

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Andrej Huč

**Digitalizacija prejete pošte in prejetih
računov**

DIPLOMSKO DELO
VISOKOŠOLSKI STROKOVNI ŠTUDIJSKI PROGRAM PRVE
STOPNJE RAČUNALNIŠTVO IN INFORMATIKA

Ljubljana 2012

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Andrej Huč

**Digitalizacija prejete pošte in prejetih
računov**

DIPLOMSKO DELO

VISOKOŠOLSKI STROKOVNI ŠTUDIJSKI PROGRAM PRVE
STOPNJE RAČUNALNIŠTVO IN INFORMATIKA

MENTOR: prof. dr. Franc Solina

Ljubljana 2012



Št. naloge: 00311/2012

Datum: 15.05.2012

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **ANDREJ HUČ**

Naslov: **DIGITALIZACIJA PREJETE POŠTE IN PREJETIH RAČUNOV**
DIGITALIZATION OF INCOMING MAIL AND INCOMING BILLS

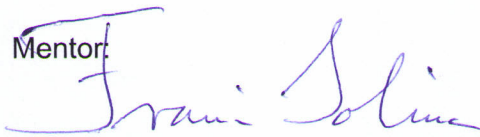
Vrsta naloge: Diplomsko delo visokošolskega strokovnega študija prve stopnje

Tematika naloge:

Opišite problematiko digitalizacije z različnih zornih kotov, kot so razvoj digitalizacije, tehnični in pravni vidiki digitalizacije.

Nato obravnavajte digitalizacijo dokumentarnega gradiva v praksi, predvsem vhodne pošte in vhodnih računov s pomočjo sistema TIS eFlow.

Mentor:


prof. dr. Franc Solina

Dekan:




prof. dr. Nikolaj Zimic

Rezultati diplomskega dela so intelektualna lastnina Fakultete za računalništvo in informatiko Univerze v Ljubljani. Za objavljanje ali izkoriščanje rezultatov diplomskega dela je potrebno pisno soglasje Fakultete za računalništvo in informatiko ter mentorja.

Besedilo je oblikovano z urejevalnikom besedil $\text{\LaTeX}$.

IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisani Andrej Huč, z vpisno številko **63070123**, sem avtor diplomskega dela z naslovom:

Digitalizacija prejete pošte in prejetih računov

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom proc. dr. Franca Soline,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 21. junija 2011

Podpis avtorja:

Kazalo

Povzetek

Abstract

1	Uvod	1
2	Opis digitalizacije	3
3	Razlogi za in proti digitalizaciji	9
4	Zgodovina digitalizacije in OCR-a	13
5	Digitalizacija in OCR danes	15
5.1	Izboljšave, ki pripomorejo k boljšemu OCR-u	15
6	Programska oprema za zajem in digitalizacijo	19
7	Zakonodaja na področju digitalizacije	23
7.1	Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA)	23
7.2	Uredba o varstvu dokumentarnega gradiva (UVDAG)	27
7.3	Zakon o elektronskem poslovanju in elektronskem podpisu (ZEPEP)	28
7.4	Enotne tehnološke zahteve (ETZ)	29

8	Digitalizacija v praksi	31
8.1	Vhodna pošta	31
8.1.1	Digitalizacija vhodne pošte	33
8.1.2	Zahteve za metapodatke	34
8.1.3	Skeniranje pri naročniku in uporaba najete storitve	36
8.1.3.1	Skeniranje	37
8.1.3.2	Pošiljanje skenogramov preko FTP-ja	39
8.1.3.3	Storitev v oblaku	42
8.1.3.4	Podrobnejši opis modulov sistema TIS eFlow	45
8.1.3.4.1	Modul FilePortal	45
8.1.3.4.2	Modul DataCapture	46
8.1.3.4.3	Modul ValidateData	46
8.1.3.4.4	Modul Exceptions	49
8.1.3.4.5	Modul Export	49
8.1.4	Preusmeritev pošiljanja pošte k izvajalcu digitalizacije	51
8.2	Vhodni računi	52
8.2.1	Zahteve za metapodatke	54
8.2.2	Klasifikacija dokumentov v sistemu TIS eFlow	56
8.2.3	Skeniranje in pošiljanje preko FTP-ja	57
8.2.4	Podrobnejši opis modulov sistema TIS eFlow	57
8.2.4.1	Modul DataCapture	58
8.2.4.2	Modul Validation	61
9	Sklepne ugotovitve	63
	Literatura	65

Slike

8.1	Postopek digitalizacije z najemom oblaka storitev.	36
8.2	Prikaz uporabe boljše in slabše ločljivosti skeniranja.	39
8.3	Primer modularne zgradbe aplikacije TIS eFlow.	43
8.4	Primer EFI, prikazan na vročilnici.	47
8.5	Validacija vročilnice v sistemu TIS eFlow.	48
8.6	Model digitalizacije, kjer naročnik ni vpleten v postopek. . . .	52
8.7	Algoritem iskanja in zapolnjevanja metapodatkov s pomočjo podatkovnih baz.	60

