju bomo opisali sistem prepoznavanja obrazcev IMiS/ICR, ki bo uporabljen v praktičnem primeru v zadnjem poglavju. Sam sem vodil razvoj produkta IMiS/ICR. V razvojni ekipi smo se odločili, da ne razvijamo lastnih algoritmov prepoznavanja znakov, temveč smo uporabili orodje FormReader podjetja ABBYY Software House iz Rusije, ki se je najbolje izkazalo ob analizi natančnosti in možnosti integracije. To orodje bom predstavil kot praktičen primer metod in algoritmov prepoznavanja znakov.

Sistem optičnega prepoznavanja znakov je hibridne narave. Lahko govorimo o vrsti prej omenjenega večnivojskega procesiranja. Gre namreč za uporabo več modulov prepoznavanja.

3.4.5.1 Glavni modul

Glavni in najkompleksnejši modul prepoznavanja uporablja sistem gradbenih ali sestavnih vzorcev, ki popolnoma opisujejo vsak razred. Vsak znak se namreč predstavi kot niz osnovnih sestavnih delov, kot so odseki, loki, krogi in točke, med seboj povezani s prostorskimi relacijami. Vsak razred opisujejo njegovi sestavni elementi in njihove medsebojne relacije. Kot je opisano v nadaljevanju, modul dobi podatek o predvidenem izhodu oziroma spisek predvidenih izhodov. To pomeni, da se potem iščejo sestavni deli predvidenega izhoda na objektu prepoznavanja, in ne gre več za tradicionalni postopek strukturne analize objekta in iskanja korelacij v bazi znanih razredov. Analiza temelji na tem, *kaj naj bi bilo* videno na

objektu prepoznavanja in ne *kaj je* videno. Tak pristop naj bi povečal natančnost prepoznavanja. Seveda je za uspešnost prepoznavanja bistvenega pomena podrobna definicija osnovnih delov in torej same strukturne analize znaka, kar je pa poslovna skrivnost proizvajalca.

Za opisan postopek prepoznavanja je bistvenega pomena dobro definirati posamezne sestavne dele znaka in njihove relacije. Relacije imajo obliko “mehke” logike (*angl. fuzzy logic*). Večina jih je opisanih kot preprosta kontrola, če vrednost pripada določenemu območju znotraj mehkih mej. Atributi tovrstnih izrazov so preproste lastnosti sestavnih delov, kot npr. dolžine, koti ali koordinate karakterističnih točk. Kot rezultat relacijskih primerjav, ki je zmnožek primerjav vseh relacij, dobimo vrednost v območju [0..1], kjer vrednost 0 pomeni popolno neujemanje, vrednost 1 pa popolno ujemanje.

3.4.5.2 Dodatna prepoznavalnika

Poleg glavnega modula sistem vsebuje še dva preprostejša prepoznavalnika. Oba bazirata na tradicionalnih metodah definicije značilk in izračuna mere ujemanja. S pomočjo teh dveh modulov se hitro določi rangiran spisek hipotez, kateri naj bi bili možni izhodi prepoznavanja. Ta spisek se uporabi kot filter v glavni modul, saj zmanjša iskanje sestavnih delov znaka. S tem se zelo poveča hitrost prepoznavanja, kot je bilo omenjeno že pri opisu večnivojskega procesiranja. Zaradi velikega števila izhodov, ki so možni pri prepoznavanju alfanumeričnih znakov, je vhod v glavni modul omejen s spiskom desetih najboljših hipotez. S tem omejimo prepoznavanje glavnega modula in povečamo hitrost. Kreiranje spiska 10-ih najboljših hipotez se izvede s pomočjo dveh komplementarnih prepoznavalnikov in komplementarnost omogoča visoko stopnjo uspešnosti.

Prvi dodatni prepoznavalnik prepoznava s pomočjo vektorizirane slike znaka. Uporablja 224 značilk. Vektorizirana slika znaka je razdeljena v 3X3 mrežo enakih kvadratov. Značilke so izvedene na osnovi končnih in vozliščnih točk, številom črno-belih prehodov v vertikalnih in horizontalnih linijah, številom in dolžino določenih potez in podobno. Izhodni vektor značilk ima dimenzijo 224X1 in vsebuje vrednosti ± 1. V obdobju učenja se izračuna vektor *p* razreda kot povprečje vseh vektorjev znakov istega razreda. Drugi enostavni prepoznavalnik (*angl. raster classifier*) pa za vhod uporablja sliko znaka standardne velikosti predstavljeno kot mrežo slikovnih elementov dimenzije 14X14 z vektorjem dimenzije 196X1. Beli slikovni elementi so kodirani z –1, črni pa z 1.

Dodatno so implementirane tudi linearne ločitvene funkcije (*angl.* LDF – *linear discriminant functions*), ki ločujejo “podobne” znake, kot je npr. “H” in “N”. To je izvedeno z značilkami, ki se med temi znaki najbolj razlikujejo (Anisimovich, 2001; ABBYY, 2004).

4. TEHNOLOGIJA PREPOZNAVANJA OBRAZCEV

Obrazcev je v vsakdanjem življenju veliko, računalniki so pa nepogrešljivi pri obdelavi podatkov in dostopu do informacij. Vmes je potreben prenos podatkov iz obrazca v informacijski sistem. Možnosti sta dve: ročen vnos ali uporaba tehnologije prepoznavanja obrazcev. Informacijski sistemi prepoznavanja obrazcev se zadnja leta vse bolj uveljavljajo. Trg narekuje hiter razvoj novih in še uspešnejših algoritmov prepoznavanja pisave in dinamičnih obrazcev, saj so uspešni projekti implementacije spremenili pogled na tehnologijo prepoznavanja obrazcev kot tehnologijo sedanjosti in ne več prihodnosti. Zaradi vse večjega povpraševanja in konkurence so se močno znižali stroški vpeljave, s čimer se je zmanjšal stroškovni rizik vpeljave in skrajšal se je čas povrnitve investicije. Pri tem je pomembno poudariti, da prej ni bilo ustvarjene nujno potrebne informacijske osnove za vpeljavo prepoznavanja obrazcev. Najprej so se namreč morali razviti in dokazati sistemi za skeniranje in elektronsko arhiviranje dokumentov, pa sistemi za avtomatsko kroženje dokumentov, torej dokumentni sistemi. Zaradi procesne potratnosti OCR in ICR algoritmov procesorji niso bili dovolj zmogljivi, napredovati je morala strojna oprema. Danes postaja vpeljava tehnologije prepoznavanja obrazcev v poslovnih procesih z veliko obrazci nujna.

4.1 Inteligentno prepoznavanje znakov ali ICR

Inteligentno prepoznavanje znakov (Intelligent character recognition, ICR) je eden fundamentov prepoznavanja obrazcev, saj so le ti pretežno izpolnjeni ročno. Ta termin se je prvič pojavil na začetku osemdesetih prejšnjega stoletja, ko je podjetje Kurzweil želelo diferencirati svoj OCR produkt od konkurence. Trdili so, da so v svojem produktu dodali inteligenco konvencionalnem, s pravili definiranem OCR-u, saj so omogočali, da uporabnik uči algoritem prepoznavati znake, pri katerih dela napake. V poznih osemdesetih so postali sistemi prepoznavanja pisave primerni za uporabo v produkcijskih sistemih in akronim ICR se je ponovno pojavil. Ti sistemi namreč za razliko od konvencionalnih OCR algoritmov, ki pretežno uporabljajo primerjalne metode in metode strukturne analize, bazirajo na metodah nevronskih mrež, katerih značilnost je učenje. Tako so proizvajalci skladno s Kurzweilovim pojmovanjem, začeli z akronimom ICR označevati sisteme, ki znajo

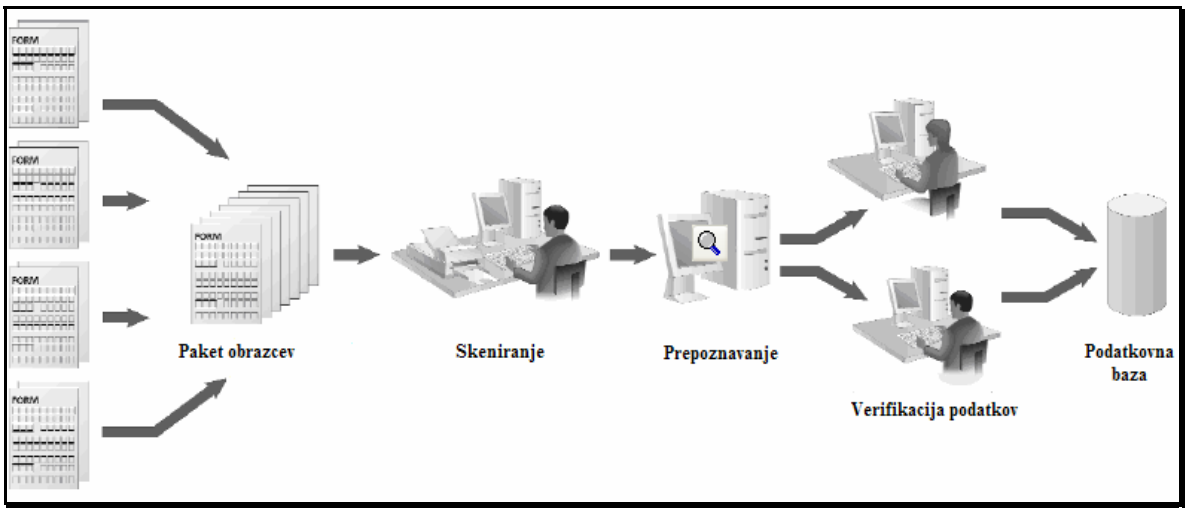
prepoznavati ročno pisavo. S tem so si »prilepili« oznako kakovosti, ki jih je diferencirala od OCR sistemov, ki so sposobni prepoznavati le tiskane znake. Prve implementacije so ICR sistemi doživeli pri procesiranju čekov, z razvojem dokumentnih sistemov se je pa hitro začelo razvijati tudi procesiranje obrazcev in s tem uporaba ICR tehnologije. Danes velik del industrije posplošeno uporablja termin ICR kar za tehnologijo prepoznavanja obrazcev (Gingrande, 2000; McCharty, 1994; Gingrande, 2003).

4.2 Postopek prepoznavanja obrazcev

Osnovna naloga sistemov prepoznavanja obrazcev je prepoznati tip obrazca in prepoznati ter izvoziti vsebino aktivnega dela. Postopek lahko razdelimo na 8 korakov, kot je prikazano na sliki 17 (Spencer, 1996; Gingrande, 2000, 2003; KOFAX):

1. skeniranje obrazca
2. analiza in predobdelava slike
3. odstranitev ozadja in pasivnih podatkov obrazca
4. segmentacija polj
5. segmentacija znakov
6. klasifikacija znakov
7. obdelava rezultata
8. verifikacija in ročno popravljanje napak

Slika 17: Proces prepoznavanja obrazcev



Pred skeniranjem je potrebno pripraviti obrazce. Priprava, in morebitno sortiranje obrazcev v veliko primerih ni časovno zanemarljivo opravilo. Skeniranje obrazcev in vpliv skeniranja smo že omenili. Kvaliteta skeniranja je še posebej pomembna,

saj zelo vpliva na kvaliteto segmentacije vsebine, pravtako pa tudi na kvaliteto prepoznavanja. Zato je potrebno zagotoviti kontrolo kvalitete skeniranih obrazcev in veliko testiranja posvetiti nastavitvam parametrov skeniranja.

V fazi predobdelave slike se izvaja identifikacija obrazcev. V primeru zavrnitve obrazca je potrebno omogočiti kontrolo obrazca s strani operaterja in morebitno ponovno skeniranje. Poleg postopkov čiščenja, prepoznavanja orientacije in avtomatične poravnave, se v fazi predobdelave slike pojavi še proces identifikacije tipa obrazca, glede na izdelane predloge. Obrazce, katerim ne moremo določiti tipa oziroma predloge, je potrebno izločiti iz procesa in jih ponovno skenirati ali ročno obdelati.

Odstranjevanje ozadja obrazca je proces, ki pred segmentacijo obrazca omogoča odstranitev vseh delov, ki so t.i. pasivni del obrazca. Metode odstranjevanja pasivnega dela so različne, najlažje to dosežemo z uporabo t.i. »drop out« barve, s katero tiskamo obrazec, kar pomeni, da jo skener sploh ne zazna. Tako obrazec lahko vsebuje razne kvadratke ali kakšne druge like, ki omogočajo lažji vpis podatkov, v digitalni obliki pa tega potem ni, saj bi le motilo proces prepoznavanja in slabšalo rezultat. Odstranimo tudi razne linije, statični tekst, slike, logotipe, itd. Pri odstranitvi pasivnih podatkov včasih poškodujemo aktivne podatke. S posebnimi postopki obdelave slike lahko podatke obnovimo pred prepoznavanjem.

Segmentacija polj pomeni lociranje polj in identifikacijo tipa aktivnih podatkov glede na tip obrazca in predlogo. V tem koraku ima zelo pomembno vlogo oblika obrazca, saj je potrebno ravno proces segmentacije polj ob oblikovanju obrazca veliko vzporedno testirati in prilagoditi izbranem algoritmu. Če bi bilo skeniranje konstantno in zamikov ne bi bilo in če bi bila kvaliteta izpolnjenih obrazcev enaka, potem je lokacijo polj enostavno določiti z absolutnimi pozicijami na obrazcu. Žal pa v realnih primerih temu ni tako in zato je potrebno obliko obrazca prirediti orodju, s katerim prepoznavamo. V večini primerov obrazec opremimo s posebnimi markerji ali lokatorji, s pomočjo katerih poteka segmentacija in lociranje polja glede na relativno pozicijo in oddaljenost polja od lokatorjev.

Segmentacijo znaka smo opisali ob predstavitvi OCR tehnologije in jo v procesu prepoznavanja obrazcev lahko še izboljšamo z natančno definicijo vsebine polj, kot je npr. fiksno število znakov ali pa z obliko polja, ki naj vsebuje ločene kvadratke za vpis posameznih znakov. S tem prisilimo izpolnjevalca, da ločuje znake, zmanjšamo možnost prekrivanja znakov in izboljšamo rezultate segmentacije znakov.

Proces prepoznavanja oziroma klasifikacije znaka ponavadi ne da le rezultata, temveč tudi podatek o zaupanju v rezultat ter morebitne alternativne rezultate. Ti dodatni podatki so zelo pomembni predvsem v kasnejšem procesu verifikacije, saj lahko za znake z manjšim zaupanjem v rezultat od določenega praga zahtevamo potrditev operaterja, pri čemer mu ponudimo tudi alternativne rezultate. Določitev praga zaupanja rezultatu mora biti izvedena s pomočjo podrobnega testiranja na vzorčni množici realno nastalih obrazcev. Seveda to ne pomeni, da so vsi znaki pod pragom nepravilno prepoznani, in tudi ne pomeni, da so vsi znaki nad pragom pravilno prepoznani. Gre predvsem zato, da je prag določen tako, da so za uporabnika rezultati dovolj dobri in uporabni. Morebitne napake identificiramo in odpravimo tudi v kasnejših korakih procesa.

Obdelava rezultata pomeni validacijo prepoznanih podatkov. Veliko obrazcev vsebuje pravila, ki povezujejo vrednosti posameznih polj na obrazcu in s tem omogočajo izdelavo logičnih kontrol, ki identificirajo napake. V tej fazi so pomembna tudi slovnična pravila, slovarji, povezave na obstoječe podatke v podatkovnih bazah in podobno. Tako lahko pridemo do informacije o kvaliteti prepoznanih podatkov, ki nam usmerja nadaljno obdelavo obrazca, katerega podatke izvozimo ali pa pošljemo v fazo verifikacije in popravljanja rezultatov.

Proces verifikacije in popravljanja napak ni neposredno del tehnologije prepoznavanja obrazcev, saj gre za delo operaterja. Vendar je potrebno opozoriti, da predstavitev podatkov s strani programske opreme in način popravljanja napak lahko zelo vplivata na hitrost in natančnost verifikacije in s tem celotnega procesa, saj je postopek identifikacije in odprave napake ozko grlo procesa prepoznavanja obrazcev. Temu delu je zato potrebno posvetiti velik del testiranja in pilotskega dela projekta. Časovna optimizacija popravljanja napačno prepoznanih podatkov ima največji vpliv na hitrost celotnega procesa. Potrebno je razviti karseda ergonomičen sistem idenitifikacije in načina odprave napake. Operaterja je potrebno čim bolj usmeriti na lokacijo napake. V primeru napake v polju z veliko znaki je včasih celo bolje ponovno vnesti celotno vrednost, saj je iskanje napačno prepoznanega znaka v veliki množici znakov lahko časovno neučinkovito.

Ugotovimo lahko, da je v bistvu le korak 6 neposredno povezan s tehnologijo optičnega prepoznavanja znaka, ostali pomenijo pripravo na prepoznavanje ali pa interpretacijo rezultata. Vsak korak v procesu pa vsebuje možnosti napak in zato je fascinatno, da lahko ICR sistemi dosegajo ali celo presegajo natančnost ročnega vnosa podatkov.

Tipe obrazcev definiramo s predlogami vnaprej. Izdelamo predlogo, s katero definiramo lokacije in vrste polj ter morebitne vzorce, ki omogočijo hitro

prepoznavanje tipa obrazca in lažje lociranje polja v primeru zamika pri skeniranju. Najlažja in priporočena metoda identifikacije obrazca je označitev tipa obrazca s pomočjo črtne kode. V zadnjem času se kot nadgradnja tovrstnim statičnim sistemom prepoznavanja tipa obrazca, pojavljajo kompleksnejši sistemi prepoznavanja dinamičnih obrazcev. To pomeni, da obrazcev ne definiramo vnaprej s predlogami in tudi ne poznamo točne lokacije posameznega polja, ki nas zanima, temveč algoritmi po obrazcu iščejo značilke, vzorce ali besede, ob katerih so željeni podatki. Gre torej za dinamično definicijo strukture obrazca. Največ raziskav se opravi na nivoju dinamičnega prepoznavanja prejetih računov, ki so med seboj različni, vendar imajo skupne elemente. Tako se na primer išče beseda TOTAL, ob kateri je ponavadi izpisan znesek računa ali pa INVOICE za številko računa in podobno. Prepoznavamo torej obrazce, katerih vsebino poznamo, a ne vemo natančno kje na obrazcu se nahaja. Današnje metode analize dinamičnih obrazcev uporabljajo tehnologijo nevronskih mrež ter zbirke sofisticiranih poslovnih pravil. Uspešni so hibridni sistemi z mešanico prepoznavanja tipa s pomočjo statičnih in dinamičnih obrazcev (Spencer, 2001a; Haverson, 2002).

Ko je polje podatka locirano, je potrebno prepoznati vsebino. Prepoznavanje vsebine oziroma vrednosti polja izvajamo s pomočjo tehnologij prepoznavanja ročne pisave (ICR), prepoznavanja tiskanih znakov (OCR), prepoznavanja označevanj (OMR) in prepoznavanja črtnih kod.

Zaradi samega postopka prepoznavanja obrazcev s prepoznavanjem tipa in vsebine, se je vpeljave tovrstnih tehnologij potrebno lotiti projektno, saj se ne da programske opreme le namestiti in pričakovati pozitivne rezultate. Potrebno je analizirati obrazce, jih izdelati oziroma spremeniti tako, da se najbolj prilegajo uporabljeni programski opremi oziroma algoritmom, ki jih vsebuje. Nato je potrebno najti optimalne nastavitve skeniranja, kar lahko dosežemo samo s poizkušanjem in testiranjem vzorčnih obrazcev. Definirati je potrebno čim več pravil in kontrol, ki omogočajo avtomatsko identifikacijo morebitnih napak prepoznavanja. Napačno prepoznane obrazce je potrebno usmeriti v proces verifikacije. Ko so podatki pravilno prepoznani, se podatki izvozijo v podatkovne baze in dokumentne sisteme, skenirane obrazce pa umestimo v elektronski arhiv.