Seznam uporabljenih kratic

OCR	Optical Character Recognition, optično prepoznavanje znakov
ZVDAGA	zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih
PDF	Portable Document Format, oblika za prenašanje dokumentov
PDF/A	Portable Document Format / Archive, oblika za elektronsko hranjenje dokumentov
VRS	Virtual Re-Scan
TIS	Top Image Systems
ETZ	Enotne tehnološke zahteve
DMS	Document Management System, sistem za upravljanje dokumentov
FTP	File Transfer Protocol, protokol za prenos podatkov
XML	Exstensible Markup Language, razširljiv označevalni jezik
EFI	Empty Form Information

Povzetek

Namen diplomske naloge je predstaviti postopke digitalizacije gradiva na splošno in ob upoštevanju vseh zakonskih predpisov, omejitev ter priporočil, izdelati celoten proces digitalizacije prejete pošte in prejetih računov kot storitev v oblaku. Dobro obvladovanje dokumentarnega gradiva v organizaciji ima velike pozitivne učinke. Za lažje in učinkovitejše obvladovanje dokumentarnega gradiva pa je treba to gradivo pretvoriti v elektronsko obliko v procesu, imenovanem digitalizacija. Zaradi velikega zanimanja za storitve v oblaku, v nadaljevanju opisujemo digitalizacijo kot storitev v oblaku. Na začetku diplomske naloge so predstavljeni vsi pomembnejši zakoni s področja digitalizacije v Sloveniji, v nadaljevanju pa podajamo dve možni obliki rešitev za najpogostejša primera uporabe digitalizacije. Programska oprema, uporabljena za prikaz primera digitalizacije vhodnih računov in vhodne pošte, je TIS eFlow [17]. Za predstavljena primera, ki predstavljata največjo količino prejetega dokumentarnega gradiva organizacije, podajamo dve različni rešitvi. Obe sta naravnani v obliki rešitve v oblaku. Prva rešitev naročnika vpleta v celoten proces digitalizacije, druga pa je v celoti izvedena s strani izvajalca digitalizacije in naročnik ni vpleten v izvedbo digitalizacije. V diplomski nalogi podajamo opis celotnega postopka digitalizacije z vsemi podrobnejšimi zahtevami za vsak posamezen korak izvedbe.

Digitalization of incoming mail and incoming bills

Abstract

The purpose of this thesis is to provide an overview of the processes involved in the digitalization of materials and having consideration of legal regulations, restrictions and recommendations, make the digitalization process of incoming mail and incoming bills as a cloud-based service. Having an efficient document management system has a significantly positive effect on an organization. In order for documents to be managed more easily and efficiently, they must be converted into an electronic format by a process called digitalization. Due to growing interest in cloud-based solutions, this thesis focuses on digitalization as a cloud-based service. We begin with an introduction to the most relevant laws governing the field of digitalization in Slovenia, followed by the descriptions of the two possible solutions for most important examples of digitalization used in practice. To illustrate the use of digitalization of incoming invoices and incoming mail, we used the TIS eFlow solution [17]. For the two examples described, which account for the larger part of the organization's incoming documents, we propose two different solutions, both of which are cloud-based. The first solution requires the client's involvement throughout the entire digitalization process. The second solution is implemented by the digitalization services provider alone, without the client's involvement. In this thesis, we provide a description of the

entire process of implementation of digitalization, including all the detailed requirements for each step of implementation.

Poglavje 1

Uvod

V času informacijske dobe in močnega razvoja informacijsko-telekomunikacijske tehnologije želimo imeti vse informacije in podatke na dosegu klika. Celotno poslovanje se iz papirnega poslovanja pomika v digitalno oziroma elektronsko poslovanje. Potreba po informacijah se je znatno povečala, pričakovani čas dostopnosti informacij pa se je močno zmanjšal. Da bi zadostili tem in vsem drugim pogojem informacijske dobe, moramo obstoječo fizično obliko pretvoriti v elektronsko, saj fizična oblika informacij ne ustreza potrebam sodobnega tempa življenja. V naši diplomski nalogi opisujemo pretvorbo dokumentarnega gradiva v elektronsko obliko, imenovano digitalizacija, kot poslovni proces v organizaciji. Za digitalizacijo smo se odločili zaradi vse večjega zanimanja za ta proces in zaradi uporabnosti postopka na vseh področjih. Zaradi zelo obsežne širine digitalizacije smo se v diplomskem delu omejili na le dve trenutno zelo aktualni področji, in sicer na digitalizacijo prejete pošte in prejetih računov. V prvem delu diplomske naloge predstavljamo opis in umestitev digitalizacije dokumentarnega in arhivskega gradiva. Opisana sta dva večja projekta, ki trenutno potekata na področju digitalizacije arhivskega gradiva in knjig. V nadaljevanju podajamo razloge za in proti digitalizaciji, začetke digitalizacije v zgodovini ter primerjavo z digitalizacijo v današnjem času. Sledi poglavje o brezplačni in plačljivi programski opremi, ki je trenutno vodilna na tržišču digitalizacije. Zelo pomembno poglavje

na področju digitalizacije je poglavje o z digitalizacijo povezani zakonodaji. Naštevamo vse pomembnejše zakone in opišemo, kako jih moramo pri procesu digitalizacije izpolnjevati in upoštevati.

V glavnem delu diplomske naloge sta opisani in izdelani rešitvi digitalizacije prejete pošte in prejetih računov, ki predstavljata največji del poslovnega dokumentarnega gradiva vsake organizacije. Za obliko rešitve smo si izbrali trenutno priljubljeno storitev v oblaku. Opisali in izdelali smo celoten postopek, od začetka digitalizacije do končne oddaje digitaliziranih podatkov naročniku. Oba postopka sta na področju digitalizacije novejša in redkeje uporabljena. Tako smo na podlagi upoštevanja vseh dejavnikov kot so zakonodaja, potrebe organizacij itd. izdelali obe rešitvi od začetka do konca. Rešitvi sta dober primer digitalizacije vhodne pošte in vhodnih računov, kateri lahko glede na specifične potrebe prilagodimo.

Poglavje 2

Opis digitalizacije

Digitalizacija je proces spreminjanja objektov iz analogne v digitalno obliko. Pojem digitalizacija je zelo širok in splošen, zato smo se v našem diplomskem delu omejili na digitalizacijo dokumentarnega gradiva, kar področje digitalizacije nekoliko zoži. Kljub temu digitalizacija dokumentarnega gradiva zajema veliko procesov, ki morajo biti izpolnjeni pred in med pretvorbo dokumentarnega gradiva v digitalno obliko. Veliko ljudi zmotno meni, da je digitalizacija dokumentarnega gradiva samo skeniranje in zajem podatkov. Vendar je to le del digitalizacije, ki ga spremljajo še drugi, zelo pomembni procesi.

Digitalizacijo lahko razdelimo na tri sklope. Prvi sklop je priprava gradiva na skeniranje. V tem sklopu je treba gradivo skladno z zakonodajo evidentirati, dodeliti (signirati), razvrstiti (klasificirati) in pripraviti na skeniranje. Drugi sklop je skeniranje oziroma pretvorba v digitalno obliko. Tretji sklop, ki je navadno povezan z digitalizacijo, pa je zajem metapodatkov in optična prepoznavna dokumentacije oziroma OCR (Optical Character Recognition). Vsi trije sklopi so med sabo tesno povezani in jih moramo enakovredno upoštevati. Pogosto se predvsem prva dva sklopa nehote zapostavljata in pozabljata. Arhiv Republike Slovenije natančno opredeljuje postopke digitalizacije. Teh postopkov se morajo dosledno držati vse javnopravne osebe, medtem ko so za druge poslovne subjekte ti postopki priporočljivi. Kasneje

bomo navedli tudi pomembnejše zakone, ki urejajo področje digitalizacije.

Digitalizacija dokumentarnega gradiva v zgodovini ni imela velikega pomena. Slab razvoj digitalizacije in redka uporaba digitaliziranega dokumentarnega gradiva sta bila posledica slabe informacijske in zakonske podpore. Da bi lahko zagotovili uporabnost digitaliziranih dokumentov, potrebujemo dobro informacijsko in zakonsko podporo. Pri digitalizaciji nastaja velika količina podatkov, ki jih moramo hraniti in z njimi upravljati v ustreznih dokumentnih sistemih. V preteklosti ni bilo programske opreme, ki bi zadostila vsem potrebnim pogojem kakovostne digitalizacije. Posledično so bili digitalizirani dokumenti slabše berljivi in večje velikosti. Z razvojem informacijskih sistemov, še posebej z razvojem dokumentnih sistemov, pa je digitalizacija dokumentarnega gradiva postala pomembnejša. Poleg tega je leta 2006 Arhiv Republike Slovenije izdal Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA) [1], ki je temelj digitalizacije in upravljanja dokumentarnega gradiva.

V preteklosti se je digitaliziralo predvsem pogosto uporabljeno gradivo. Med to gradivo lahko štejemo razne publikacije, slike, pogosto uporabljene dokumente itd. Danes digitalizacijo uporabljamo na vseh področjih družbe. S pomočjo digitalizacije dokumentarno gradivo postane lažje dostopno in bolj uporabno. Najpomembnejše pa je zniževanje stroškov, povezanih s hrambo, iskanjem in upravljanjem fizičnega dokumentarnega gradiva. Digitalizacija, skupaj z drugimi spremljevalnimi storitvami in sistemi za vodenje ter hrambo digitalnega dokumentarnega gradiva, postaja pomemben del gospodarstva.

Glede na delitev Zakona o varstvu dokumentarnega in arhivskega gradiva ter arhivih [1], ki bo predstavljen v nadaljevanju in je temeljni zakon na področju digitalizacije, poleg dokumentarnega gradiva poznamo tudi arhivsko gradivo. Po tem zakonu je arhivsko gradivo dokumentarno gradivo, ki ima trajen pomen za znanost in kulturo ali trajen pomen za pravno varnost oseb v skladu s strokovnimi navodili pristojnih arhivov. Arhivsko gradivo je lahko izvirno in reproducirano pisano, risano, tiskano, fotografirano, filmano, optično ali magnetno, kakor koli zapisano dokumentarno gradivo, nastalo

pri delu pravnih oziroma fizičnih oseb [2]. Arhivsko gradivo razdelimo na javno arhivsko gradivo, filmsko arhivsko gradivo in zasebno arhivsko gradivo. Področja bodo podrobneje opisana v nadaljevanju. Največ arhivskega gradiva nastane ob pomembnih dogodkih, kot so na primer svetovne vojne, osamosvojitve države, predsedovanje Evropski uniji itd.