4.3 Natančnost ICR sistemov

Natančnost ICR sistemov opredelimo z naslednjimi termini (Gingrande, 2000):

- **delež sprejetih** (*angl. acceptance rate*) izražamo kot odstotek sprejetih znakov ali polj, za katere sistem nedvoumno podaja rezultat, oziroma je zaupanje v pravilnost klasifikacije dovolj veliko

- **delež zavrnjenih** (*angl. rejection rate*) izražamo kot odstotek znakov ali polj, katere sistem zavrne, saj je nezaupanje v pravilnost klasifikacije dovolj veliko
- **delež pravilnih** (*angl. accuracy rate*) izražamo kot odstotek znakov ali polj, ki so bili s strani sistema pravilno prepoznani
- **delež nepravilnih** (*angl. substitution error rate*) izražamo kot odstotek znakov ali polj, ki so bili s strani sistema nepravilno prepoznani

Seštevek deležev sprejetih in zavrnjenih zajema vse znake in polja, torej 100 odstotkov. Delež sprejetih in zavrnjenih je v tesni povezavi s podatkom in pragom zaupanja v pravilnost klasifikacije znaka. Ta prag je potrebno določiti z obsežnim testiranjem. Kot sem že omenil, ta prag ne pomeni absolutne meje med pravilno in nepravilno prepoznanimi znaki, zato pri opredelitvi natančnosti uporabimo tudi delež pravilnih in nepravilnih. Delež sprejetih in zavrnjenih dobimo v fazi prepoznavanja znakov, delež pravilnih in nepravilnih pa v fazi verifikacije in popravljanja rezultatov, če nam narava procesa to sploh omogoča. Iz natančnosti na nivoju znakov lahko izračunamo natančnost na nivoju polja z n-to potenco natančnosti znaka, pri čemer je n enak predvidenem številu znakov v polju. Podobno lahko izračunamo tudi predvideno natančnost na nivoju obrazca, pri čemer je pa potreben podatek o uporabnosti polj, oziroma o vplivu napake polja na napako obrazca.

4.3.1 Napake v primerjavi z ročnim vnosom

Tudi če za vnos podatkov ne bi uporabili ICR sistemov, temveč profesionalne operaterje vnašalce, bi imeli določen delež napak in ta je ponavadi celo večji kot v primeru ICR sistemov. Seveda govorimo o deležu napak po verifikaciji rezultatov.

Sam vzrok in tip napake je v obeh primerih različen (Spencer, 1996):
- Najpogostejši vzrok napak ljudi je nepravilna izbira tipke na tipkovnici (*angl. transposition errors*), kar je ponavadi posledica utrujenosti vnašalca podatkov. ICR tovrstnih napak ne dela, saj je rezultat percepcije tudi dejanski rezultat prepoznavanja.
- Ljudje redko napačno prepoznajo znak, medtem ko je to edini vzrok napačno prepoznanega znaka pri ICR sistemih.
- Zelo pomembno je, da ICR sistemi v koraku prepoznavanja najprej izolirajo in potem klasificirajo znak, kar pomeni, da ne upoštevajo narave podatkov ali polja, v katerem se nahaja obravnavani znak. Tu imajo ljudje veliko prednost, saj z analizo oziroma percepcijo vsebine podatkov lahko zelo zmanjšajo napako klasifikacije.

4.3.2 Kako izboljšati natančnost

V fazi klasifikacije znakov je človek uspešnejši od računalnika. Predvsem z integriranjem narave in vsebine podatkov v proces interpretacije rezultatov lahko zelo zmanjšamo delež napak ICR sistemov in s tem do neke mere emuliramo človeški način prepoznavanja znaka. Poglejmo naslednje besedilo (pridobljeno kot šala v elektronskem sporočilu):

»The pweor of the hmuan mnid. Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a total mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter byistlef, but the wrod as a wlohe.

Amzanig huh?«

Kljub temu, da nobena beseda ni napisana slovnično pravilno, lahko človek besedilo pravilno prebere in razume. V fazi branja možgani avtomatično odpravljajo napake v tekstu s pomočjo vpliva vsebine na percepcijo znakov in besed. Če tovrstni princip uspemo integrirati v postopke vrednotenja in validacije ter tudi morebitne translacije vsebine polja, potem lahko odpravimo veliko napak avtomatično. Zato danes poznamo različne načine vstavljanja pravil in procedur verifikacije, ki omogočajo povečanje natančnosti, predvsem v smislu odkrivanja napačno prepoznanih znakov, ki so imeli dovolj veliko zaupanje v rezultat (Gingrande, 2000):

- **Analiza povezanosti znakov** (*angl. Context analysys*), ki vsebuje kakršnokoli informacijo o tem, kateri znak lahko pričakujemo na določenem mestu. Znotraj te skupine poznamo:
    - integracijo različnih slovničnih pravil in pogostosti pojavljanja znakov v določenih besedah.
    - vnaprej določena oblika polja v smislu alfanumerične sintakse, če so znaki na določenih mestih znotraj polja omejeni na številko, črko ali kako drugo zaključeno množico. Tako se lahko izognemo napakam zaradi podobnosti velikega »O« in ničle ali majhnega »l« in številke ena, itd.
    - uporaba slovarjev kot možnih vrednosti polj. Seveda tu ne gre za denimo SSKJ, ki je za tovrstne naloge v večini primerov neuporaben, temveč za vnaprej določene zaključene množice vrednosti polj, kot so lahko npr. šifranti pacientov v obrazcih v zdravstvu in podobno.
    - računanje paritete podatkov, kot npr. v poljih številk kreditnih kartic.

- **Procedure preverjanja veljavnosti podatkov** (*angl. Data validation routines*), ki preverjajo rezultate na osnovi predefiniranih standardov v smislu primernosti podatkov. Tovrstne procedure se uporabljajo tako v ICR sistemih, kot v primerih človeškega vnosa podatkov. Znotraj te skupine poznamo 3 osnovne tipe procedur:
    - primerjanje z obstoječimi podatki v bazi, pri čemer ne gre le za preverjanje absolutne pravilnosti, temveč tudi za procedure korekcije podatkov. Tako je npr. v poljih, ki vsebujejo naslove dovolj, da se ujema 85 odstotkov znakov, preostalih 15 odstotkov se pa popravi glede na vrednosti naslovov v bazi podatkov.
    - omejitev vrednosti znotraj določenega obsega. Naj omenim smiselnost omejitve polj, kjer je vpisana starost, ali pa preverjanje smiselnosti datumskih vrednosti, itd.
    - preverjanje odnosov med polji, kot na primer preverjanje vsote po stolpcih, kar zviša zaupanje v pravilnost prepoznavanja.

Zelo uspešen pristop k izboljševanju natančnosti je motiviranje in usmerjanje vnašalcev v smeri pravilnega in natančnega izpolnjevanja obrazcev. Predvsem je potrebno čim bolj specificirati obliko podatkov, kot npr. oblika zneska (nepisanje pik ali vejic,...), uporabo ameriške ali evropske številke 4, celo priporočiti tip pisala, itd. Tako so v Rusiji javno objavili in reklamirali natančno izpolnjevanje obrazcev za povračilo subvencij tako, da so obljubili hitrejše izplačevanje subvencij v primeru upoštevanja navodil pri izpolnjevanju, kar je pomenilo manj napak prepoznavanja. Prvo leto po uporabi tega pristopa se je število napak zmanjšalo za skoraj 50 odstotkov (ABBYY, 2001).

Seveda je kriterij natančnosti potrebno vezati na predmet prepoznavanja. V primeru prepoznavanja telefonskih številk nam 90 odstotna natančnost na nivoju znaka, kar je sicer zelo dober rezultat, ne pomaga kaj dosti, saj je nujno potrebna 100 odstotna natančnost. Tako lahko govorimo o kriteriju uporabnosti podatkov in v različnih primerih je le-ta vezan na različne natančnosti.

4.4 Izdelava obrazca

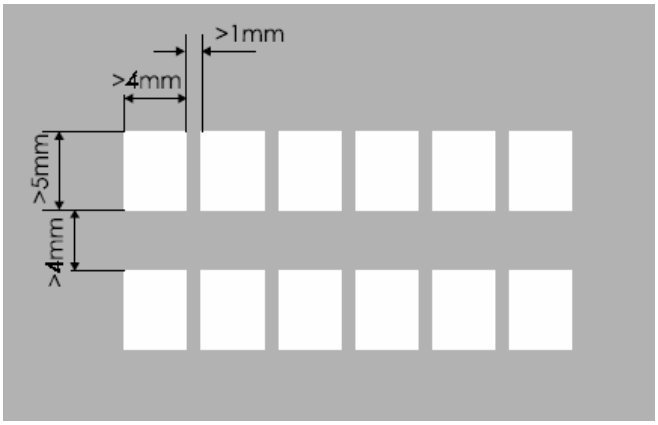
Zelo pomemben dejavnik, ki vpliva na uspešnost vpeljave ICR sistema, je oblika obrazca. Po mojem mnenju gre celo za najbolj pomemben dejavnik, kateremu je potrebno posvetiti veliko časa in testiranja, saj študije kažejo, da je kar 80 odstotkov natančnosti prepoznavanja odvisne od oblike obrazca (Spencer, 1996). Osnovni pogoj je možnost vplivanja na obliko obrazca. To žal velikokrat ni možno in v takih primerih je uspešnost vpeljave ICR sistema vprašljiva .

Med proizvajalci ICR programske opreme se je uveljavil izraz **prepoznavanju prijazen obrazec** (*angl. ICR-friendly form*, *Machine-readable form*). V nadaljevanju bomo podali splošne napotke za izdelavo tovrstnih obrazcev, podrobnosti so pa predvsem odvisne od samega proizvajalca. Veliko orodij omogoča izdelavo obrazcev kar znotraj svojih produktov, kar olajša izdelavo primernega obrazca in predloge zanj (ABBYY, 2003).

Osnovni namen je doseči ločenost znakov, kar omogoča lažjo izolacijo in segmentacijo znakov ter s tem preprečitev prekrivanja znakov in možnosti hipersegmentacije. Z upoštevanjem osnovnih napotkov lahko to lažje dosežemo. Poudariti je treba, da uporabnika ne moremo prisiliti v korektno izpolnjevanje, lahko pa ga motiviramo in usmerimo.

Polja naj vsebujejo separatorje znakov (*angl. input field place-holders*), ki naj jih skener s pomočjo »drop out« barve ignorira. Vsak proizvajalec orodij v navodilih natančneje opisuje tovrstne prijeme. Na sliki 18 so prikazani navodila za izdelavo separatorjev v primeru orodja ABBYY FormReader.

Slika 18: Primer navodila za definiranje polja na obrazcu



Vir: ABBYY Software House, 2004

Robovi obrazca naj bodo dovolj široki, tako da se tudi v primeru zamika pri skeniranju podatki ne izgubijo. Vsi produkti namreč v fazi predobdelave slike obrazca omogočajo avtomatsko poravnavo obrazca do 15° naklona, pri čemer v večini primerov uporabljajo izračun povprečnih kotov nagiba po vseh vrsticah in po potrebi nevtralizacijo tega kota. Poravnava naklonov večjih od 15° je nesmiselna, saj pride do prevelike deformacije obrazca, in v tem primeru ponavadi niso vsi podatki na zamaknjenem obrazcu.

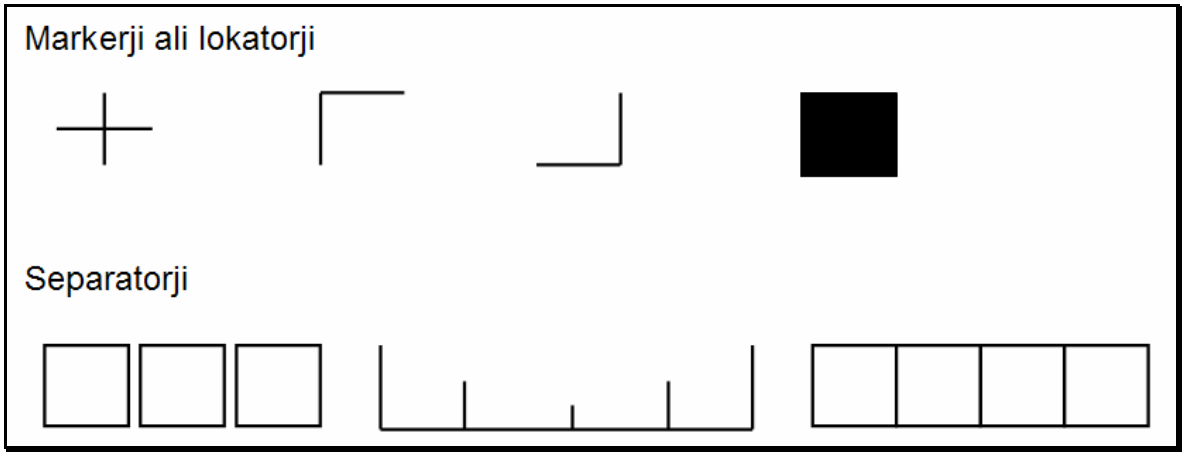
Podatki potrebni za identifikacijo obrazca naj bodo na vsakem obrazcu na istem mestu. S tem je omogočena hitra in zanesljiva identifikacija tipa obrazca.

Priporočena je uporaba črtne kode, sicer se v glavnem uporablja numerično polje, ali pa statični tekst in grafika.

Polja z aktivnimi podatki naj bodo med seboj dovolj narazen, kar lahko vidimo tudi na priporočilih na sliki 18. V predlogi namreč ponavadi določimo večje območje polj, kot je dejansko, saj vseh enakih slik in s tem lokacij v procesu skeniranja ne moremo doseči. Večja območja nam torej omogočijo korekcijo manjših napak segmentacije polj in zamikov na skeniranem obrazcu. Tudi labele in drugi grafični pripomočki, ki identificirajo polja, morajo biti iz istega razloga dovolj oddaljeni od polja. Enako velja za pasivne podatke, ki jih odstranjujemo s pomočjo orodij za odstranjevanje ozadja in pasivnih podatkov, saj sicer lahko pride do deformacije aktivnih podatkov, ki lahko vodijo v napačno klasifikacijo ali napako podsegmentacije znaka.

Obrazci naj vsebujejo posebne znake, markerje (*angl. position marker*), ki omogočajo natančno relativno lokacijo polj. Ti se sicer od proizvajalca do proizvajalca razlikujejo, vendar je osnovna ideja pri vseh enaka. Algoritmi segmentacije najprej iščejo te znake in v naslednji fazi lokacije polj izračunajo glede na lokacijo markerjev na skeniranem obrazcu. S tem nevtraliziramo zamike skeniranega obrazca. Pod tovrstne znake štejemo tudi ravne črte, ki omogočajo hitrejši izračun morebitnega naklona skeniranja. Liho število lokatorjev omogoča hitrejše ugotavljanje pravilne orientacije skeniranega obrazca. Na sliki 19 so prikazani najbolj pogosti grafični elementi obrazcev

Slika 19: Pogosti grafični elementi na obrazcih



To so osnovni napotki izdelave obrazcev. Naj veljajo tudi v primeru, ko se podatki iz obrazcev ročno vnašajo v sistem, saj tudi človeška percepcija preferira ločene znake. Osnovni pogoj izboljšanja natančnosti z obliko obrazca je možnost vpliva

na obliko obrazcev. Seveda je to možno le, ko gre za interne obrazce, oziroma ko obrazci nastajajo znotraj okolja, kjer se izvaja tudi procesiranje obrazcev. Taka so na primer podjetja, ki izvajajo ankete ali popise, ki so zelo dovzetna za spreminjanje obrazcev in vpeljavo sistemov za prepoznavanje obrazcev. S tem lahko učinkoviteje, hitreje in na dolgi rok ceneje zbirajo podatke, kar je osnovna dejavnost tovrstnih podjetij. Banke, zavarovalnice in druge finančne institucije so nekoliko manj fleksibilne organizacije, ki zelo veliko pretoka informacij in podatkov izvedejo prek obrazcev, in obrazci nastajajo znotraj teh organizacij. Te so se ob uvedbi prepoznavanja obrazcev v ZDA zelo upirale spreminjanju obrazcev. Namesto da bi sodelovale in upoštevale analize, ki so kazale na povečanje uspešnosti in s tem zmanjšanje stroškov prepoznavanja, so s svojo togostjo silile proizvajalce ICR sistemov naj prilagajajo algoritme obstoječim obrazcem. Rezultat so bila velika vlaganja na področju algoritmov analize slik in odstranjevanja ozadja, ki so pa prinesla relativno malo koristi in razočaranje. Tudi to je eden razlogov, da so se ICR sistemi uveljavili kasneje, kot bi se lahko. Šele s posameznimi uspešnimi implementacijami ICR sistemov, pri katerih so analize in končna poročila pokazala na veliko znižanje stroškov popravljanja in izboljšavo natančnosti s prilagajanjem obrazcev ICR orodjem, so spremenila poglede kupcev ICR sistemov. Tako je danes prva projektna naloga integratorja ICR sistema, da prepriča kupca v pomembnost prilagajanja obrazcev, najbolje kar s pilotskim projektom na vzorcih. Velik vpliv oblike bom pokazal tudi na praktičnem primeru v nadaljevanju naloge (Gingrande, 2000).

4.5 IMIS/ICR programska oprema kot primer sistema za prepoznavanje obrazcev

IMiS/ICR moduli so del družine produktov IMiS podjetja Imaging Systems d.o.o. iz Ljubljane. IMiS je podsistem za računalniško upodabljanje in arhiviranje dokumentov, ki ga integriramo v informacijski sistem, ki »potrebuje« funkcije skeniranja in elektronskega arhiva. IMiS je zasnovan modularno in IMiS/ICR moduli so namenjen prepoznavanju obrazcev.

IMiS/ICR ne uporablja lastnih algoritmov prepoznavanja obrazcev in znakov, temveč je integrirano orodje ABBYY FormReader SDK podjetja ABBYY Software House iz Rusije. IMiS/ICR sestavljata modula IMiS/ICR strežnik in IMiS/ICR odjemalec. Prvi je namenjen paketnemu procesiranju poskeniranih obrazcev, drugi pa verifikaciji in popravljanju rezultatov, pri čemer oba modula uporabljata API klice ABBYY FormReader produkta. Osnovna enota je paket (*angl. batch*) obrazcev. Znotraj paketa je definiran eden ali več tipov obrazcev, vsak tip ima

definirano predlogo, s postopki identifikacije pa skeniranemu obrazcu določimo tip in s tem predlogo.