Prav tako ima digitalizacija arhivskega gradiva velik pomen v gospodarstvu in kulturi. Z digitalizacijo dosežemo večjo dostopnost gradiva, predvsem večjo dostopnost zasebnega arhivskega gradiva. Fizično arhivsko gradivo je navadno v lasti ene osebe ali organizacije. Brez uporabe digitalizacije bi gradivo poznali samo določeni posamezniki, z digitalizacijo pa lahko to gradivo preprosto in poceni javno delimo in prikazujemo.

Primer digitalizacije zasebnega arhivskega gradiva z namenom povečevanja razpoznavnosti in širjenja informacij je projekt Europeana 1914–1918 [3]. Projekt je nastal na pobudo univerze v Oxfordu, ki je Britance pozvala, naj prinesejo pisma, slike in druge spomine, da bi jih digitalizirali in javno objavili. Po uspehu na lokalni ravni se je projektu pridružila še Europeana (Evropski digitalni arhiv, knjižnica in muzej) [4], ki je k sodelovanju z univerzo v Oxfordu povabila še druge države. Projektni partner je med drugimi tudi Narodna in univerzitetna knjižnica Slovenije.

Drugega večjega projekta digitalizacije arhivskega gradiva z imenom Google Books (knjige) [5] pa se je lotilo podjetje Google. Namen projekta je digitalizacija knjižničnega gradiva in s tem olajšanje iskanja knjig na spletu. Gre predvsem za knjige, ki bi jih drugače težko našli, to so recimo knjige, ki jih ne tiskajo več. Seveda pa je v tem primeru treba upoštevati avtorske pravice. Če gradivo ni več zaščiteno z avtorskimi pravicami, spletni prikazovalnik omogoči ogled celotne knjige. Če pa je gradivo še vedno zaščiteno z avtorskimi pravicami, se prikažejo le podatki o gradivu in nekaj odlomkov ter povezave na prodajna mesta gradiva ali knjižnice, ki gradivo imajo.

Zaradi digitalizacije fizično arhivsko gradivo kljub večkratnemu pregledovanju ostane nepoškodovano. Veliko fizičnega arhivskega gradiva je namreč v zelo slabem fizičnem stanju, včasih je že bolj ali manj poškodovano. Stare li-

stine in fotografije so zaradi starosti zelo občutljive. Večkratno pregledovanje tega gradiva bi tako poslabšalo že tako slabo fizično stanje.

Digitalizacija arhivskega gradiva je sicer podobna digitalizaciji dokumentarnega gradiva, vendar ima določene posebnosti. Pri digitalizaciji dokumentarnega gradiva poskušamo digitalizirati vse pomembne informacije, nepomembne pa poskušamo odstraniti. Pomembna informacija v primeru dokumentarnega gradiva je besedilo, ozadje dokumenta pa je manj pomembno oziroma nepomembno. Tako je dokument lažje berljiv, znaki so bolj optično prepoznavni, končna velikost datoteke pa je manjša. Pri digitalizaciji dokumentarnega gradiva imajo velik pomen metapodatki, saj se na podlagi le-teh išče in upravlja digitalno dokumentarno gradivo. Digitalizacija arhivskega gradiva pa navadno zahteva točne detaile gradiva. Arhivsko gradivo je tudi veliko bolj občutljivo, kar moramo pri ravnanju z njim upoštevati. Digitalizacija arhivskega gradiva glede na tip zahteva naslednje specifikke:

- Dokumentarno gradivo oziroma papirno gradivo: med papirno gradivo prištevamo različne listine, knjige in drugo knjižno gradivo. Starejše gradivo je pogosto slabo ohranjeno in zelo občutljivo. Vezane knjige skeniramo v obstoječi obliki in jih ne razvežemo, gradivo mora ostati v prvotni obliki. Za skeniranje gradiva moramo uporabiti strojno opremo, ki ne poškoduje gradiva. Po navadi za papirno gradivo uporabljamo ploskovne knjižne skenerje; ti ne oddajajo toplote, ki bi lahko poškodovala gradivo. Pri delu z gradivom uporabljamo posebne bombažne rokavice, da se med skeniranjem na gradivo ne bi prenesle snovi z naših rok, kot na primer pot, kisline itd. Kakovost skeniranja mora biti na visokem nivoju, saj so podrobnosti pomembne. Poznamo več različnih detajlov kakovosti, kot so: pravilnost zajetih barv, pravilna svetilnost, ločljivost itd. Postopki za zagotavljanje kakovosti so zahtevni in zahtevajo dobro in drago strojno in programsko podporo. Tudi prostor, v katerem skeniramo, mora biti ustrezno opremljen. Po beljen mora biti s posebno barvo, ki ne odbija svetlobe. Vsa boljša strojna oprema za skeniranje podpira nastavitve pravilnih profilov, ki

omogočajo najboljši približek digitalne verzije dokumenta fizični dokumentaciji.