4.5.1 Definiranje predlog obrazcev

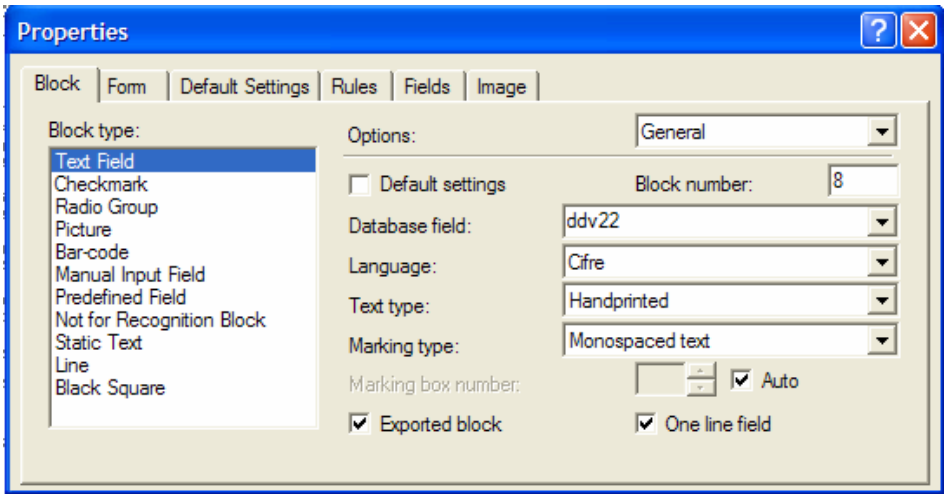
Predlogi obrazcev se definirajo s pomočjo produkata ABBYY template designer in tako definirane predloge uporabljata tako strežnik kot odjemalec. Najprej je na obrazcu potrebno definirati identifikacijske dele. V primeru orodja ABBYY FormReader so priporočeni markerji črni kvadratki, velikosti 4x4 mm do 7x7mm, oddaljeni morajo biti najmanj 3mm od kateregakoli drugega objekta. Optimalno naj jih bo 5, pri čemer naj štirje med sabo povezani tvorijo pravokotnik, peti naj se pa nahaja med dvema kot pomoč pri ugotavljanju orientacije. Priporočljivo je dodati horizontalne ali vertikalne črne črte, ki omogočajo natančnejšo nevtralizacijo zamika. V primeru podobnih obrazcev znotraj istega paketa je priporočljivo definirati še nekaj statičnega teksta na obrazcu, ki pripomore k razločevanju podobnih obrazcev različnih tipov. Primer tako izdelanega in definiranega obrazca je spremenjeni DDV-0 obrazec na sliki 37, stran 75. Nato v predlogi označimo lokacijo in lastnosti polj, ki jih prepoznavamo. Velikost označenega območja polja je potrebno dobiti s testiranjem, saj premajhno območje lahko povzroči izgubo podatka zaradi zamika skeniranja, preveliko območje pa lahko zajame podatke, ki ne spadajo v polje. Tip polja je lahko:

- navadno besedilo
- potrditveno polje (*angl. checkmark*)
- skupina izbirnih gumbov (*angl. radio group*)
- slika
- črtna koda
- polje ročnega vnosa
- preddefinirano polje (tu lahko zapišemo statistične podatke, kot so: številka strani, ime predloge, število napak, komentar, kreatorja, itd.)
- identifikacijsko polje

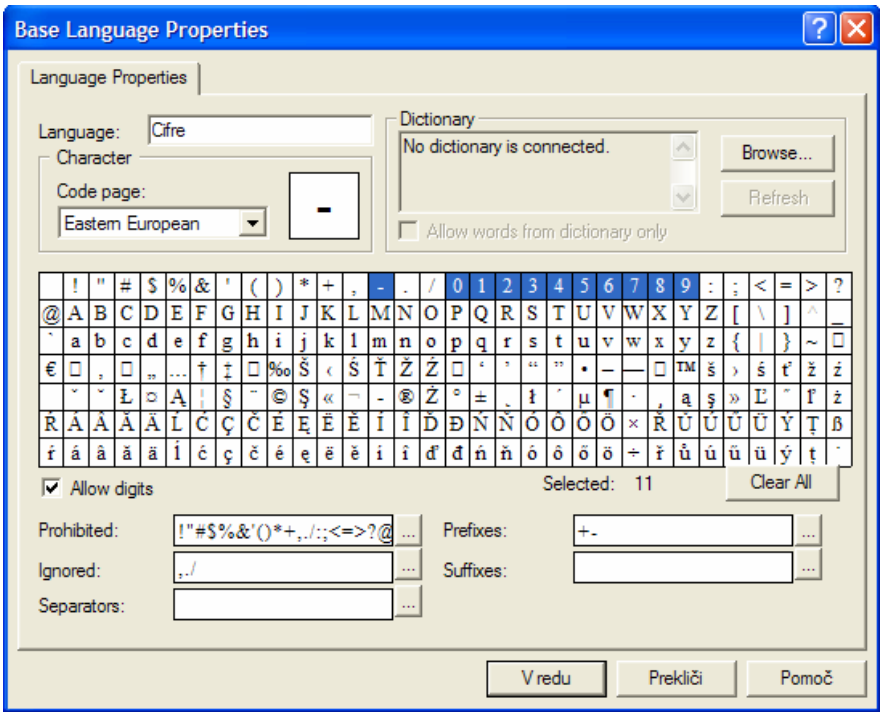
Vmesnik za definicijo lastnosti polja je prikazan na sliki 20. Nadalje lahko v poljih tekstovnega tipa določimo še nabor znakov (jezik), tip pisave (avtomatično, matrični tiskalnik, tipkalni stroj, ročna pisava ali mešan tip), obliko separatorjev znakov, število vrstic, ali dovolimo presledke, nivo označevanja napak, uporabo vzorca naučenih znakov in nivo verifikacije znakov ali polj. Zelo pomembna je pravilna definicija nabora znakov, ki lahko zelo vpliva na hitrost in natančnost prepoznavanja. Znotraj nabora moramo namreč zajeti le tiste znake, ki se lahko pojavijo znotraj polja, eksplicitno določiti znake, ki so prepovedani, itd. Vmesnik za določitev nabora znakov je prikazan na sliki 21. Pri črtnih kodah zaradi hitrejše

identifikacije določimo tip črtne kode, če je le-ta znan in konstanten na vseh obrazcih.

Slika 20: Vmesnik za določanje lastnosti polja



Slika 21: Vmesnik za določanje nabora znakov v jeziku prepoznavanja



Na nivoju obrazca znotraj predloge definiramo tudi enostavna pravila. Na voljo je več tipov pravil, naj jih nekaj naštejem:

- primerjava s podatkovno bazo
- izraz
- sestavljanje vrednosti polj
- normalizacija in preverjanje datumskih vrednosti

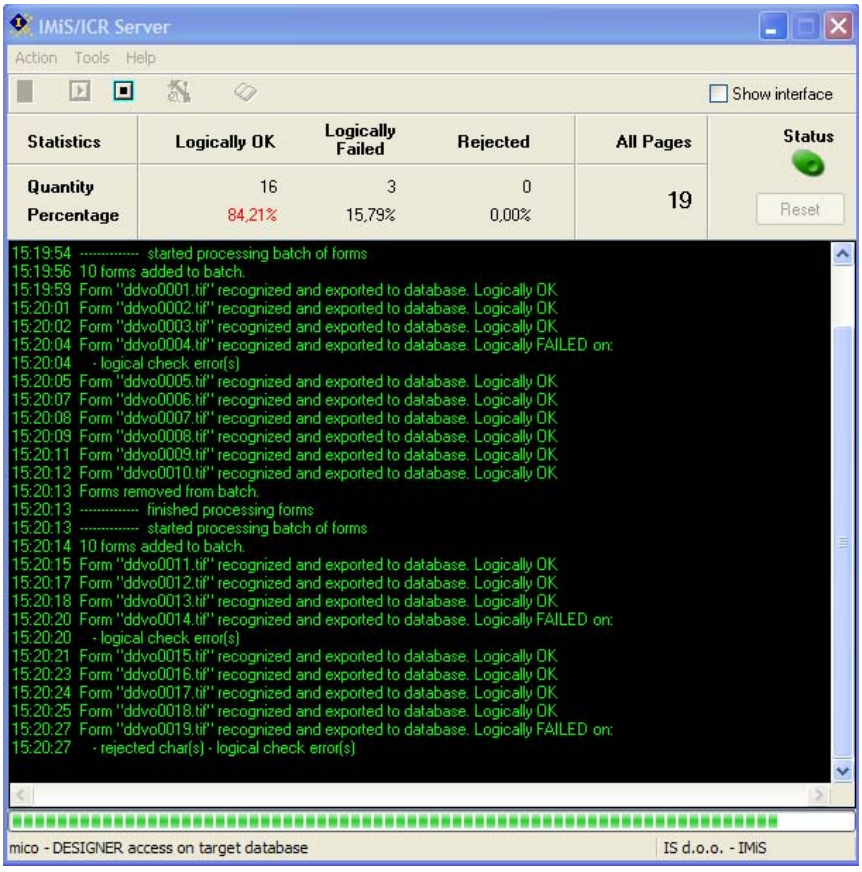
- zamenjava znakov
- itd.

Verjetno najbolj uporabna so pravila s pomočjo izrazov, kjer lahko določimo npr. fiksno število znakov nekega polja, preverjamo relacije med polji, itd.

4.5.2 IMiS/ICR strežnik

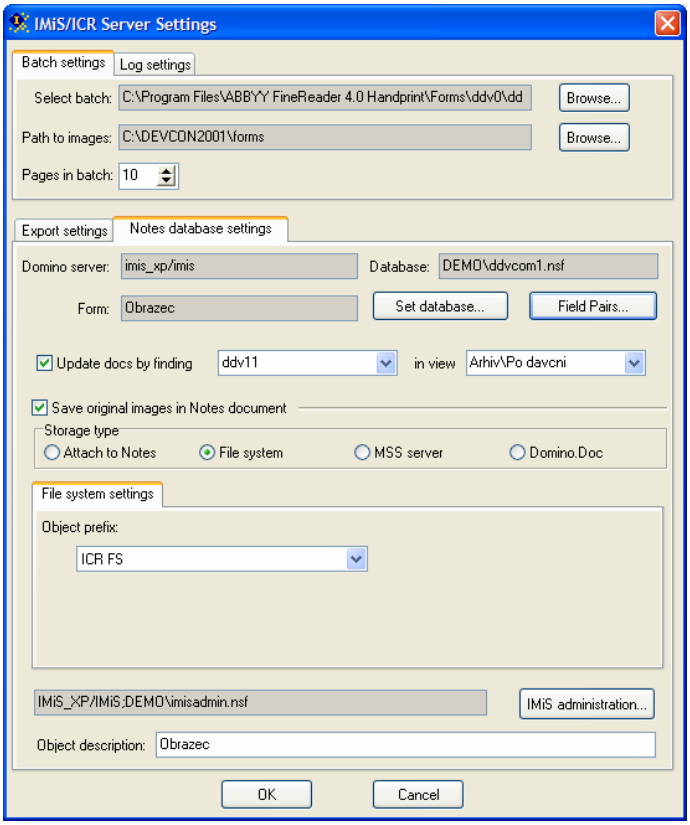
IMiS/ICR strežnik je namenjen paketnemu procesiranju obrazcev. Strežnik na določen časovni interval pregleduje lokacijo, kjer skenirne postaje odlagajo skenirane obrazce. V paketu po 10 obrazcev (kar je nastavljiva vrednost) jih obdela tako, da identificira tip obrazca, prepozna vsebino polj, preveri pravilnost podatkov in podatke izvozi. Kot vidimo na sliki 22, se v strežniški konzoli sproti izpisujejo podatki o prepoznanih obrazcih, o pravilnosti in tipu napak, vodi se tudi statistika natančnosti prepoznanih obrazcev. Podatki iz obrazcev se takoj izvozijo v ciljno bazo podatkov, pravtako se arhivirajo na arhivski strežnik dokumentnega sistema in so takoj na voljo uporabnikom.

Slika 22: IMiS/ICR strežnik



Na sliki 23 je prikazano okno z možnostmi IMiS/ICR strežnika. Določiti je potrebno datoteko z definiranimi predlogami obrazcev (»Select batch« polje), katero predhodno kreiramo s pomočjo produkta ABBYY FormReader template designer. Nato določimo pot do skeniranih obrazcev ter število obrazcev, ki jih obdelujemo v enem paketu. V spodnjem delu okna so prikazane nastavitve v primeru izvoza podatkov v IBM Domino okolje in pa nastavitve za arhiviranje skeniranih obrazcev. Arhiviranje obrazcev je zelo pomemben del, saj s tem omogočimo kasnejši hitri dostop do slike obrazcev ter omogočimo verifikacijo rezultatov z IMiS/ICR odjemalcem.

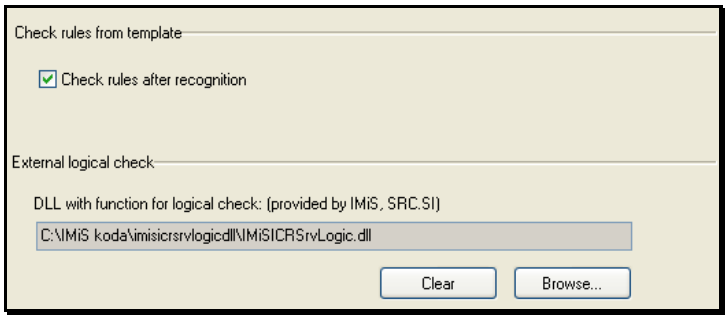
Slika 23: Vmesnik z nastavitvami IMiS/ICR strežnika



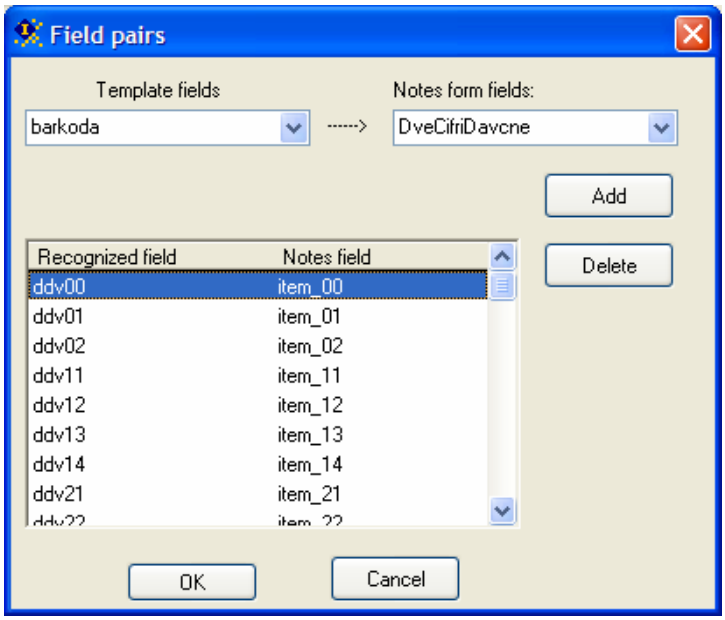
Kot je prikazano na sliki 24, preverjanje pravilnosti podatkov lahko izvedemo s pomočjo pravil definiranih na predlogi obrazca, za morebitne kompleksnejše izraze in pravila pa lahko izdelamo posebno knjižnico.

Poleg baze podatkov je potrebno nastaviti tudi pare polj, torej polja v bazi podatkov, ki se napolnijo z vsebino polj iz obrazca. Nastavitve parov polj vidimo na sliki 25.

Slika 24: Nastavitve za preverjanje podatkov po prepoznavanju



Slika 25: Vmesnik za določanje parov polj med obrazcem in podatkovno bazo



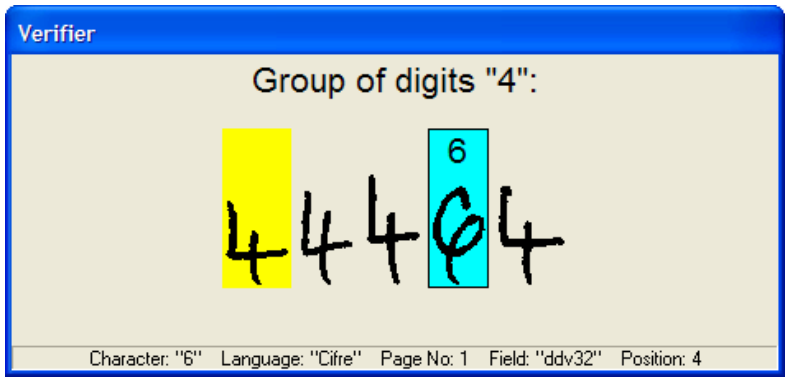
4.5.3 IMiS/ICR odjemalec

Ker je utopično doseči 100 odstotno natančnost, je potrebno po prepoznavanju obrazcev s strani IMiS/ICR strežnika podatke verificirati in te v primeru napak popraviti. IMiS/ICR odjemalec je bil izdelan z namenom čim hitrejše in čim lažje verifikacije rezultatov prepoznavanja.

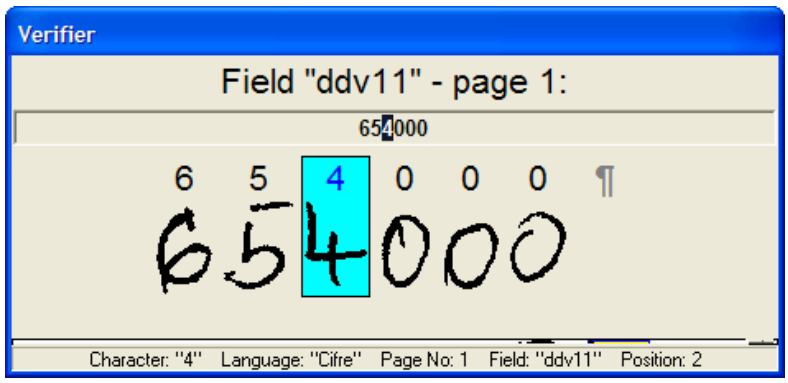
Ergonomsko nesmiselno bi bilo popravljati podatke s primerjanem podatkov na papirju, saj bi bil proces iskanja napak v primeru veliko polj zelo počasen. Enako lahko ugotovimo za iskanje napake s primerjanjem podatkov na skeniranem obrazcu. Zato je bilo potrebno najti način hitrega odkrivanja lokacije napake ter njeno odpravo. IMiS/ICR odjemalec omogoča dvofazno odkrivanje napak. V prvi fazi prenese skeniran obrazec iz elektronskega arhiva in ga ponovno prepozna. Nato združi vse znake, ki so bili klasificirani v isti razred in uporabniku prikaže skupine znakov eno za drugo. Ta način uporablja metodo skupinskega preverjanja

znakov (*angl. group checking*) in je prikazan na sliki 26. V tem primeru so prikazane slike vseh znakov, ki so bili prepoznani kot številka 4. Napako lahko zelo hitro najdemo in jo takoj popravimo. Če nismo prepričani, lahko znak označimo (glej prvi znak 4 na sliki 26) in po končanih preverjanjih vseh skupin znakov se nam bo pojavilo polje z označenim znakom, kot je prikazano na sliki 27. Ko preverimo vse skupine znakov, se ponovno preverijo vsa logična pravila nad podatki. S tem je prva faza verifikacije končana. V primeru pravilno validiranih podatkov, se podatki izvozijo v ciljno podatkovno bazo.