- Fotografije: fotografije imajo podobne lastnosti kot papirno gradivo. Pomemben dodatek je odpravljanje odseva med skeniranjem. Poleg tega moramo fotografijo pred skeniranjem očistiti, saj so na njej po navadi razni delci in prstni odtisi.
- Slike in plakati: v primeru slik in plakatov moramo še bolj paziti na pravilnost zajetih barv, saj so te najbolj pomembne. Dodatna ovira pri skeniranju slik in plakatov je njihova velikost, pogosto namreč presegajo velikost A0.
- Drugo gradivo: poznamo še veliko drugih oblik gradiva, na primer filmske negative, diapozitive, kasete VHS, mikrofilme in podobno [30]. Skeniranje tega gradiva ni bistveno drugačno od prej opisanega, specifični sta le strojna in programska oprema. Drugačen je tudi način preverjanja in zagotavljanja kakovosti skeniranja.

V nasprotju s pomembnostjo metapodatkov pri dokumentarnem gradivu, v primeru arhivskega gradiva ti niso ključnega pomena. V primeru arhivskega gradiva so veliko pomembnejši naslednji podatki:

- avtor gradiva,
- čas nastanka gradiva,
- kraj nastanka oziroma objave gradiva,
- namen nastanka gradiva (na primer sklenitev premirja v svetovni vojni),
- opis gradiva (na primer, katere osebe so prikazane na sliki).

Digitalizacija se prav zaradi pocenitve tehnologije vse pogosteje uporablja tudi v zasebne namene. Veliko ljudi želi svoje fotografije digitalizirati, saj so tako bolj obstojne in jih lažje urejajo. Fotografijo v digitalni obliki lahko

preprosto pošljemo drugi osebi ali jo dodamo v spletni album. Fizično gradivo se z leti uničuje in postaja slabše kakovosti, medtem ko v digitalni obliki ves čas ostaja nespremenjeno.

Vse bolj priljubljena je tridimenzionalna (3D) digitalizacija. V primeru 3D-digitalizacije lahko objekte digitaliziramo z vseh strani in tako dobimo popoln 3D-digitalni objekt. 3D-digitalizacijo sta omogočili napredna strojna in programska oprema. Objekt, skeniran s 3D-skenerjem, lahko kasneje pregledujemo v digitalni obliki, prednost tega pa je pregledovanje objekta z vseh strani. Objekt lahko poljubno obračamo in raziskujemo. Trenutno 3D-digitalizacijo uporabljajo ne le v industrijski proizvodnji temveč vse več tudi v kulturi (kiparstvo, arhitektura, arheologija). Prihodnost 3D-digitalizacije je njena uporaba v medicini za izdelavo nadomestnih protez. Zdrav ud skeniramo s pomočjo 3D-skenerja in nato v skladu s skenirano sliko izdelamo nadomestno protezo, ki je zelo podobna zdravemu udu.

Poglavje 3

Razlogi za in proti digitalizaciji

Digitalizacija prinaša veliko pozitivnih učinkov, ki niso zanemarljivi. Med njimi je zmanjševanje količine fizične dokumentacije v arhivih in s tem posledično zmanjševanje stroškov, povezanih z arhiviranjem fizične dokumentacije. Arhiviranje fizične dokumentacije je precej drago, saj je treba zagotoviti prostor, opremo in osebje, ki z arhivom upravlja. Z digitalizacijo lahko te stroške zmanjšamo, saj moramo po zakonu ZVDAGA hraniti le trajno in arhivsko dokumentarno gradivo. Vse drugo gradivo lahko po predpostavkah, ki jih podaja ZVDAGA, po digitalizaciji uničimo. Drugi pozitivni razlogi za digitalizacijo so:

- Posledica zmanjševanja količine fizične dokumentacije je lažje upravljanje arhiva, vendar pa moramo zagotoviti vsa sredstva za vzdrževanje elektronskega arhiva. To je ob dobri podpori informatike preprostejše.
- Iskanje po arhivu je lažje, čas dostopanja do dokumentov je skrajšan. Čeprav je fizični arhiv dobro urejen, se dokumentarno gradivo išče dalj časa. Poleg tega dokumente po navadi išče za to usposobljena oseba (arhivar). Iskanje po elektronskem arhivu je hitro in enostavno. V nasprotju s fizičnim lahko po elektronskem arhivu išče vsak, ki ima pravico do dostopa. Poleg tega nam elektronski arhiv po navadi omogoča iskanje po ključnih besedah (metapodatki) ali celo po celotni vsebini

dokumentacije, če je bila pri pretvorbi opravljena tudi optična prepoznavna znakov (OCR) celotnega dokumenta.

- Lažje in hitreje pošiljanje dokumentov med uporabniki. Če želimo fizični dokument oddati dvema uporabnikoma, moramo narediti kopijo, v primeru pošiljanja dokumenta pa moramo velikokrat uporabljati dostavne službe ali pošto. Tako je pošiljanje dokumenta drago in časovno zamudno. V primeru elektronske verzije dokument preprosto pošiljamo različnim uporabnikom, ne glede na oddaljenost.
- Do elektronske dokumentacije lahko dostopamo, če nam dokumentni sistem to dopušča, od koder koli in kadar koli.
- Elektronskemu dokumentu lahko sledimo in imamo nad njim popoln nadzor, pri fizični dokumentaciji pa tega velikokrat ne moremo zagotoviti.
- Pregledovanje elektronskih dokumentov je hitreje in enostavnejše. Ponostavljeno je predvsem pregledovanje veliko različne dokumentacije hkrati. Predstavljajte si, da bi morali sočasno pregledati in primerjati 15 različnih knjig v fizični obliki.
- Večja poslovna učinkovitost in zmanjševanje stroškov. Vsi zgoraj naštetih dejavniki privedejo do povečane učinkovitosti pri delu z dokumentarnim gradivom, zmanjšanja časa, namenjenega dokumentarnemu gradivu, in zmanjševanja stroškov, povezanih z obvladovanjem dokumentarnega gradiva v fizični obliki.
- Ekološka usmerjenost podjetja. Z uporabo elektronske dokumentacije se izognemo fizičnemu produciranju in reproduciranju dokumentov ter tako poskrbimo za čistejše okolje in neokrnjeno naravo.

Vendar digitalizacija ne prinaša samo pozitivnih učinkov. Dejavniki, ki govorijo proti digitalizaciji, so:

- Postavitev infrastrukture in usposabljanje ljudi za delo z elektronskim arhivom. Vsako podjetje ima fizični arhiv dokumentacije, uporabniki so navajeni na uveljavljeni proces in uporabo fizične dokumentacije. Pri digitalizaciji in uporabi elektronske oblike dokumenta je treba spremeniti celoten poslovni proces. Najprej moramo zagotoviti informacijsko infrastrukturo, skupaj z usposobljenimi administratorji, ki bodo sistem upravljali in vzdrževali. Usposobiti moramo tudi vse uporabnike elektronskih dokumentov, kar v nekaterih primerih lahko povzroča težave, saj uporabniki niso vešči uporabe računalnika.
- Hitro zastaranje tehnologije. Da tehnologija ne bi zastarala, potrebujemo neprestano vlaganje in izboljšave. Preprečiti moramo tudi zastaranje nosilcev digitalnih informacij in formata zapisa informacij.
- Do pred kratkim je bila največja težava neobstoječa ali pomanjkljiva pravna podlaga za digitalizacijo dokumentarnega gradiva.
- Težavno ugotavljanje avtorstva in dodeljevanje avtorskih pravic v primeru arhivskega gradiva. Največje težave se pojavljajo predvsem pri ugotavljanju avtorstva slikovnega gradiva.
- Podjetje mora imeti klasifikacijski in signirni načrt. Veliko podjetij, predvsem srednja in manjša podjetja, ima slabo določeno organizacijsko strukturo in s tem povezano signiranje dokumentov. Prav tako pogosto vidimo slabo razvrščanje dokumentov. Tako se velikokrat zgodi, da za posamezno dokumentarno gradivo ne vemo, komu pripada ali v katero kategorijo oziroma tip bi ga uvrstili.

Z digitalizacijo, še bolj pa z vpeljavo sistema za upravljanje z dokumenti, podjetje lahko pridobi na konkurenčnosti. Hkrati z vpeljavo digitalizacije ima podjetje priložnost definicije poslovnih procesov, kot bi morali potekati, in ne kot potekajo.

Poglavje 4

Zgodovina digitalizacije in OCR-a

Digitalizacija dokumentarnega gradiva in optična prepoznavanje znakov (Optical Character Recognition) sta povezani področji. Digitalizacija kot širši pojem zajema vse storitve, od začetka prejema dokumenta do končne elektronske oblike dokumenta, OCR pa je le del procesa digitalizacije. Navadno se OCR uporablja za avtomatski zajem metapodatkov na dokumentu ali za izdelavo elektronskih dokumentov z možnostjo iskanja po vsebini dokumenta. Najpogostejši primer takih dokumentov so dokumenti PDF, ki omogočajo iskanje po vsebini dokumenta.

Digitalizacija je v računalništvu dokaj star pojem. Vanjo lahko uvrstimo vsak fizični element, pretvorjen v digitalno obliko. Tako lahko med digitalizacijo štejemo digitalno fotografiranje, prepis v digitalno obliko, skeniranje itd. Seveda je digitalizacija tesno povezana z razvojem računalništva. Z razvojem računalništva in strojne opreme je tudi digitalizacija močno napredovala. Prvi skenerji, ki so bili povezani z računalnikom, so bili tako imenovani bobnasti skenerji. Prvega so izdelali leta 1957 v nacionalnem uradu za standarde v ZDA. Skener je imel sposobnost skeniranja 176 slikovnih pik na stran. Danes poznamo skenerje, ki imajo možnost zajemanja 1200 ali več slikovnih pik na palec. Glede na vrsto gradiva se uporabljajo različni specializirani skenerji

kot so 3D skener, skener za diapozitive, občutljivo dokumentarno gradivo itd.