Slika 26: Skupinsko preverjanje znakov



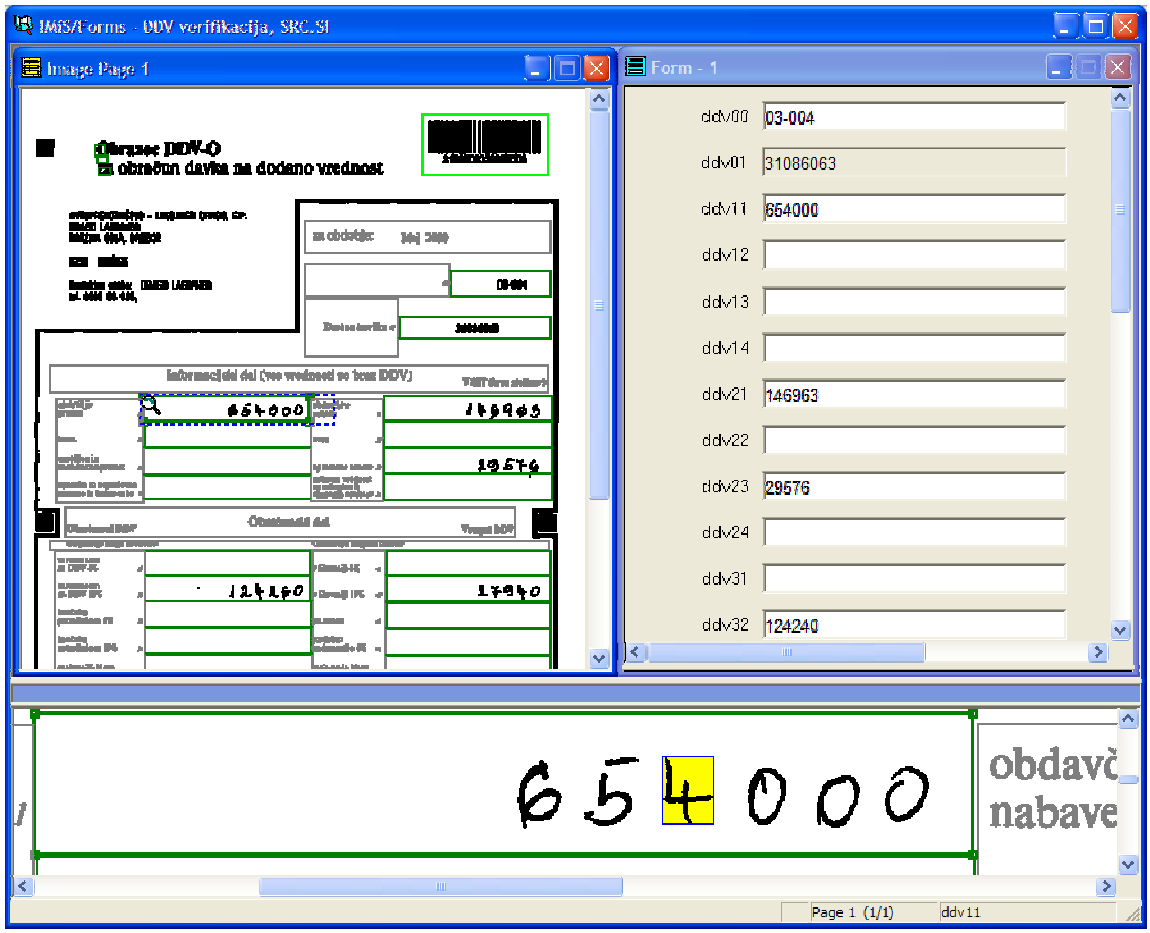
Slika 27: Prikaz označenega znaka v kontekstu polja



Če v prvi fazi ne odpravimo vseh napak, nam IMiS/ICR odjemalec omogoči iskanje napake v drugi fazi. Ta ne uporablja več skupinskega preverjanja, temveč vodi uporabnika od polja do polja. Odpre se nam grafični vmesnik, ki je prikazan na sliki 28. Kot vidimo, je ekran razdeljen na 3 dele. Na levi strani je prikazana slika celotnega obrazca, pri čemer je označeno polje v katerem se trenutno nahajamo. Na desni strani so prepoznane vrednosti polj, kjer lahko podatke popravljamo, spodaj je pa povečana slika dela obrazca, kjer se nahaja trenutno polje. S premikanjem po poljih lahko primerjamo vsa polja in odkrijemo ter popravimo napako.

IMiS/ICR odjemalec omogoča hitro lociranje in odpravo napake. Poleg tega preveri pravilnost popravljenih podatkov in popravke izvozi v ciljno bazo podatkov.

Slika 28: Druga faza verifikacije obrazca



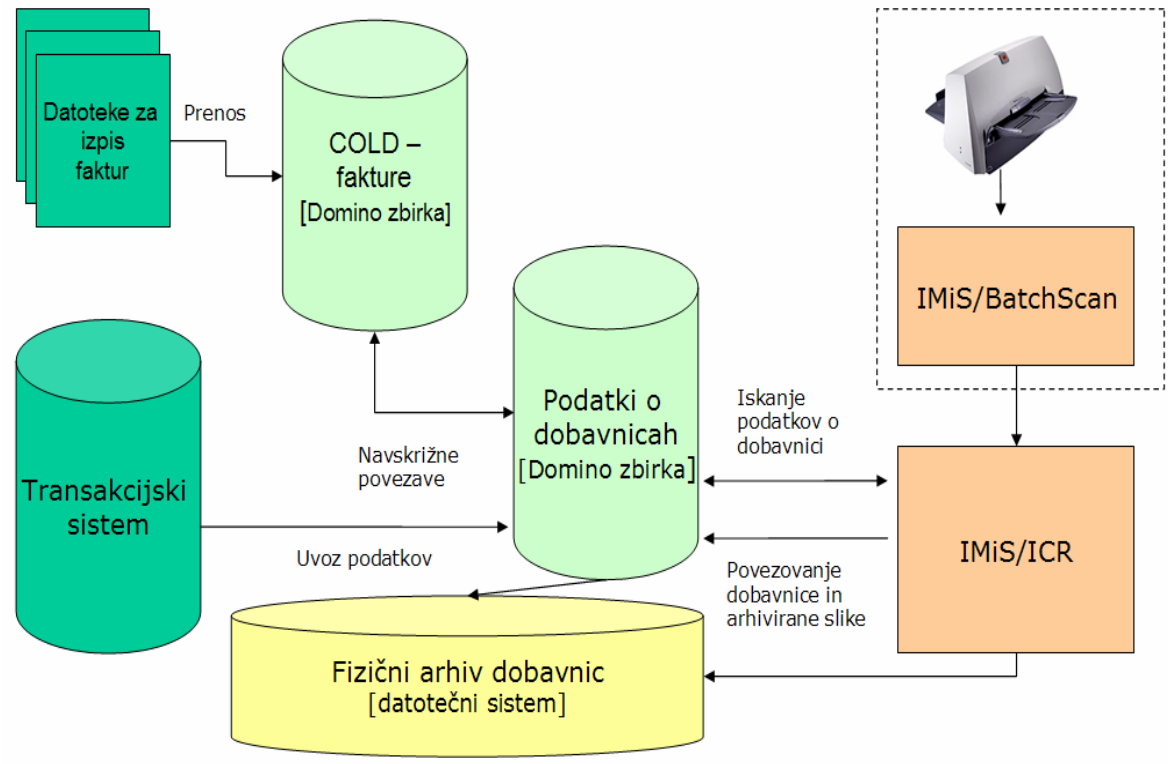
4.6 Primer: Ljubljanske mlekarne d.d.

Konec leta 2003 je bil izveden pilotski projekt v podjetju Ljubljanske mlekarne, ki je vključeval tehnologijo prepoznavanja obrazcev. Na danem primeru bom skušal prikazati tehnološke težave pri vpeljavi prepoznavanja obrazcev, predvsem problem oblike obrazca. Analiza stanja je pokazala, da v poslovnih procesih podjetja Ljubljanske mlekarne d.d. nastaja veliko število papirnih dokumentov, ki v večini nastanejo v enem izmed informacijskih sistemov podjetja. Ti papirni dokumenti po eni strani predstavljajo zamudno delo, po drugi strani pa za arhiviranje teh dokumentov porabijo zelo veliko prostora. Ker je arhiviranje papirno, je iskanje dokumentov ustrezno počasno. Ocenjeno je bilo, da je kar nekaj dokumentov arhivirano večkrat, za potrebe posameznih sektorjev. V okviru

projekta je bilo potrebno analizirati vpeljavo skeniranja in elektronskega arhiviranja dobavnic, prepoznavanje in vnos indeksiranih podatkov, povezavo s transakcijskim sistemom, arhiviranje izpisov faktur v elektronski obliki (COLD) ter povezovanje faktur in pripadajočih dobavnic v obe smeri.

Sam sem sodeloval v projektu in bil zadolžen za izpeljavo skeniranja in prepoznavanja dobavnic. Iz obstoječega sistema transakcijskega sistema smo prevzeli podatke za 1.708 faktur, 22.837 dobavnic ter 902 papirnih dobavnic, kar je po podatkih iz analize približno 5% količine za dekado. Za zajem teh dokumentov smo zgradili COLD sistem za iskanje in pregledovanje izpisanih faktur. Postavili smo tudi sistem za paketno skeniranje dobavnic ter sistem prepoznavanja obrazcev, kjer smo želeli prepoznati številko dobavnice, ki nam bi omogočila avtomatsko umestitev skenirane dobavnice v dokumentni sistem. S posamezne fakture je bilo moč doseči povezane dobavnice, iz teh pa tudi povezano fakturo. Shema sistema je predstavljena na sliki 29.

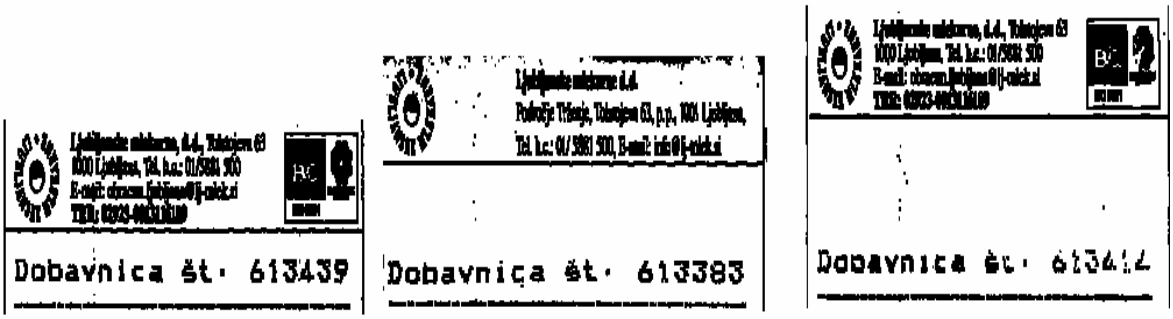
Slika 29: Shema procesa v podjetju Ljubljanske mlekarne d.d.



Za pravilno in hitro umestitev skenirane dobavnice, je bilo potrebno prepoznati številko dobavnice, torej le eno polje na celotnem obrazcu. Dobavnice so lahko večstranske, število strani pa ni fiksno. Za razdelitev dobavnic iz paketa v svoje datoteke s pravilnim številom strani poskrbi programska oprema paketnega skeniranja. Uporabili smo IMiS/BatchScan programsko opremo. Analiza skeniranih

dobavnic je pokazala osnovno težavo prepoznavanja obrazcev. To je segmentacija obrazca in pravilna lokacija polja številke dobavnice na skeniranem dokumentu. Lokacija številke dobavnice je zaradi različnih postavitev papirja in začetka tiskanja zelo nihala. Na sliki 30 lahko vidimo tri relativno zelo različne lokacije številke dobavnice glede na logo podjetja.

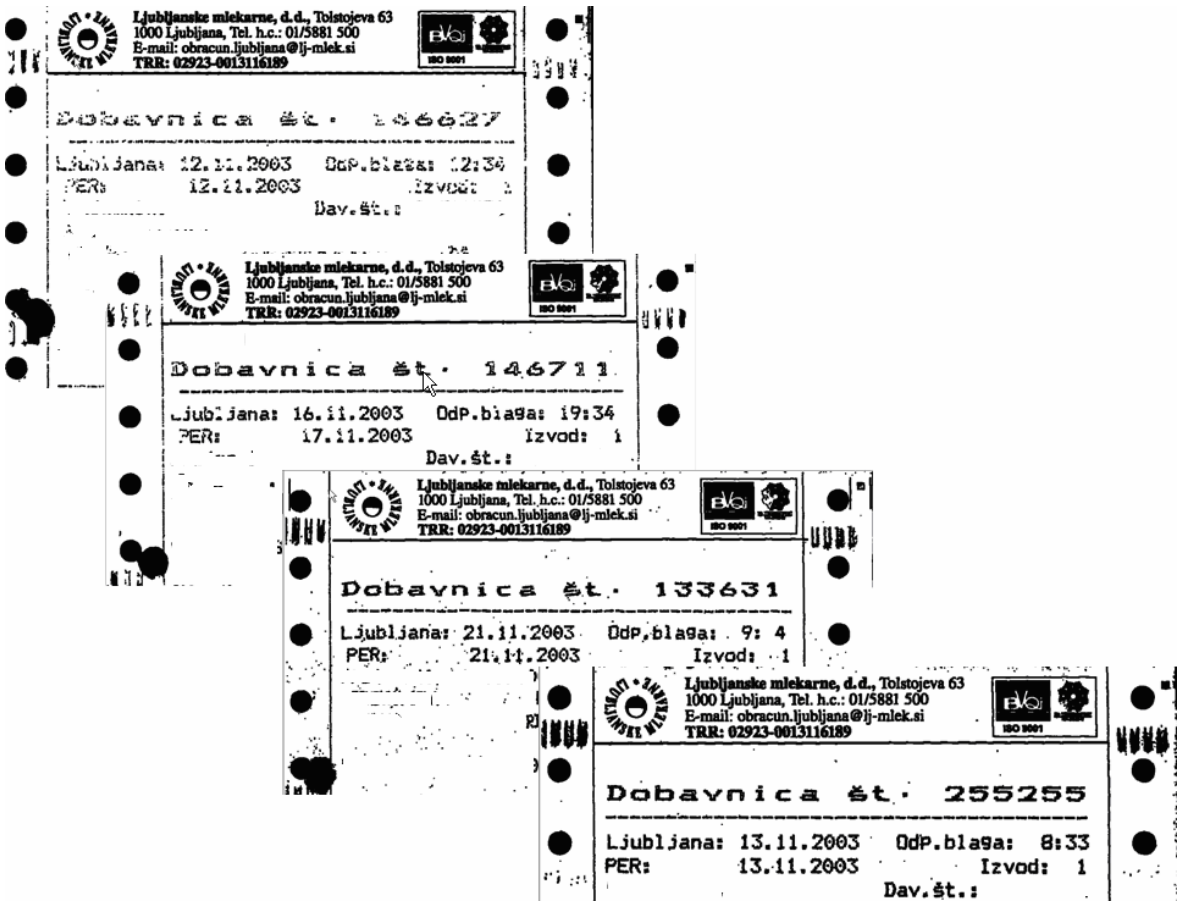
Slika 30: Različne lokacije polja številke dobavnice



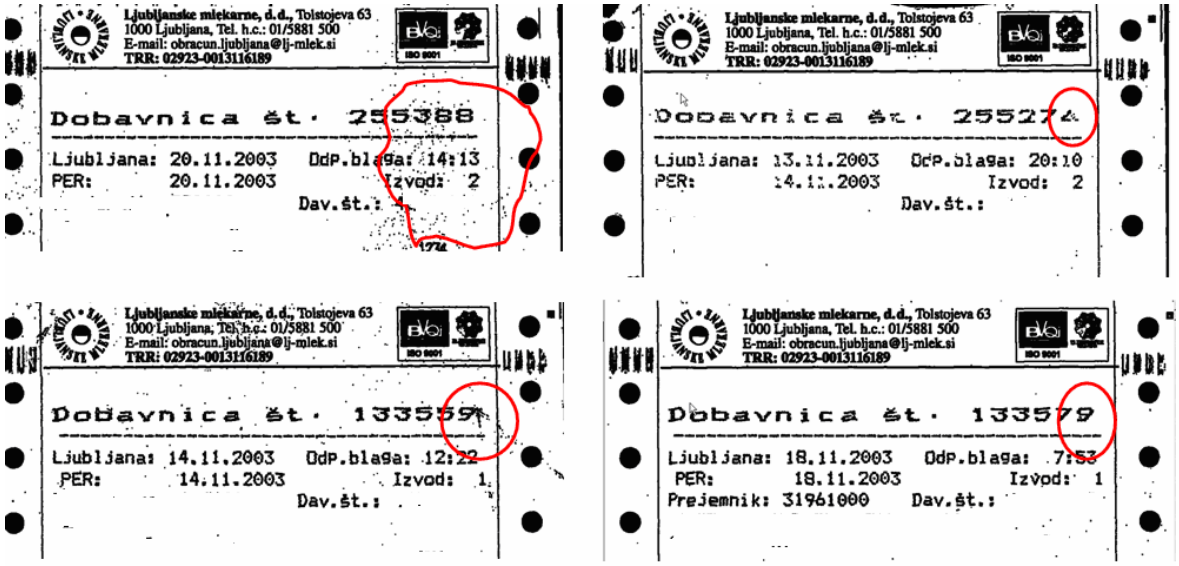
Žal obrazec ni omogočal izbire lokatorjev, ki bi omogočali pravilno lokacijo polja številke glede na lokator, spremembe oblike obrazca pa v fazi pilotskega projekta niso bile mogoče. Zato smo definirali tri različne predloge obrazcev, ki so do neke mere rešile problem segmentacije polja številke. Naslednji problem so bile velike razlike v kvaliteti natisnjenih dobavnic, saj se tiskajo na različnih tiskalnikih z različno kakovostjo. Te razlike so prikazane na primerih na sliki 31, kjer je zadnja dobavnica tiskana z mnogo močnejšo barvo kot naprimer prva. Razlike v kvaliteti natiskanih dobavnic so povzročile težave pri iskanju optimalnih nastavitev skeniranja. Temnejše skeniranje bi pri kvalitetno natiskanih dobavnicah povzročilo prisotnost motenj, kar lahko povzroča hipersegmentacijo znakov. Nasprotno pa presvetlo skeniranje pri slabo tiskanih dobavnicah izpušča črne pike in lahko povzroča lomljenje oziroma podsegmentacijo znakov. V ta namen smo uporabili možnost učenja znakov, v katerem smo zajeli čim več cifer, katerih razlike so bile posledica različne kvalitete tiska.

Kljub temu so bili rezultati na vzorčni množici 902 dobavnic presenetljivo dobri. 791 številk dobavnic je bilo pravilno prepoznanih in povezanih s podatki v dokumentnem sistemu. V 74 primerih je bil zavrnjen tip obrazca (napaka identifikacije), v 37 primerih prepoznavanje ni bilo pravilno. Na sliki 32 so prikazane nekatere dobavnice, ki niso bile prepoznane pravilno. Rezultat prepoznavanja je bil skoraj 87 odstoten, pri čemer lahko pričakujemo, da bi na konkretnem primeru s spremembo oblike obrazca prišli do rezultata med 95 in 100 odstotki.

Slika 31: Različne kvalitete tiska dobavnic



Slika 32: Primeri napačno prepoznanih številk dobavnic



Kot smo ugotovili, v danem primeru večji problem predstavlja segmentacija obrazca, kot pa samo prepoznavanje samih znakov. Prepoznavanja znakov lahko sistem »naučimo«, medtem ko je prepoznavanje tipa in lokacije polj obrazca brez prilagoditve oblike obrazca ali še bolje črtne kode, težje in manj zanesljivo. Validacija podatkov je bila izvedena s primerjanjem prepoznane številke s podatkom iz podatkovne baze oziroma dokumentnega sistema. Tako smo lahko učinkovito izločili nepravilno prepoznane dobavnice in jih nato verificirali s pomočjo IMiS/ICR odjemalca.

Ob dodatnih znanjih, ki smo jih pridobili s povečanjem vzorčne množice, smo v zadnji iteraciji dosegli celo rezultat 93,23 odstotkov pravilnih številk na nivoju dobavnice. Po analizi rezultatov smo ugotovili, da lahko še dodatno izboljšamo natančnost prepoznavanja, saj se številke dobavnice v sistemu pojavljajo znotraj določenega obsega vrednosti, kar pomeni da lahko pričakujemo rezultat v dovolj ozkem območju vrednosti in ga tako tudi omejimo oziroma algoritmu sporočimo prve 2 ali tri cifre v številki. Pilotski projekt je pokazal, da bi utečen sistem v produkciji deloval blizu 100 odstotne natančnosti.

5. SMOTRNOST UPORABE PREPOZNAVANJA OBRAZCEV

Smotrnost implementacije prepoznavanja obrazcev je odvisna od samega poslovnega procesa ter količine in vpliva obrazcev nanj. V nadaljevanju bom analiziral stroške in koristi uporabe prepoznavanja obrazcev ter podal splošne smernice za tovrstne projekte. Zmanjšane stroške in koristi je potrebno primerjati z stroški investicije in ugotoviti čas povrnitve investicije. Ob koncu poglavja bom podal dva primera iz tujine, ki potrjujeta uspešnost prepoznavanja obrazcev.

5.1 Analiza stroškov

Stroške investicije je potrebno ocenjevati na konkretnem primeru. Poudariti je potrebno, da niso stroški investicije le nakup strojne in licenčne programske opreme, temveč je potrebno oceniti še stroške izobraževanja uporabnikov in predvsem stroške uvajanja sistema. Predvsem uvajanje in integracija sistemov prepoznavanja obrazcev zahteva veliko projektnih ur, saj je potrebno veliko testiranja in analiz vzorčnih množic obrazcev, kar omogoča prilagoditve oblike obrazca in s tem večjo natančnost prepoznavanja ter večje stroškovne koristi produkcijskega sistema. Tudi ko sistem prepoznavanja deluje v produkciji, je v

začetnem obdobju potrebno spremljanje delovanja procesa in iskanje morebitnih dodatnih sprememb, ki bi povečale hitrost prepoznavanja in obdelave podatkov.

5.2 Analiza koristi

Stroškovne koristi so v primeru uspešnih vpeljav prepoznavanja obrazcev lahko zelo velike in v tem primeru govorimo o relativno kratkem času povrnitve investicije. Z analizo zmanjšanih stroškov opredelimo **kratkoročne učinke** vpeljave prepoznavanja obrazcev.

V primeru neposrednih stroškovnih koristih (*angl. Hard-dollar benefits*) govorimo o prihrankih, ki so takoj očitni in jih lahko merimo. Izhajajo iz vpeljave aplikacije oziroma informacijskega sistema in pomenijo neposredno zmanjšanje stroškov, ki nastane s preoblikovanjem procesa. Ob uspešni vpeljavi prepoznavanja obrazcev je najbolj očiten prihranek pri strošku dela. Število operaterjev, ki so pred vpeljavo na roke vnašali podatke, se namreč občutno zmanjša. Delovna mesta vnašalcev podatkov se reducirajo in preoblikujejo v delovna mesta za nadzor skeniranja obrazcev ter verificiranja in popravljanja prepoznanih podatkov. Relativno zmanjšanje stroškov dela je odvisno od same narave procesa in natančnosti prepoznavanja, ki pogojuje stroške verifikacije napak, vendar se v večini primerov število delovnih mest, v z obrazci intenzivnih procesih, zmanjša za vsaj 60 odstotkov (Gingrande, 2000).

Nadalje lahko neposredno merimo prihranke pri stroških delovnih prostorov, ki izhajajo neposredno iz zmanjšanja števila delovnih mest. Tako analiziramo prihranke v zmanjšanju potrebnega delovnega prostora, zmanjšanju stroškov storitev (elektrika, čiščenje, itd.) in tudi zmanjšanju stroškov materiala. Zaradi zmanjšanja števila delovnih mest vnašalcev podatkov je namreč potrebnih manj delovnih postaj, s čimer prihranimo pri strojni in programski opremi, itd.

Zaradi hitrejšega prenosa podatkov iz papirja v podatkovne baze lahko praviloma prihranimo tudi pri stroških financiranja projekta. Če je na primer naš tipični projekt analiza zadovoljstva kupcev našega naročnika, kjer so izvor podatkov izpolnjeni obrazci, lahko s hitrejšim rezultatom zmanjšamo stroške financiranja projekta, saj hitreje dobimo plačano storitev. Tako govorimo o hitrejši uporabi prihodkov iz projekta.

Čas povrnitve investicije v vpeljavo sistema za prepoznavanje obrazcev je v tipičnih študijah primerov manj kot dve leti. Praviloma se donosnost procesov poveča z zmanjšanjem stroškov, v primeru večje natančnosti in zmožnosti

procesiranja večje količine obrazcev se pa lahko čas povrnitve investicije še zmanjša.

Analiza zmanjšanja stroškov je neposredno merljiva in da konkreten rezultat, pri analizi posrednih koristi pa ni tako. Gre bolj za posredne in s tem težje merljive učinke vpeljave prepoznavanja obrazcev, ki vplivajo na preoblikovanje procesa in s tem na učinkovitost. Te lahko delimo na dve skupini (Gingrande, 2000; Spencer, 1996):

- posredne koristi, ki izhajajo iz prenove poslovnega procesa (*angl. Soft-dollar benefits*),
- strateške koristi, ki izhajajo iz povečanja poslovne uspešnosti podjetja.

Zaradi preoblikovanega poslovnega procesa in zmanjšanega ali odpravljenega ročnega vpisovanja podatkov zmanjšamo stroške poškodb in bolezni, ki sledijo iz narave delovnega mesta. Tipične poškodbe zaradi tipkanja so poškodbe zapestja ali tunelskega živca. V študijah primerov najdemo podatke o zmanjšanju stresa, ki nastane zaradi monotonosti ročnega vpisovanja podatkov ter zmanjšanju poškodb in bolezni oči. S tem se poveča udobje in prijaznost delovnih mest ter zadovoljstvo delojemalcev. V ZDA posvečajo analizi tovrstnih koristi veliko prostora, saj so stroški zdravljenja ali celo tožb delojemalcev do delodajalcev prisotni v večji meri kot pri nas.

Zanimiv je podatek, da je v procesih z veliko obrazci vlaganje v dokumentne sisteme z elektronskim arhivom skeniranih dokumentov stroškovno vprašljivo, če ne vpeljemo tudi sistema za prepoznavanje obrazcev. Dokazano je namreč, da je pretipkavanje podatkov iz poskeniranih slik mnogo počasnejše kot iz papirja. Analize so pokazale, da izkušeni operaterji iz papirja pretipkajo v povprečju dvakrat več znakov na časovno enoto kot iz identičnega poskeniranega dokumenta. Tako v bistvu zmanjšamo hitrost in natančnost ročnega vnosa podatkov (Gingrande, 2000; Spencer, 1996).

Nenazadnje je posredna korist tudi večja natančnost in hitrost prenosa podatkov iz papirja v podatkovno bazo. Lahko govorimo o povečani produktivnosti v smislu količine obdelanih podatkov, o skrajšanju trajanja procesa za obdelavo enake količine podatkov in hitrejši odzivnosti do strank.

V primerih, ko je potreben tudi kasnejši vpogled v originalne obrazce, se izkaže prednost dokumentnega sistema z elektronskim arhivom poskeniranih obrazcev. Iskanje in prikaz željenega obrazca je neprimerno hitrejši proces kot v klasičnem arhivu, pa tudi stroški arhiva so manjši. Ob analizi koristi, ki izhajajo iz

spremenjenega delovnega procesa, torej govorimo o **srednjeročnih učinkih** vpeljave prepoznavanja obrazcev.

V primeru strateških koristi je potrebno analizirati koristi vpeljave prepoznavanja obrazcev na nivoju podjetja in govorimo o **dolgoročnih učinkih**. Predvsem gre za povečano konkurenčno prednost, ki izhaja iz:

- optimizacije in boljšega izkoristka delavcev zaradi eliminacije tradicionalnih ponavljajočih se delovnih opravil,
- izboljšane ponudbe,
- učinkovitejše razporeditve sredstev,
- hitrejšega odziva do strank,
- povečane produktivnosti.

Strateške koristi sicer niso neposredno merljive, kot je primer prihrankov z zmanjšanjem neposrednih stroškov, vendar se pa odražajo v parametrih, ki so bistvena usmeritev vsakega podjetja in njegovega poslovanja. Poslovno opravičujejo implementacijo sistema prepoznavanja obrazcev, ki omogoča preoblikovanje poslovnih procesov. Kot najpomembnejše neposredno zmanjšanje stroškov omenjam strošek dela, do katerega pa zaradi socialne komponente ob odpuščanju ali potrebe na drugih delovnih mestih dostikrat ne pride, temveč se izvede reorganizacija delovnih mest in prerazporeditev delavcev. Zato se strošek dela neposredno ne zmanjša, temveč se pozitivni učinki reorganizacije kažejo v boljšem izkoristku človeških virov, kar je strateški učinek na nivoju podjetja.

5.3 Splošne smernice pri uvajanju tehnologije prepoznavanja obrazcev

Ko analiziramo študije vpeljav projektov prepoznavanja obrazcev lahko ugotovimo, da projekti v veliko primerih niso uspešni. Razlogov je seveda več, vendar pa imam občutek, da gre velikokrat za nezaupanje v tehnologijo zaradi same mladosti tehnologije ter pomanjkanja uspešnih implementacij. Tako se do neke mere vrtimo v krogu, saj zaradi v preteklosti neuspešnih projektov težje izpeljemo oziroma upravičimo potencialno uspešen projekt. Pravtako je ravno v času prvih uspešnih implementacij prepoznavanja obrazcev vzporedno prišlo do ekspanzije uporabe interneta in poslovanja prek interneta, kar je začasno zajezilo projekte prepoznavanja obrazcev, saj je trend razvoja kazal v smer poslovanja z manj papirja. Pa je temu res tako? Seveda se vse bolj uporablja elektronska izmenjava dokumentov in podatkov, ter tudi elektronsko izpolnjevanje obrazcev, vendar pa analize kažejo, da absolutno gledano uporabljamo papir kot vir informacije vse več

in ne vse manj (Gingrande, 2000). Če pogledamo vsakdanje življenje tako na delovnem mestu kot doma, lahko temu pritrdimo.

Vpeljave sistema prepoznavanja obrazcev se je potrebno lotiti projektno. Tovrstnih sistemov in rešitev se torej ne da uporabiti kot izdelanih produktov, temveč so vedno potrebne analize procesa, obrazcev, rezultatov in sprotni popravki ter prilagoditve sistema. Sama natančnost prepoznavanja znaka niti ni odločilni dejavnik, saj danes komercialno dosegljivi produkti dosegajo podobne natančnosti. Bistvena kvaliteta izbranega sistema naj bo možnost integracije in zato je zelo pomembna tudi izbira morebitnega zunanjega partnerja. Še bolj kot drugje je potrebno upoštevati reference podjetja, ki mu bomo zaupali izpeljavo projekta. Tu ne gre le za reference na področju vpeljave prepoznavanja obrazcev, temveč je potrebno tudi znanje in izkušnje o poslovnih procesih, načinu in organizaciji dela in podobno. Le na ta način lahko iščemo pravilne smeri prilagoditve sistema, ter z znanjem o poslovnem procesu tudi določimo in pravilno ovrednotimo prihranke in koristi integracije sistema (Rittman, 2001).

Kot smo že poudarili pri obravnavi tehnologije prepoznavanja obrazcev, je zelo pomemben dejavnik uspeha sama oblika obrazca. Zato je za pilotski projekt priporočeno izbrati proces, v katerem imamo kontrolo nad obliko obrazca. Ponavadi je koristno prva testiranja izvesti na originalnem, nespremenjenem obrazcu, ter nato postopoma spreminjati obrazec in dokazovati vpliv oblike na rezultate. Na tak način lahko merimo in analiziramo vpliv posameznih sprememb oblike obrazca na natančnost in rezultate prepoznavanja.

Bistvo vsakega projekta vpeljave informacijskih sistemov je v stroškovni upravičenosti in koristnosti in projektni vodje morajo posvetiti velik del analize ravno tem kazalcem. Pomembno je, da si že pred začetkom projekta zastavimo cilje in določimo merljive parametre, ki bodo kazalci uspešnost ali neuspešnost projekta. Poudarek naj bo na merljivih neposrednih stroških, kot so stroški dela, stroški delovnih prostorov, zmanjšanje operativnih stroškov procesa ter morebitna hitrejša pridobitev prihodkov iz procesa. Če se le da, naj iz teh kazalcev prikažemo tudi posredne oziroma srednjeročne ter strateške oziroma dolgoročne učinke.

Da pa bi lahko dokazali stroškovno upravičenost in koristnost prepoznavanja obrazcev, je nujno izpeljati pilotski projekt, katerega rezultati naj pokažejo smotrnost vpeljave tehnologije. Pilotski projekt je torej analiza procesa, ki naj dokaže ali ovrže upravičenost preoblikovanja procesa z vpeljavo prepoznavanja obrazcev. Za pilotski projekt je potrebno izbrati proces, ki najbolj vpliva na stroške podjetja in je v samem srcu osnovne dejavnosti podjetja, kar pomeni, da lahko z rezultati tudi najbolj prepričamo srednji in visoki management v podjetju. Konec

koncev je vsak projekt »političen«, zato je v visokem managementu priporočljivo imeti botra projekta, ki bo zagovarjal in s svojo avtoriteto omogočal nemoteno izvedbo projekta. Proces mora biti papirno intenziven in hitrost oziroma časovni aspekt pretvorbe informacije v elektronsko obliko naj odločilno vpliva na rezultat procesa. Seveda lahko prihranke simuliramo s pomočjo modela procesa in predvidene natančnosti in hitrosti prepoznavanja, vendar iz izkušenj lahko trdim, da je pilotski projekt nujno potreben za pridobitev podatkov o natančnosti na vzorčni množici, ki omogočajo analizo prihrankov. Rezultate analize je potrebno do neke mere preslikati na druge procese in oddelke ter tudi oceniti stroške tovrstnih preslikav. Ti so seveda dosti manjši od inicialnih stroškov, saj nam je že na voljo infrastruktura in znanje za vpeljavo prepoznavanja obrazcev. Pridobljene podatke in rezultate primerjamo s podatki iz procesa pred preoblikovanjem ter s sprejeto statistiko iz raznih analiz. Naj navedemo nekaj splošnih, vendar mnogokrat primernih podatkov za primerjavo rezultatov, povzetih po organizaciji AIIM:

- 25% časa ročnega vnašanja porabimo za prinašanje in odnašanje papirja
- k času za ročni vnos je potrebno dodati 17% zaradi utrujenosti
- v povprečju porabimo 2,5 delovnih ur na teden za iskanje dokumentov
- 90% dokumentacije v podjetju je še vedno na papirju
- okoli 32% dokumentacije v podjetju je stalno v uporabi
- okoli 3% papirne dokumentacije založimo, oziroma se izgubi
- povprečno iščemo založeni dokument 10-30 minut
- strošek iskanja založenega dokumenta znaša 6$ – 120$ (analiza za ZDA)

Pilotski projekt mora zajemati vse komponente procesa v produkciji in mora biti skalabilen, razširljiv. Tudi če posameznih komponent zaradi stroškov ne povežemo kot v realnem procesu v produkciji in zato nekatere celo simuliramo, je treba predvideti vpliv prenosa tehnologije iz pilotskega projekta v produkcijo.

Poleg kratkoročnih in srednjeročnih je potrebno analizirati tudi dolgoročne, torej strateške učinke preoblikovanja procesa. Morebitna uspešnost obdelave obrazcev naj ne vpliva le na zmanjšanje stroškov samega procesa, temveč naj pokaže morebiten vpliv na rast prihodkov, povečanje konkurenčne prednosti ter večanje tržnega deleža. To je sicer težko merljivo in življenska doba pilotskega projekta je v večini primerov mnogo prekratka za tovrstne konkretnejše analize, vendar se uspeh vpeljave prepoznavanja obrazcev kaže tudi v strateških koristih na nivoju podjetja. Če nam uspe dokazati vpliv preoblikovanega procesa na strateške kazalce podjetja, potem je mnogo lažje najti »botra« v visokem managementu podjetja, ki bi podprl projekt.

5.4 Prihodnost prepoznavanja obrazcev

Razvoj internet tehnologij je prispeval k manjšemu zanimanju za sisteme prepoznavanja obrazcev, saj se je pričakovalo preoblikovanje papirno intenzivnih procesov z uporabo elektronskih obrazcev in direktnega vnašanja podatkov v obrazce prek interneta ter z razvojem elektronske izmenjave podatkov. Vendar se je izkazalo, da se kljub vpeljavi sistemov elektronskih obrazcev, količina papirnih obrazcev ni zmanjšala, temveč se je celo povečala, zmanjšal se je le trend rasti (Gingrande, 2000). Problem je v večini primerov v sami naravi procesov, saj informacije včasih zahtevajo papirno obliko. Po drugi strani je za identifikacijo izpolnjevalca obrazca potrebna vpeljava tehnologije elektronskega podpisa in zakonske spremembe, ki pa se kljub informacijskem razvoju dogajajo počasi. Ravno zakonska podlaga oziroma pravna veljavnost papirja v primerjavi s poskeniranim dokumentom, kot smo omenili v poglavju 2.4.3, v veliki meri preprečuje vpeljavo elektronskih obrazcev. Drugače je z internimi obrazci v podjetju. Ti se predvsem v velikih podjetjih vse bolj preoblikujejo v elektronsko obliko, vendar tovrstni obrazci niso nikoli bili vir za sisteme prepoznavanja obrazcev. Tipičen izvor podatkov, ki jih obdelujemo v sistemih prepoznavanja obrazcev, je izven podjetja, pri čemer podjetje ima vpliv na obliko obrazca, oziroma ga kreira in distribuira. In ti obrazci so v večji meri še vedno na papirju. Sistemi elektronskih obrazcev so le preprečili še večjo rast števila papirnih obrazcev (Hayes, 2001; Spencer 2001b).

Papir je globoko zakoreninjen v naši kulturi in po mojem mnenju še nekaj generacij ne bo pripravljeno opustiti papirja kot nosilca informacije. Človek ga uporablja že 2000 let in papir ima veliko prednosti: prenosljivost, nadomestljivost, standardiziranost, ne potrebuje električne energije, ne potrebuje posebnega pregledovalnika, itd. Tovrstni razlogi v današnjem informacijskem svetu delujejo nekoliko smešno, vendar vse večja poraba papirja potrjuje kulturno vkoreninjenost. Ljudje želijo imeti izbiro pri vrsti komunikacije. Dovolj velik dokaz je tudi uporaba elektronskega davčnega poslovanja in možnost oddaje dohodnine za leto 2003 v Sloveniji prek sistema edavki. Za elektronsko oddajo se je odločilo 16.745 uporabnikov, kar je relativno malo in celo trikrat manj, kot so optimistično napovedali na Davčni upravi Republike Slovenije (Kovšca, 2004). Pravgotovo je elektronsko davčno poslovanje velik in zelo pozitiven korak v delovanju e-države, vendar ta podatek kaže, da imamo kljub vpeljavi elektronskega poslovanja še vedno veliko obrazcev (v tem primeru dohodninskih napovedi) v papirni obliki in je torej razmišljanje o integraciji sistemov prepoznavanja obrazcev, navkljub sistemu elektronskega obrazca, upravičeno.

Naj omenim še projekt e-SLOG. Gospodarska zbornica Slovenije vodi projekt e-SLOG, v okviru katerega so nastala priporočila za elektronsko povezovanje v okviru Slovenije. Bistvena prednost sistema je izvorna elektronska oblika papirja, s čimer se izognemo papirni obliki. Tako smo dobili tudi XML shemi za kompleksni in enostavni račun oziroma standarda na področju oblike elektronskega računa. Gre torej za standardizacijo elektronske izmenjave podatkov in standardizacijo e-dokumenta, ki z veljavnimi elektronskimi podpisi potrjuje pravno veljavnost dokumenta. Podjetja, ki bodo implementirala ta standard, si bodo lažje izmenjevala podatke, kar bo zmanjšalo porabo papirja in informacije na papirju, vendar ti podatki nikoli niso bili predmet prepoznavanja obrazcev.