OCR oziroma Optical Character Recognition je zelo širok pojem, zato smo se v našem diplomskem delu omejili predvsem na zajem oziroma prepoznavo znakov iz dokumenta. Proces prepoznavanja znakov je del digitalizacije, ki se ga vse pogosteje uporablja. Vse pogosteje se srečujemo z zahtevami, da mora biti fizično dokumentarno gradivo spremenjeno v elektronsko obliko z dodanim OCR-om vsebine dokumenta.

Zgodovina OCR-a je zelo dolga. OCR je bil prvotno namenjen lažšanju življenja slepih in slabovidnih. Leta 1914 je Emanuel Goldberg razvil stroj, ki je prebral znake in jih pretvoril v standardno telegrafsko kodo. V istem času je Edmund Fournier d'Albe izdelal napravo, ki je lahko prebrala znake in jih spremenila v zvok. Leta 1949 so inženirji RCA (Radio Corporation of America) prvič uporabljali primitivni OCR, ki je temeljil na računalniku. Želeli so prebrati znake s pomočjo računalnika in jih nato predvajati kot črke. Zaradi visokih stroškov se projekt ni nadaljeval. Leta 1965 sta podjetji Reader's Digest in RCA izdelali OCR, ki je bral serijske številke na vnaprej natisnjenih oglaševalskih kuponih. Hitrost branja je bila okrog 1500 dokumentov na minuto. Pošta Združenih držav Amerike OCR od leta 1965 uporablja za razvrščanje pošilk [6].

Poglavje 5

Digitalizacija in OCR danes

Dandanes je trend OCR 100% prepoznavna znakov. Četudi uporabljamo naj-novejšo tehnologijo, tega ne moremo doseči. Avtomatski zajem latinskih znakov doseže med 71% do 98% pravilnosti [6], odvisno od kvalitete skenirane slike. 100% točnost lahko zagotovimo šele po človeškem pregledu prepoznanih znakov. V primeru ročne pisave zanesljivost OCR močno pade. Ravno tako je v primeru Vzhodnih jezikov kot je Kitajščina, Arabščina itd. OCR precej slabše kvalitete. Ravno tako včasih najdemo težave v primeru Slovenščine, saj uporabljamo šumnike, kateri so bili do nedavnega težko prepoznani. Zaradi vseh možnih napak pri prepoznavanju znakov dokumenta, moramo v elektronskem arhivu vedno imeti originalno skenirano elektronsko verzijo dokumenta, katera mora biti identična originalnemu dokumentarnemu gradivu. V primeru slabše prepoznavne znakov lahko na podlagi originala elektronskega skeniranega dokumenta še vedno pridemo želenih informacij.

5.1 Izboljšave, ki pripomorejo k boljšemu OCR-u

Proces OCR ni omejen le na prepoznavanje znakov, ampak se začne že s samim skeniranjem in zajemanjem slike. Boljšo sliko dobimo pri skeniranju, boljši končni rezultat OCR-a lahko pričakujemo. Prav izboljšanje skeniranja

in s tem povezana boljša kakovost skenirane slike je najboljša optimizacija, ki pripomore k boljšemu OCR-u. Zaradi tega danes večina komercialnih programov za zajem slik oziroma skeniranje uporablja različne operacije za izboljševanje kakovosti slike dokumenta. V uporabi je več različnih tehnik izboljšav dokumentov, kot so:

- Avtomatsko odstranjevanje ozadja: program za zajem slike avtomatsko analizira in odstrani ozadje dokumenta. Skenirani dokument je brez ozadja, OCR je opravljen samo na besedilu oziroma ostanku dokumenta. Zaradi ozadja na sliki pogosto nastanejo pike in nečistoče, ki kasneje znižajo kakovost OCR-a. Poleg tega je elektronski dokument brez ozadja manjši in zasede manj prostora.
- Avtomatsko obračanje dokumenta: program za zajem slike avtomatsko analizira dokument in ga obrne glede na usmerjenost pisave. Pri samem OCR-u potem nimamo težav zaradi napačno obrnjenega dokumenta, zato je kakovost OCR-a višja. Težava pri avtomatskem obračanju dokumenta nastane, kadar je del pisave obrnjen drugače kot ostali del dokumenta. Tak dokument lahko program za zajem obrne narobe.
- Avtomatsko obrezovanje dokumenta: program za zajem slike avtomatsko obreže dokument, če je dokument manjši od pričakovanega. V nasprotnem primeru bi tak dokument imel črno obrobo, kar slabša kakovost OCR-a. Poleg tega je zaradi obrezovanja elektronski dokument manjše velikosti.
- Nastavitev svetilnosti in kontrasta: nastavitev teh dveh parametrov je zelo pomembna, saj z njima urejamo intenzivnost barv in s tem črk. Najti je treba pravilno ravnovesje med intenziteto, saj premajhna oziroma prevelika svetilnost ali kontrast lahko posvetlita ali potemnita dokument do nerazpoznavnosti. Če je dokument presvetel, bodo znaki pri OCR-u izpuščeni, če je pretemen, pa se na njem pojavijo dodatne nečistoče, kar zniža kakovost OCR-a. Veliko komercialnih programov

nastavitev kontrasta in svetilnosti prilagaja avtomatsko, glede na skenirano dokumentacijo. Kasneje lahko sami popravljamo dokument do zelene oblike.