Sisteme prepoznavanja obrazcev uvrščamo znotraj dokumentnih sistemov. Investicije v tovrstne sisteme pravgotovo niso majhne, zato je prva ovira za vpeljavo prepoznavanja obrazcev njegov temelj, torej dokumentni sistem in znotraj njega sistem za skeniranje in arhiviranje dokumentov. Relativno veliko podjetij se namreč še ne zaveda pomena dokumentnih sistemov ter njihovih prednosti, vendar trend kaže osveščanje v tej smeri. Implementacije poslovnih informacijskih sistemov kot osnovne informatizacije podjetja so v večini podjetij končane, tako da imajo projektne ekipe na voljo tako časovne kot finančne vire za naslednjo stopnjo informatizacije – dokumentni sistem. Ko so tovrstni sistemi uspešno integrirani v podjetju, se pokažejo potrebe po obvadovanju vsebine iz papirja in takrat je čas za sisteme optičnega prepoznavanja znakov in prepoznavanja obrazcev. Pogoj za vpeljavo tehnologije prepoznavanja obrazcev je torej relativno visoka stopnja informatizacije podjetja. Navkljub razvoju elektronskega poslovanja in s tem uporabe sistemov elektronskih obrazcev, se uporaba papirnih obrazcev ne zmanjšuje. To pomeni, da se trg za sisteme prepoznavanja obrazcev ne krči, temveč se veča. Procesi, ki zahtevajo široko množico izpolnjevalcev obrazcev so primerni za implementacijo tako elektronskih obrazcev (če narava procesa in zakonska podlaga to dopušča) kot tudi sistemov prepoznavanja obrazcev. Skupaj omogočajo preoblikovanje in informatizacijo procesa, kar prinaša prihranke in hitrejšo obdelavo podatkov ter hitrejšo odzivnost in mnogo večjo učinkovitost procesa. Od samega procesa je pa odvisno v kakšnem obdobju se investicija povrne.

5.5 Primeri

V 6. poglavju je natančneje opisan pilotski projekt prepoznavanja obrazcev za prijavo Davka na dodano vrednost, v tem poglavju bomo pa nekoliko bolj grobo predstavili nekaj drugih primerov implementacije tehnologije prepoznavanja obrazcev.

5.5.1 Finansbank (Turčija)

Ob prošnji za pridobitev kreditne kartice banke, mora komitent priložiti nekaj dodatnih dokumentov, kot so fotokopije potnega lista, plačilne liste, itd. Podatki, ki jih vnese na vlogi, se morajo vnesti v bančno bazo podatkov skupaj s poskeniranimi prilogami. Pred implementacijo sistema za prepoznavanje obrazcev je bilo zaposlenih 80 operaterjev, ki so lahko v enem delovnem dnevu obdelali skupaj 1000 obrazcev. Zaradi vse večjega dnevnega obsega obrazcev in s tem povezanih višjih stroškov dela, prostora in materiala, je Finansbank začela iskati učinkovitejšo rešitev avtomatske obdelave obrazcev. Po pilotskem testiranju je bil vpeljan sistem prepoznavanja obrazcev podjetja ABBYY. Obrazci in priloge se poskenirajo, produkt ABBYY FormReader pa prepozna obrazec. Verifikacija poteka avtomatsko s pomočjo pravil in s pomočjo operaterjev, kar zagotavlja 100 odstotno pravilno izvožene podatke. Skenirani objekti se registrirajo v sistemu NovaManage EDM System, ki omogoča vpogled v skenirane objekte v vsakem trenutku, podatki pa se izvozijo v informacijski sistem banke, ki temelji na podatkovnih bazah podjetja Oracle. Rezultati kažejo na velik uspeh projekta. V povprečju je potrebnih 90 sekund za skeniranje, prepoznavanje, verificiranje, izvoz podatkov in vnos skeniranih dokumentov v EDMS. Zaposlenih je le še 10 operaterjev, ki v preoblikovanem procesu lahko obdelajo 10.000 obrazcev dnevno (ABBYY Case studies & success stories [URL:http://www.abbyy.com], 2004).

5.5.2 Popis prebivalstva v ZDA leta 2000

Leta 2000 se je v ZDA izvajal popis prebivalstva in U.S. Bureau of the Census (v nadaljevanju BOC) je agencija, ki je zadolžena za projekt popisa prebivalstva. Takratni popis je bil prvi v zgodovini ZDA, pri katerem se je uporabila tehnologija prepoznavanja obrazcev. Cilj je bil ustvariti elektronski arhiv izpolnjenih obrazcev in elektronsko obdelati podatke v polovico krajšem času kot jim je to uspelo ročno leta 1990. S projektom so začeli leta 1993!

BOC je izbral podjetje Lockheed Martin kot glavnega izvajalca projekta in po pogodbi naj bi bilo delo opravljeno v 171 dneh, pri čemer je bila kvantitativna ocena okoli 150 milijonov obrazcev. Lockheed Martin je zbral ekipo takrat vodilnih podjetij na področju tehnologij upravljanja z dokumenti, ter skeniranja in prepoznavanja dokumentov: Kodak je prispeval skenerje, Kofax je poskrbel za programsko opremo za skeniranje, kroženje dokumentov je bilo razvito v Staffware Corp., obdelava slikovnih dokumentov v TMSSequoia Inc, optično prepoznavanje

znakov je prispevalo podjetje CGK, prepoznavanje označb Optimum Solutions Inc., programska oprema za verificiranje in popravljanje podatkov je bila pa razvita v podjetju Captiva Software Corp.

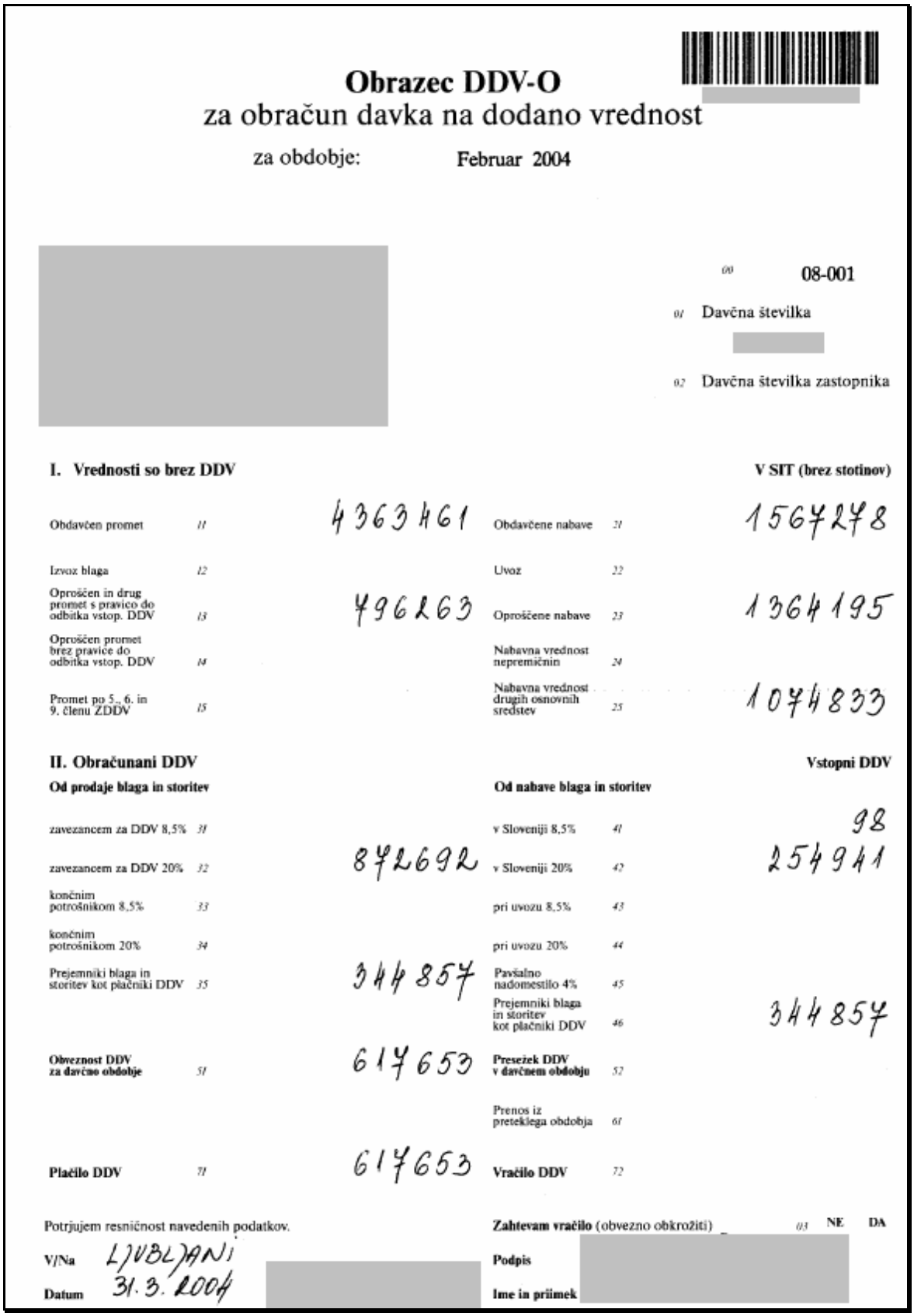
BOC je za potrebe popisa postavil štiri regionalne centre (Data Capture Center), ki so opravljali enake procese. S pomočjo skenerjev, postaj in strežnikov so opravljali proces pretvorbe podatkov iz obrazcev v elektronsko bazo podatkov. Dnevno so se obdelani podatki prenašali v centralno bazo podatkov BOC v Washington D.C.. Vsak regionalni center je obdelal najmanj 35 milijonov obrazcev, prejemali so jih v obdobju enega meseca, v nekaterih dneh je prišlo tudi do 6 milijonov obrazcev dnevno. Na vsaki lokaciji je bilo zaposlenih več kot 2000 ljudi, ki so delali v dveh izmenah 24 ur na dan.

Končni podatki so fascinatni. V popisu je sodelovalo 284,6 milijona prebivalcev, obdelanih je bilo 140 milijonov obrazcev, kateri so skupaj vsebovali 1,5 milijard A4 strani. Pri tem je bilo uporabljenih 50 strežniških sistemov in 10.000 računalnikov, katere je upravljalo 8.000 operaterjev. Delo je bilo opravljeno v 170 dneh. Željena natančnost 98 odstotkov na nivoju alfanumeričnega in 98,5 odstotkov na nivoju numeričnega polja, kar je pomenilo 98,3 odstotkov na nivoju vseh polj, je bila dosežena. Natančnost na nivoju obrazca je bila 38 odstotkov. Projekt je bil ocenjen kot zelo uspešen (Content Management Online; TechInfoCenter; Kofax).

6. ANALIZA SMOTRNOSTI UPORABE NA PRIMERU DDV OBRAZCEV

V prvi polovici leta 2000 je Davčna uprava Republike Slovenije (v nadaljevanju DURS) objavila javni razpis za idejni projekt avtomatskega zajema in prepoznavanja obrazcev za prijavo davka na dodano vrednost (v nadaljevanju DDV). Obrazec ima oznako DDV-0 (glej sliko 33). Podjetje SRC.SI d.o.o. se je prijavilo na razpis in bilo tudi izbrano. Sam sem bil tehnični vodja projekta s strani izvajalca, odgovoren za razvoj programskega modula, ki naj bi omogočil prepoznavanje podatkov in izvoz podatkov v ciljni simtem za upravljanje baz podatkov. Takrat se je začel razvoj IMiS/ICR modula z namenom doseči splošno uporaben produkt na področju prepoznavanja obrazcev. Namen razpisa je bil ugotoviti tehnične možnosti za izvedbo avtomatskega zajema in prepoznavanja podatkov iz DDV obrazcev, morebitne prihranke pri tem in kot rezultat ugotoviti smotrnost objave javnega razpisa za izdelavo tovrstnega sistema. Pilotski projekt naj bi pokazal, ali je tehnologija prepoznavanja obrazcev zrela za uporabo in ali prinaša dovolj velike prihranke časa in stroškov, ki opravičujejo implementacijo.

Slika 33: Primer izpolnjenega DDV-0 obrazca



Obrazec DDV-O
za obračun davka na dodano vrednost

za obdobje: Februar 2004

00 08-001

01 Davčna številka

02 Davčna številka zastopnika

I. Vrednosti so brez DDV — **V SIT (brez stotinov)**

Obdavčen promet	11	4363461	Obdavčene nabave	21	1567248
Izvoz blaga	12		Uvoz	22	
Oproščen in drug promet s pravico do odbitka vstop. DDV	13	496263	Oproščene nabave	23	1364195
Oproščen promet brez pravice do odbitka vstop. DDV	14		Nabavna vrednost nepremičnin	24	
Promet po 5., 6. in 9. členu ZDDV	15		Nabavna vrednost drugih osnovnih sredstev	25	1074833

II. Obračunani DDV — **Vstopni DDV**

Od prodaje blaga in storitev			**Od nabave blaga in storitev**		
zavezancem za DDV 8,5%	31		v Sloveniji 8,5%	41	98
zavezancem za DDV 20%	32	872692	v Sloveniji 20%	42	254941
končnim potrošnikom 8,5%	33		pri uvozu 8,5%	43	
končnim potrošnikom 20%	34		pri uvozu 20%	44	
Prejemniki blaga in storitev kot plačniki DDV	35	344857	Pavšalno nadomestilo 4%	45	
			Prejemniki blaga in storitev kot plačniki DDV	46	344857
Obveznost DDV za davčno obdobje	51	617653	**Presežek DDV v davčnem obdobju**	52	
			Prenos iz preteklega obdobja	61	
Plačilo DDV	71	617653	**Vračilo DDV**	72	

Potrjujem resničnost navedenih podatkov.

V/Na LJUBLJANI

Datum 31. 3. 2004

Zahtevam vračilo (obvezno obkrožiti) 03 **NE** **DA**

Podpis

Ime in priimek

6.1 Opredelitev problema

Na izpostave DURS-a po Sloveniji pride vsak mesec nekaj manj kot 70.000 izpolnjenih obrazcev DDV-0. Primer izpolnjenega DDV-0 obrazca je podan na sliki 33. Podatke, ki jih na obrazcu izpolni obvezanec, in predtiskane podatke je potrebno vsak mesec v petih dneh po prejetju vnesti v DURS-ov informacijski sistem. Seveda je to naporno delo, ki ga v večini primerov celo opravljajo inšpektorji, ki so zadolženi in kvalificirani za preverjanje pravilnosti izkazanih podatkov. V fazi priprave, obdelave in vnosa podatkov iz obrazcev sodeluje vsak mesec okoli 120 ljudi, za vnos pa porabijo okoli 3700 delovnih ur. Gre torej za zamudno in naporno delo, za katerega so s pilotskim projektom želeli najti alternativo. Pravzaprav je bil namen pilotskega projekta preizkus tehnologije prepoznavanja obrazcev in smortnosti uporabe le-te. Na osnovi rezultatov pilotskega projekta naj bi se odločili za morebitno vključitev sistema prepoznavanja obrazcev v e-DDV projekt.

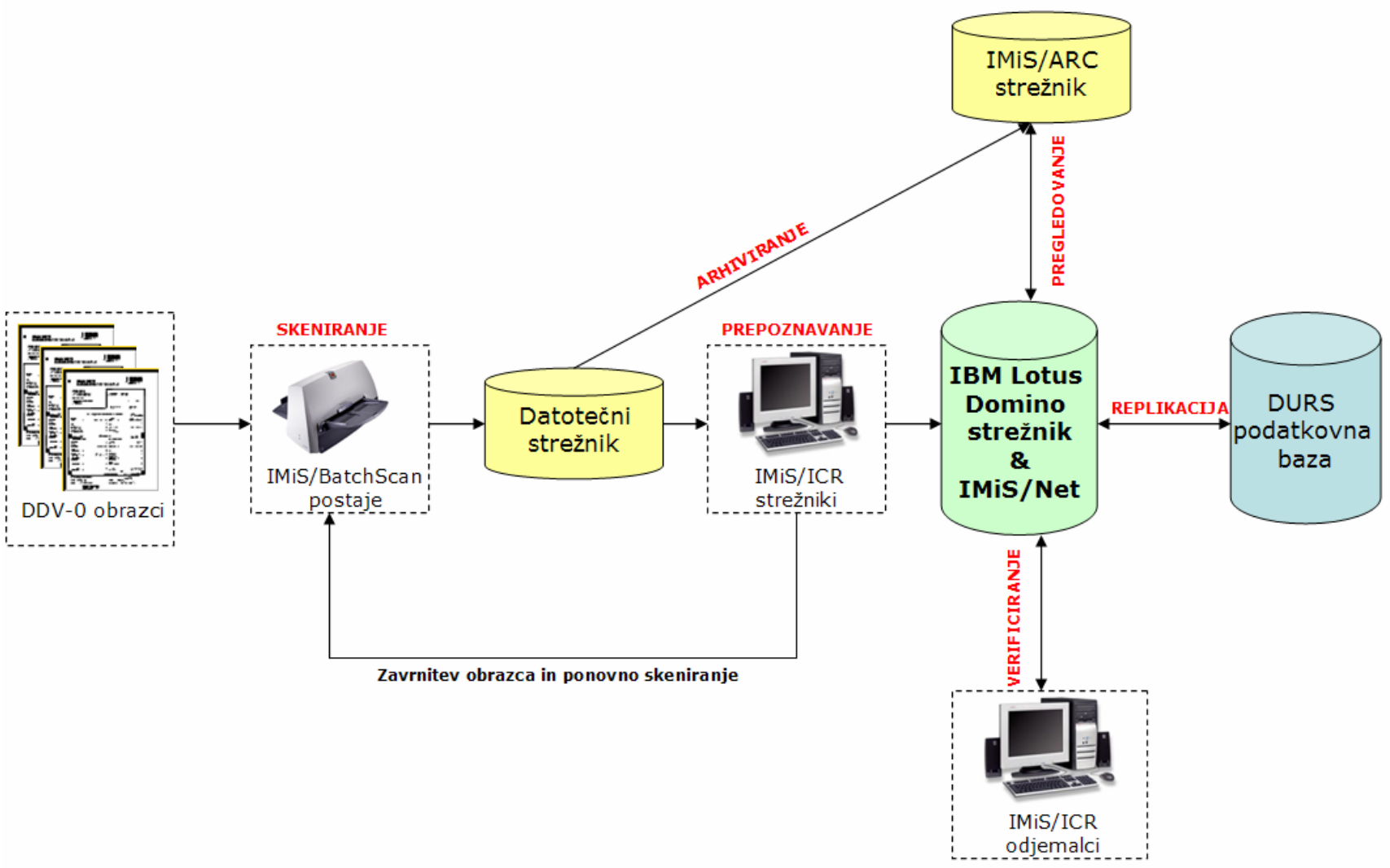
6.2 Izbira gradnikov in zasnova sistema

Projektna skupina, ki se je lotila razvoja idejnega sistema, se je odločila za naslednje gradnike sistema:

- IMiS/Scan odjemalci na skenirnih postajah,
- IMiS/ICR strežnik za prepoznavanje obrazcev,
- IMiS/ICR odjemalec za verifikacijo prepoznanih rezultatov,
- IBM Lotus Domino strežnik za ciljno zbirko podatkov,
- IMiS/ARC strežnik za shranjevanje in arhiviranje slikovnih datotek,
- IMiS/Net modul za pregledovanje originalnih obrazcev prek internetnega brskalnika,
- IMiS/View modul za pregled originalnih obrazcev prek Lotus Notes odjemalca.

Celoten sistem je zasnovan modularno in je shematsko prikazan na sliki 34. Izbrani IMiS moduli so plod lastnega razvoja in razen ICR modulov so uspešno preizkušeni na večini ministrstev in organih v sestavi vlade Republike Slovenije, na upravnih enotah ter tudi v tujini. Pravtako je IBM Lotus Domino platforma standard za dokumentne sisteme v državni upravi, torej je bila izbira s stališča kompatibilnost z ostalimi sistemi v državni upravi upravičena. Koncept IBM Lotus Domino platforme ter na njej zgrajenih aplikativnih rešitev in IMiS sistema za skeniranje in arhiviranje dokumentov torej zagotavlja primerno osnovo za izgradnjo celovitega sistema za upravljanje dokumentov, ki je v skladu s politiko Centra vlade za informatiko

Slika 34: Predvidena shema sistema prepoznavanja obrazca DDV-0



Obrazce iz papirnate oblike v elektronsko s postopkom paketnega skeniranja pretvorijo IMiS/Scan postaje (kasneje je bil za namen paketnega skeniranja razvit poseben modul IMiS/BatchScan). Skenirane datoteke se shranjujejo na datotečni strežnik, oziroma natančneje na izvorno pot, ki jo imajo kot izvor skeniranih obrazcev nastavljeno IMiS/ICR strežniki. Takoj, ko datoteka pride na izvorno pot, jo zazna in da v svoj paket eden od IMiS/ICR strežnikov (pravilo najhitrejšega) in jo prepozna. Kot je bilo opisano, IMiS/ICR strežnik preveri logično pravilnost rezultatov, ki je izvedena s pomočjo 2 kriterijev:

- ali obrazec vsebuje kakšen zavrnjen oziroma neprepoznan znak,
- izračun finančnih formul (podane s strani DURS-a), ki morajo veljati na obrazcu.

Nadaljna pot podatkov je odvisna od rezultata kontrole. V primeru pravilnih rezultatov logične kontrole podatkov, se le-ti izvozijo v Domino podatkovno bazo s pravilnim statusom in so tako vidni le znotraj arhivskega dela podatkovne baze. Ti dokumenti so takoj dosegljivi prek intraneta in pripravljeni za izvoz v transakcijski informacijski sistem DURS-a. V primeru nepravilnega vsaj enega od obeh kriterijev kontrole, pa se podatki izvozijo v isto podatkovno bazo v del za verifikacijo, kjer jih takoj začnejo verificirati operaterji s pomočjo IMiS/ICR odjemalca. Po verifikaciji in korekciji napak se dokumenti avtomatsko prenesejo v arhivski del baze.

Poleg podatkov, ki so rezultat prepoznavanja, se v IBM Lotus Notes evidenčni dokument vpiše tudi kazalec na skenirano datoteko obrazca v elektronskem arhivu in obrazci so prek IMiS/View odjemalca (znotraj Lotus Notes odjemalca) ali IMiS/Net strežnika (v primeru dostopa prek internetnega brskalnika) dostopni vsem pooblaščenim osebam znotraj DURS-ovega razprostranega omrežja v Sloveniji.

Tako lahko pooblaščene osebe poleg podatkov s preprostim klikom pogledajo tudi originalni obrazec, ki ga je izpolnila pravna oseba. Skenirani obrazci se nahajajo sprva na datotečnem strežniku, po prenosu pa na IMiS/ARC arhivskem strežniku s hierarhično arhitekturo. Tovrstni princip kasnejšega prenosa na elektronski arhiv je izveden zaradi ločitve postopka prepoznavanja in arhiviranja, kar omogoči nemoteno in hitrejše prepoznavanje obrazcev. Lokacija skeniranega obrazca je za uporabnika transparentna in nepomembna, saj sistem s podatkom v dokumentu vedno ve, kje se le-ta nahaja.

6.3 Analiza natančnosti in hitrosti na vzorčni množici

6.3.1 Metoda vrednotenja natančnosti

Rezultate prepoznavanja obrazcev smo vrednotili na osnovi dveh kriterijev:

- delež pravilno prepoznanih obrazcev – to pomeni, da vse finančne formule za preverjanje prepoznanih polj, ki so bile podane s strani DURS-a, dajo pravilen rezultat,
- delež pravilno prepoznanih znakov.

Pravilno prepoznani so tisti obrazci, na katerih ni bil noben znak zavrnjen kot neprepoznan ali preveč sumljivo prepoznan in na katerih so logične kotrole dale pozitiven rezultat. Pri tem je potrebno poudariti, da smo prepoznavali v povprečju 120 znakov na obrazec od katerih je bilo v povprečju 80 odstotkov ročno napisanih numeričnih znakov. Pravilno prepoznani znaki so znaki, ki jih sistem ne zavrne kot neprepoznane, oziroma za njih ugotovi dovolj veliko verjetnost pravilne klasifikacije nad pragom.

Predvidevali smo okoli 40 odstotni delež pravilno prepoznanih obrazcev v primeru nespremenjenega obrazca in okoli 60 odstotni delež pravilno prepoznanih obrazcev v primeru spremenjene oblike obrazcev, prilagojene sistemu prepoznavanja obrazcev. 60 odstotni delež se tudi v literaturi pojavlja kot mejni delež, pri katerem je v večini primerov smotrno implementirati sistem prepoznavanja obrazcev. V našem primeru smo delež predvidevali ob 99 odstotni natančnosti na nivoju znaka, kar ob 100 znakih na dokument teoretično pomeni 37 odstotni delež na nivoju dokumenta, vendar smo zaradi predvidene neenakomerne razporeditve napak po dokumentih predvideni delež dvignili na 60 odstotkov.

Skenirali smo pri ločljivosti 300 dpi v črno-beli tehniki. Kontrast in osvetljenost skeniranja smo določili s preizkušanjem na prvi vzorčni množici.

6.3.2 Prvi preizkus na 1440 nespremenjenih obrazcih

Prva testiranja so se izvedla na 1440 realnih obrazcih. Ti obrazci so bili že obdelani s strani DURS-a, kar pomeni, da so bili žigosani in popisani. Tovrstne anomalije so seveda zelo moteče tako za algoritme segmentacije kot za algoritme prepoznavanja. Tak preizkus torej ni primeren za pridobitev realnih natančnosti,

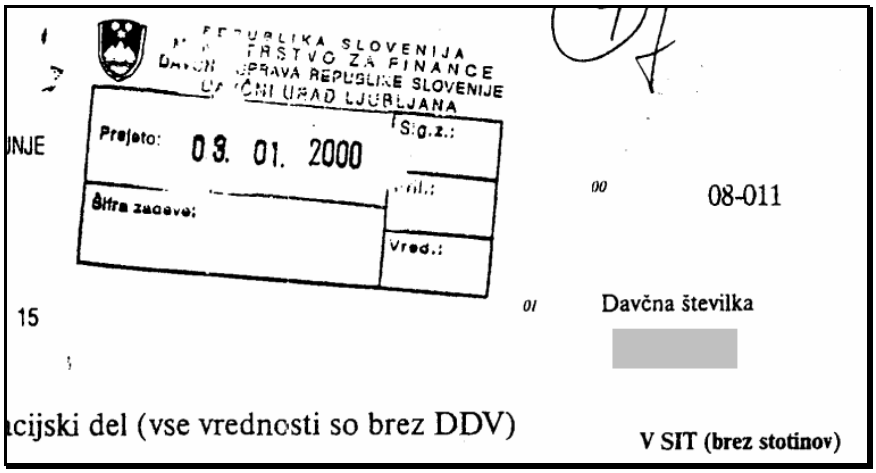
temveč naj služi zgolj kot primerjalni podatek za nadaljne preizkuse. S tem lahko dokazujemo neprimernost obrazcev in pomembnost prilagoditve oblike obrazca.

Rezultati so bili naslednji:

- število vseh obrazcev : 1440
- delež pravilno prepoznanih obrazcev: 557 ali **38,7 %**
- delež dovolj zanesljivo prepoznanih znakov: **94,7 %**

Rezultat prepoznavanja je bil precej slabši od dejanske zmogljivosti celotnega sistema. Razlog je bil predvsem v tem, da so bili vsi dokumenti žigosani, nekateri tudi dodatno popisani (različne »kljukice«). Najbolj moteči so bili žigi MF/DURS s podatkom o datumu prejema obrazca, ki so segali v polje »Davčna številka« zaradi česar je bilo veliko dokumentov, kjer se davčna številka sploh ni prepoznala ali pa se je prepoznala le delno. V nekaterih primerih je prišlo celo do zavrnitve obrazca s strani sistema. Zaradi kontrastno močnih žigov, ki so bili prevladujoči element na obrazcu, je algoritem segmentacije in poravnave izvedel avtomatsko poravnavo obrazca glede na horizontalno zamaknjen žig, kar je pomenilo poševno zamaknjen obrazec na vhodu sistema in kot tak je bil zavrnjen, saj sistem na njem ni našel značilk iz predloge obrazca. Takih primerov obrazcev je bilo kar okoli 5%. Primer žigosanega obrazca, ki je zaradi zamaknjenosti sprožil zavrnitev obrazca, je prikazan na sliki 35.

Slika 35: Primer zavrnjenega obrazca



Žigi so torej povzročili tako napake poravnave in s tem identifikacije obrazca, kot tudi napake segmentacije in prepoznavanja znakov. Vpliv žigov je bil prevelik, da bi rezultate lahko verodostojno obravnavali.

6.3.3 Spremembe vzorčne množice obrazcev in ugotovitve

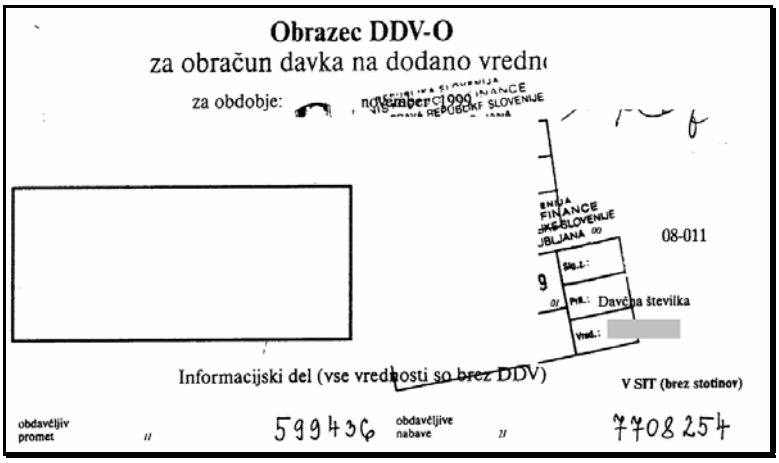
Zaradi neprimernosti tako oblike kot stanja obrazcev, je bilo potrebno naročnika prepričati, da bi sprememba oblike obrazcev in prepoznavanje takih obrazcev brez žigov pomenila bistveno izboljšanje rezultata. Zato smo se odločili za programsko korekcijo skeniranih obrazcev, saj smo programsko skušali očistiti poskenirane obrazce z odstranitvijo žigov in ostalih motečih faktorjev. Del obrazcev, kjer se je v večini primerov nahaja žig, smo programsko odstranili in dodali močno poudarjen pravokotnik, ki je odstranil vpliv žiga na poravnavo obrazca. Seveda nismo korigirali vsakega posebej, temveč smo na vseh obrazcih programsko popravili enako območje. Primer programsko korigirane slike obrazca je prikazan na sliki 36.

Testiranje prepoznavanja na tako popravljenih obrazcih je dalo mnogo boljši rezultat:

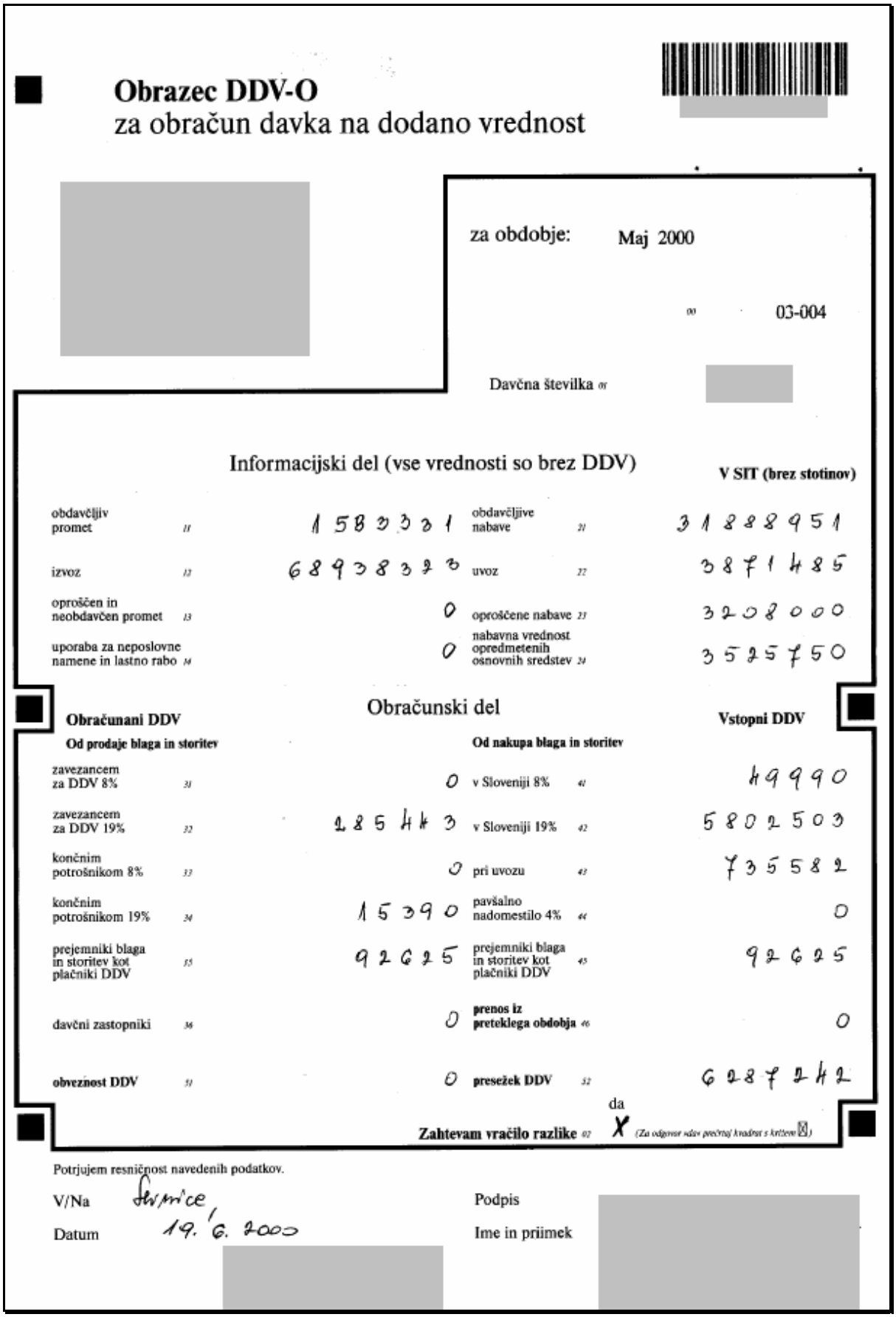
- število vseh obrazcev : 1440
- delež pravilno prepoznanih obrazcev: 876 ali **60,8 %**
- delež dovolj zanesljivo prepoznanih znakov: **98,9 %**

S tem preizkušanjem smo želeli preveriti, kako lahko odstranitev žigov vpliva na rezultat prepoznavanja. Ta problem je namreč predstavljal bistveni razlog, da smo pri prvem preizkušanju dobili tako slabe rezultate. Rezultati tega preizkusa so dejansko dokazovali, da je možno doseči še boljši rezultat, če spremenimo strukturo obrazca in ga priredimo orodju, ki prepoznava obrazec. Naročnik in izvajalec sta se odločila za še en preizkus s spremenjeno obliko obrazca.

Slika 36: Programsko korigiran obrazec za nevtralizacijo vpliva žiga



Slika 37: Preoblikovan obrazec DDV-0



Obrazec DDV-O
za obračun davka na dodano vrednost

za obdobje: Maj 2000

00 03-004

Davčna številka 01 [redacted]

Informacijski del (vse vrednosti so brez DDV)

V SIT (brez stotinov)

obdavčljiv promet	11	1 583 331	obdavčljive nabave	21	31 888 951
izvoz	12	689 383	uvoz	22	387 485
oproščen in neobdavčen promet	13	0	oproščene nabave	23	3 208 000
uporaba za neposlovne namene in lastno rabo	14	0	nabavna vrednost opredmetenih osnovnih sredstev	24	3 525 750

Obračunski del

Obračunani DDV – Od prodaje blaga in storitev			Od nakupa blaga in storitev		Vstopni DDV
zavezancem za DDV 8%	31	0	v Sloveniji 8%	41	49 990
zavezancem za DDV 19%	32	285 443	v Sloveniji 19%	42	5 802 503
končnim potrošnikom 8%	33	0	pri uvozu	43	73 582
končnim potrošnikom 19%	34	15 390	pavšalno nadomestilo 4%	44	0
prejemniki blaga in storitev kot plačniki DDV	35	92 625	prejemniki blaga in storitev kot plačniki DDV	45	92 625
davčni zastopniki	36	0	**prenos iz preteklega obdobja**	46	0
obveznost DDV	37	0	**presežek DDV**	32	6 287 242

Zahtevam vračilo razlike 07 da X (Za odgovor »da« prečrtaj kvadrat s križcem ☒)

Potrjujem resničnost navedenih podatkov.

V/Na Šmarje

Datum 19. 6. 2000

Podpis [redacted]

Ime in priimek [redacted]

6.3.4 Preizkus 1818 obrazcev s prilagojeno obliko

Sprememba oblike obrazcev je pomenila sodelovanje izvajalca, naročnika in izdelovalca obrazcev. Predlog je bil sprejet in testno je bilo izdelanih 2000 obrazcev (za primer spremenjenega obrazca glej sliko 37), ki jih je davčni urad Brežice poslal pravnim osebam na območju izpostave Brežice.

Na obrazcu smo opravili naslednje spremembe:

- polja smo med seboj bolj razmaknili
- oblikovali smo separatorje znakov, ki jih algoritem segmentacije preferira
- dodali smo markerje za relativni izračun lokacije polj
- dodan je bil močno poudarjen okvir, ki določa pravilno poravnavo obrazca

Na dan preizkusa so bili obrazci tako poskenirani kot prepoznani. Pomembno je bilo tudi merjenje porabe časa skeniranja, ki je vključevalo tudi pripravo obrazcev z zlaganjem v pakete in v podajalec. Vseh 1818 obrazcev je bilo na dveh zajemnih mestih poskenirano v 1 uri in 18 minutah, pri čemer sta bila uporabljena dva že nekaj let rabljena skenerja srednjega ranga s hitrostjo cca 15 strani na minuto in s podajalnikom za 50 listov. Skenirali smo z ločljivostjo 300 dpi. Za ICR strežnik je bil uporabljen Compaqov strežnik s taktom 866 MHz, za verifikacijo z ICR odjemalci smo pa uporabili Compaq postaje z 500 oziroma 600 MHz taktom procesorjev.

Zadnje testiranje je zajemalo prepoznavanje 1818 realno izpolnjenih in nepoškodovanih obrazcev. Prepoznavanje s pomočjo IMiS/ICR strežnika je dalo naslednje rezultate:

- čas prepoznavanja na eni delovni postaji, hkrati z validacijo rezultata in izvozom podatkov ter skeniranih obrazcev: **31 minut**
- povprečen čas prepoznavanja, validacije in izvoza podatkov na obrazec: **0,8 sek**
- število vseh obrazcev: 1818
- delež pravilno prepoznanih obrazcev: 1346 ali **74,0 %**
- delež dovolj zanesljivo prepoznanih znakov: **99,6 %**
- črtna koda je bila pravilno prepoznana na vseh obrazcih

V 1 uri in 54 minutah so bili na dveh delovnih mestih z IMiS/ICR odjemalcem popravljeni vsi obrazci z napako prepoznavanja. Potrebno je poudariti, da se je ob verificiranju ugotavljal vzrok napak, način odkrivanja in popravljanja se je v podrobnosti razlagal komisiji, tako da bi bila poraba časa za verificiranje v produkciji krajša. V približno 90 odstotkih primerov je identifikacija in odprava vseh napak na enem obrazcu trajala 10-15 sekund. Zalo veliko je bilo obrazcev, kjer je

zavrnitev celotnega obrazca povzročil le en ali dva nepravilno prepoznana znaka (natančno število takih obrazcev žal ni znano). Največ časa smo porabili za obrazce, kjer smo odpravili vse napake prepoznavanja ali pa napak sploh ni bilo, a logična kontrola vseeno ni bila pozitivna. Takih obrazcev je bilo 42 in so že upoštevani v izračunanih končnih deležih.

6.4 Tehnološka in stroškovna analiza

Rezultati so bili nekoliko nad pričakovanji vseh strani in še enkrat dokazali pravilnost odločitve o preoblikovanju strukture obrazca v smeri doseganja boljših rezultatov. Rezultati dobijo še večjo težo, če jih primerjamo z rezultati popisa prebivalstva v ZDA leta 2000, kjer je njihov sistem avtomatskega prepoznavanja obrazcev pravilno prepoznal 38 odstotkov obrazcev. Seveda je šlo tam za različne oblike obrazcev, a po drugi strani so projekt pripravljali več let in bili z rezultatom zadovoljni, oziroma je bil projekt implementacije sistema prepoznavanja obrazcev uspešen in po njihovi končni analizi smotrn.

Analiza rezultatov je postregla z naslednjimi ugotovitvami (z upoštevanjem pričakovanega obsega 70.000 obrazcev):

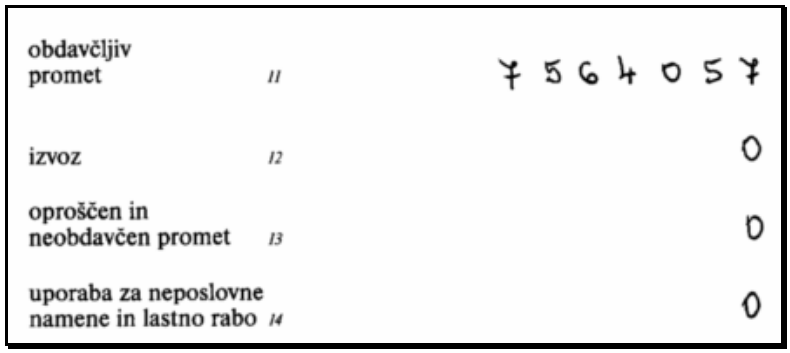
- Skeniranje bi trajalo 20 delovnih ur (okoli 7 ur v primeru treh postaj za skeniranje). Višji rang skenerjev in s tem hitrejše skeniranje ni potrebno. Bolje je zaradi kontrole izpada imeti več slabših kot en zelo dober skener.
- Prepoznavanje obrazcev bi trajalo 16 delovnih ur (6 ur v primeru treh IMiS/ICR strežnik postaj). Hitrost prepoznavanja se z višanjem zmogljivosti procesorja in dodatni prireditvi oblike obrazca povečuje.
- verifikacija 30 odstotkov nepravilno prepoznanih obrazcev bi trajala okoli 130 ur (okoli 13 ur v primeru 10 operaterjev na IMiS/ICR odjemalcih). Čas verifikacije se z višanjem zmogljivosti procesorja in dodatni prilagoditvi oblike obrazca manjša.

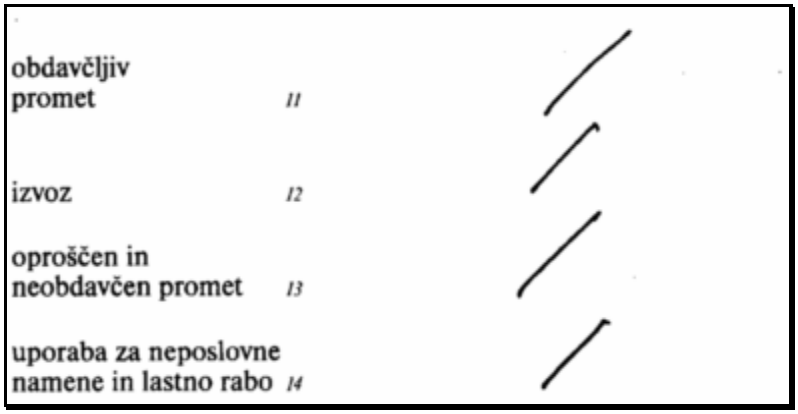
Ker vse tri opisane aktivnosti v procesu potekajo vzporedno, lahko govorimo o 150 delovnih urah potrebnih za skeniranje, prepoznavanje in verifikacijo vseh 70.000 obrazcev, kar pomeni skoraj **25-kratni prihranek časa** v primerjavi z 3700 urami, ki jih porabijo sedaj za pripravo in vnos podatkov. Ker je proces prepoznavanja obrazcev predvsem procesorsko potraten proces, se z višanjem zmogljivost procesorjev čas prepoznavanja še zmanjša. V času od pilotskega projekta do danes so se izboljšali algoritmi segmentacije obrazca in prepoznavanja znakov, kar bi še povečalo natančnost in zmanjšalo čas celotnega procesa. Pravgotovo enomesečna analiza in testiranja niso privedla do optimalne oblike obrazca, tako da bi dodatne analize in spremembe tako obrazca kot oblike referenčnih podatkov,

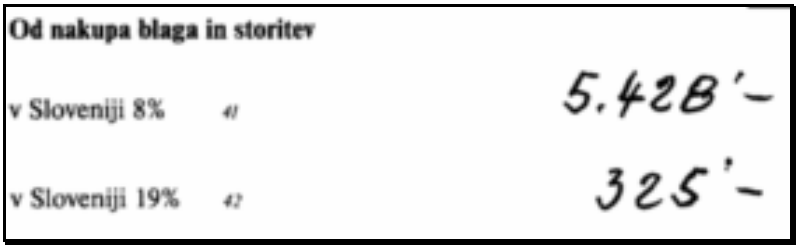
izboljšale natančnost prepoznavanja in s tem zmanjšale odstotek obrazcev, ki ne ustrezajo logični kontroli.

Nadalje bi boljše rezultate dosegli z usmerjanjem izpolnjevalcev na določena pravila pisanja, ki pripomorejo natančnejšemu prepoznavanju. Na obrazec bi npr. natiskali primere željenih oblik vseh cifer. Predvsem je to pomembno za cifre 4 in 7. Uporabnike bi morali opozorit, da naj v primeru, da zneska v polju ni,pustijo polje prazno. Prva dva primera na sliki 38 bi bila pravilno prepoznana obrazca, če bi bila polja brez zneskov prazna. Pravtako naj ne vpisujejo nikakršnih ločil, predpon ali pripon. V zadnjem primeru na sliki 38 so ločila in pripone povzročile napako neprepoznanega znaka. Groba ocena je, da bi v primeru upoštevanja tovrstnih pravil dosegli 80 odstotno natančnost na nivoju obrazca.

Slika 38: Primeri napačno vpisanih izpolnjenih obrazcev







Zaradi decentraliziranega postopka pridobivanja obrazcev (po davčnih uradih) bi bila primerna decentralizacija sistema s postavitvijo dveh ali več ekvivalentnih sistemov na večjih davčnih uradih, s čimer lahko razporedimo obremenitev med sistemi, tak redundantni pristop pa omogoča tudi nemoteno delo v primeru izpada enega od sistemov. Nadalje bi bilo smotrno na vseh ali vsaj na večjih davčnih uradih postaviti skenirna mesta, da ne bi bilo potrebno pošiljati v centralo napovedi v papirni, temveč že v elektronski obliki. Na ta način bi se zmanjšali čas in stroški transporta.

Zmanjšanje števila delovnih ur v primeru DURS-a sicer ne bi neposredno pomenilo zmanjšanje stroška dela inšpektorjev, kajti ti bi ostali zaposleni, bi pa omogočalo skoraj 25 odstotno povečanje razpoložljivega delovnega časa inšpektorjev za kontrolo prijavljenih DDV podatkov. Lahko govorimo o večji učinkovitosti in produktivnosti delovnih mest inšpektorjev. Poleg inšpektorjev v procesu vpisovanja podatkov sodelujejo tudi honorarni sodelavci in njih število bi se zelo zmanjšalo, kar v tem segmentu dela pomeni neposredno zmanjšanje stroška dela. Poudariti je potrebno, da vsi obrazci niso naenkrat na voljo in prihajajo v obdobju treh dni, zato se obremenjenost sistema ustrezno razporedi.

Čeprav je že 25-kratni prihranek delovnih ur dovolj velik razlog za vpeljavo sistema, pa premislimo še o drugih stroškovnih prihrankih in koristih. Prihranke lahko iščemo v strošku delovnih prostorov in materiala, saj z novim sistemom potrebujemo le še prostor za 10 delovnih mest in nekaj strežnikov. S samo implementacijo sistema prepoznavanja obrazcev bi se zaradi fundamentov, ki izhajajo iz dokumentnega sistema vzpostavila infrastruktura za izdelavo dodatnih dokumentnih rešitev DURS-a in celotnega Ministrstva za finance (medtem je v letu 2003 Ministrstvo za finance že implementiralo dokumentni sistem na platformi IBM Lotus Domino in IMiS za skeniranje in elektronski arhiv). Poleg tega je sistem modularen in nastavljiv ter enostavno in hitro razširljiv na prepoznavanja obrazcev za napoved dohodnine in ostalih formularjev davčne uprave. Za potrebe prepoznavanja DDV-0 obrazcev bi sistem deloval le 3 do 5 dni na mesec, kar pomeni veliko »prostega časa«, ki bi ga lahko zapolnili za potrebe prepoznavanja drugih vrst obrazcev. Glede na število različnih obrazcev v državni upravi imamo na voljo številne delovne procese z obrazci, v katere bi enostavno vključili prepoznavanje obrazcev in s tem omogočili preoblikovanje procesa ter dodatne prihranke. Ne gre pozabiti tudi na večjo učinkovitost preoblikovanih procesov, h kateri stremi vladni projekt reforme državne uprave. Še enkrat se bom vrnil na že omenjeni projekt elektronske oddaje dohodnine. Napoved za odmero dohodnine za leto 2003 je prek sistema edavki oddalo 16.745 davčnih zavezancev oziroma 1,45 odstotka vseh pričakovanih dohodninskih napovedi. Pričakovanih je bilo vsaj 50.000 elektronsko oddanih napovedi, saj je bilo do sredine marca v Sloveniji

skupaj izdanih preko 100.000 kvalificiranih elektronskih potrdil, ki so pogoj za elektronsko oddajo dohodnine. Bolj kot 1,45 odstotka vseh, je zaskrbljujoč podatek, da je elektronsko pot uporabilo le okoli 10 odstotkov tistih, ki imajo izpolnjene vse tehnične pogoje za oddajo. Znane so bile težave s kar tremi javnimi razpisi za izvedbo tega projekta in na koncu je državo projekt stal 318 milijonov tolarjev (Kučić, 2004). S tako nizkim odstotkom se investicija v e-dohodnino (kar pomeni seveda tudi v e-ddv) ne bo povrnila še dolgo. Prepričan sem, da bi se v primeru vzporedne uporabe sistema prepoznavanja obrazcev znotraj projekta edavkov (seveda sam trdno zagovarjam tudi možnost elektronske oddaje davkov) investicija povrnila mnogo prej.

Kot primer naj navedem, da so se za hibridno rešitev vzpostavitve tako sistema za prepoznavanje obrazcev, kot možnosti elektronske oddaje podatkov kakor tudi elektronske izmenjave podatkov prek XML standarda, odločili tudi na davčnem uradu Kalifornije in po treh letih se je še vedno velika večina obrazcev obdelala s pomočjo sistema prepoznavanja obrazcev (Lechner, 2002). Davčni urad Lee County na Floridi je leta 1999 uvedel celotni dokumentni sistem za prihajajoče davčne obrazce vključno s sistemom prepoznavanja obrazcev. Slednji se je izkazal kot kritični in najpomembnejši del celotnega sistema, ki omogoča največje prihranke. Celotna investicija v višini 250.000$ je bila povrnjena v treh do štirih mesecih po začetku delovanja (Jones, 2000). Po lastni oceni bi se vrednost investicije celotnega projekta edavkov zaradi sistema prepoznavanja obrazcev povečala največ za 20 odstotkov, celotna investicija v projekt pa bi bila povrnjena v manj kot treh letih.

Komisija DURS-a je podala pozitivno poročilo o poteku in rezultatih idejnega projekta in sprejela tehnologijo prepoznavanja obrazcev kot dovolj zrelo za uporabo. Kasnejši razpisi edavkov tehnologije prepoznavanja obrazcev niso vsebovali.

7. SKLEP

Sistemi za prepoznavanje obrazcev omogočajo velike stroškovne in časovne prihranke v procesu prenosa podatkov iz papirne oblike v sisteme za upravljanje baz podatkov. Prvi korak je skeniranje in arhiviranje obrazcev, zato je prepoznavanje obrazcev vpeto v sisteme za upravljanje z dokumenti. Opisal sem osnove dokumentnih sistemov, nato sem obdelal tehnologije optičnega prepoznavanja znakov in lastnosti sistemov za prepoznavanje obrazcev. Na osnovi opisanih lastnosti sem analizirali stroškovne prihranke in ostale koristi ter opisano analiziral na praktičnem primeru prepoznavanja obrazca za prijavo davka na dodano vrednost (obrazec DDV-0) Davčne uprave Republike Slovenije.

Implementacije tovrstnih sistemov se je potrebno lotiti projektno, saj je potrebno primerno analizirati obrazec, ga preoblikovati v smeri boljše segmentacije in natančnosti ter analizirati morebitne ostale spremembe v procesu, ki bi izboljšale rezultate prepoznavanja. Pravtako je potrebno postaviti primeren arhivski sistem za arhiviranje elektronske (skenirane) oblike obrazcev, ki naj omogoča dostop do skeniranih originalnih obrazcev v realnem času.

V zadnjem času je zaznati porast elektronske izmenjave podatkov in tudi elektronskih obrazcev, kar naj bi prineslo zmanjšanje uporabe papirja za prenos informacij. Vendar analize kažejo, da se uporaba papirja ne zmanjšuje in je še vedno daleč najpogosteje uporabljen medij za prenos informacij. V večini primerov je navkljub sistemu elektronskih obrazcev ali celo elektronske izmenjave podatkov smiselno vzporedno postaviti še sistem za prepoznavanje obrazcev, s čimer obvladujemo vse izvore podatkov in tako omogočamo učinkovitejše procese, čas povrnitev investicije je pa relativno kratek.

Analiza praktičnega primera je nakazala velike stroškovne prihranke, ki jih omogoča vpeljava sistema za prepoznavanje obrazcev in s tem pokazala, da je vpeljava sistema smotrna. Implementacija sistema za obrazec DDV-0 bi pomenila postavitev infrastrukture, ki bi omogočala tako vzpostavitev in širitev dokumentnega sistema kot tudi implementacijo prepoznavanja različnih obrazcev Davčne uprave Republike Slovenije in drugih organov državne uprave.

8. LITERATURA IN VIRI

Literatura:

1. Avedon Don M.: Introduction to Electronic Imaging. AIIM Catalog No. C125, 1995. 214 str.
2. Gingrande Arthur: Forms Automation: From ICR to E-Forms to the Internet. AIIM, 2000, 155 str.
3. Gingrande Arthur: Introduction to Forms Processing. e-doc. September/Oktober 2003, str. 36-39
4. Haverson Debra: Move Beyond Manual Data Entry. Transform Magazine. Julij 2002. str 19-24
5. Hayes Mark: Are Paper Forms Disappearing?. e-doc. Januar/Februar 2001. str 29-32
6. Jain Anil K., Mao Jianchang: Artificial Neural Networks: A Tutorial. IEEE, Marec 1996. str. 31-44
7. Jones Clayton: A Taxing Situation. e-doc. Marec/April 2000. str 41-44
8. Kampffmeyer Ulrich: E-Documents – It' All Legal, Or Is It?. e-doc. September/Oktober 2000. str. 31-34
9. Koulopoulos Thomas, Frappaolo Carl: Electronic Document Management Systems : a portable consultant. McGraw-Hill, 1995. 135 str.
10. Kovšča Blaž: Po 1. marcu bo »vse« drugače: E-oddaja dohodnine. Delo, Ljubljana, 23.2.2004, str. 10
11. Kučić Lenart J.: Kje so koristi e-dohodnine?. Delo, Ljubljana, 3.4.2004, str. 5
12. Lechner Bruce, Oliver Randy: A Taxing problem. e-doc. Januar/Februar 2002. str. 52-55
13. Lunt Penny: New directions in Data Capture. Imaging & document solutions, Vol 9, Number 3, Marec 2000. str. 26-37
14. Mantelman Lee: Mantelman’s Imaging buyers Guide. Flatiron Publishing, 1995. 576 str.
15. McCarthy Anne S.: Image Character Recognition. AIIM Catalog No. R058, 1994. 102 str.
16. Meyers Jason: Automating Forms at Hallmark. e-doc. Januar/Februar 2003. str. 41-43
17. Mori Shunji, Suen Ching Y., Yamamoto Kazuhiko: Historical Review of OCR Research and Development. IEEE, Vol. 80, No. 7, Julij 1992, str. 1029-1058
18. Nadler Morten, Smith Eric P: Pattern Recognition Engineering, John Wiley & Sons, 1993. 587 str.
19. Rapaport Lowell: OCR finally gets the Recognition it deserves, Imaging Magazine, December 1997

20. Rittmann Rosemarie: High-Speed Data Entry: Automated Invoice Processing At ABB AG. e-doc. September/Oktober 2001. str. 53-56
21. Rogelj Peter: Optično odčitavanje serijskih številk števcev, Diplomska naloga, Fakulteta za elektrotehniko, 1998. 51 str.
22. Schantz Herbert F.: Recognition Technology in the information Industry. The National Federation of Abstracting and Information Services, 1992. 107 str.
23. Schantz Herbert F.: The History of OCR. Recognition Technologies Users Associaton, 1982. 113 str.
24. Spencer Harvey: Automated Forms Processing. Flatiron Publishing, 1996. 247 str.
25. Spencer Harvey: Free-Forms Processing. e-doc. Marec/April 2001. str. 36-39
26. Spencer Harvey: E-Forms. e-doc. November/December 2001. str. 46-52
27. Štrumbl Žarko: Roki hrambe dokumentarnega gradiva – pravni in ekonomski vidiki. DOK_SIS 2004. str. 32V-38V
28. Thompson Chris: Enterprise Content Capture. e-doc. Marec/April 2001. str. 40-42
29. Wang Jin, Machine-Printed Character Recognition with Neural Networks, Wright State University, 1993. 110 str.

Viri:

1. ABBYY Case studies & success stories. [URL:http://www.abbyy.com]
2. AIIM. [URL:http://www.aiim.org]
3. Anisimovich Konstantin, Rybkin Vladimir, Shamis Alexander: Using Combination of structural, feature and raster classifiers for character recognition. Abbyy Software House. (poslano po elektronski pošti), 2001
4. Content Management Online. [URL:http://www.contentmgtonline.com]
5. Finansbank uses ABBYY FormReader for automatic input of credit card application forms [URL:http://www.abbyy.com/articles/1286_SS_Finans.pdf]
6. FineReader Engine-Developer's Guide. Abbyy Software House. (poslano po elektronski pošti), 2004
7. GartnerGroup: State of the Document Technologies Industry: 1997-2003
8. KOFAX: Forms processing solutions. [URL:http://www.kofax.com/forms.asp]
9. KOFAX: The Dynamics of Cost In Document and Data Capture [URL:http://www.kofax.com/learning/whitepapers/ac5_whitepaper_dynamics.pdf]
10. Ponikvar Uroš: Koncept dokumentnih sistemov. (poslano po elektronski pošti), 2003
11. TechInfoCenter. [URL:http://binary.techinfocenter.com]
12. What is forms processing? [URL:http://www.abbyy.com/formreader/?param=30764]

13. Žerko Bine: Imaging & archiving v teoriji in praksi (1995-2003). [URL:http://www.imis.si/Web/IMiSSLO.nsf/e49e2d9e0dcaba38c1256b1200365294/ee36251e0c355962c1256d270029afa4?OpenDocument]