- Čiščenje dokumenta: s čiščenjem dokumenta dosežemo večjo berljivost, izboljšamo obliko znakov in odstranimo nepotrebne informacije. S čiščenjem dokumenta lahko bistveno izboljšamo kasnejšo prepoznavo znakov.
- Popravljanje debeline pisave: pogosto se zgodi, da se po čiščenju dokumenta ali spreminjanju drugih parametrov znaki na dokumentu stanjšajo in postanejo slabše razpoznavni. V tem primeru lahko uporabimo izboljšanje debeline pisave, ki popravi znake in s tem omogoči boljši OCR.

Zgoraj omenjene izboljšave so v naprednih aplikacijah za skeniranje velikokrat že avtomatsko uporabljene. Tako pri čiščenju dokumenta ni treba nastavljanja debeline pik, ki naj se brišejo, katera barva ozadja naj se briše itd. Aplikacija avtomatsko prepozna in obdela dokument, kar v večini primerov zadostuje potrebam skeniranega dokumenta. Najbolj znana aplikacija oziroma modul na tem področju je VRS (Virtual Re-Scan) [7] podjetja Kofax, ki je eno vodilnih podjetij na področju skeniranja, zajema, obdelave in elektronskega vodenja dokumentarnega gradiva. Modul omogoča vse zgoraj omenjene funkcionalnosti. Poleg tega uporaba samodejnega čiščenja privede do odličnih rezultatov, zaradi česar je prepoznavanje znakov zelo dobro.

Naslednja pomembna izboljšava, ki je doprinesla predvsem k pravilnosti prepoznavne besednih zvez, je uporaba slovarjev. Vsi prepoznani znaki, besede in besedne zveze se preverjajo v slovarjih. Če besedne zveze v slovarju ni, lahko aplikacija, ki izvaja OCR, spremeni parametre in poskuša doseči boljši OCR. Težava, ki se ob tem pojavlja, je nenehno dodajanje besednih zvez v slovarje. Poleg tega morajo dobre aplikacije OCR vsebovati večino slovarjev, ki se uporabljajo po vsem svetu. Težko si predstavljamo, koliko se v enem letu spremeni slovnica slovenskega jezika, če pa te spremembe pogledamo na

svetovnem nivoju, je praktično nemogoče slediti dnevnim trendom. Kljub vsem možnim dodatnim pomagalom in slovarjem kakovost oziroma točnost OCR-a še vedno ni 100 %.

Poglavje 6

Programska oprema za zajem in digitalizacijo

Poznamo več različne programske opreme za digitalizacijo in zajem metapodatkov dokumentarnega gradiva. Brezplačna programska oprema navadno ponuja omejeno funkcionalnost ali podpira le nekaj funkcij. Brezplačno programsko opremo po navadi dobimo že z nakupom naprave za zajem, kar je v primeru dokumentarnega gradiva skener. Pri vseh večjih proizvajalcih skenerjev dobimo programsko opremo, ki omogoča zajem dokumentarnega gradiva. Kot brezplačno programsko opremo pojmujeemo tudi opremo, ki smo jo dobili poleg naprave ob nakupu, vendar je kot samostojen produkt plačljiva.

Brezplačna programska oprema je navadno namenjena zajemu slike oziroma skeniranju. V večini primerov možnosti zajemanja metapodatkov, prepoznavanja znakov (OCR), avtomatske klasifikacije dokumentov itd. niso ali so slabo podprte.

V ta namen obstajajo kompleksnejše in dražje rešitve, ki skeniranje dokumentarnega gradiva in kasnejšo obdelavo poenostavijo. Te programske rešitve so seveda plačljive, pogosto zelo drage. Znotraj posamezne skupine programskih rešitev imamo glede na potrebe več možnosti. Naročnik, ki ne potrebuje natančnega zajemanja metapodatkov ali ima za to že drug sis-

tem, lahko uporabi cenejšo, manj kompleksno in manj obsežno programsko opremo. Med tako programsko opremo štejemo:

- Kofax Express [8] je produkt podjetja Kofax, namenjen predvsem profesionalnemu skeniranju in zajemanju majhnega števila metapodatkov. Produkt velja za enega najboljših na področju kakovosti skeniranja dokumentarnega gradiva. Kakovost zagotavlja z uporabo integrirane rešitve, imenovane VRS (Virtual Re-Scan) [7]. Rešitev VRS lahko uporabimo tudi samostojno, skupaj s skenerjem in drugimi produkti za zajem dokumentarnega gradiva, kot dodatek k izboljšani kakovosti elektronske verzije dokumentarnega gradiva. Poleg tega Kofax Express omogoča avtomatsko zajemanje črtnih kod, preverjanje integritete metapodatkov s podatkovnimi bazami, avtomatski vnos nekaterih metapodatkov itd.
- Kodak Capture Pro [9] je produkt podjetja Kodak in je prav tako namenjen profesionalnemu skeniranju in zajemu manjšega števila metapodatkov.
- Abbyy Scan Station [10] je produkt podjetja Abbyy in je namenjen zgolj profesionalnemu skeniranju.
- Kofax Desktop [11] je produkt podjetja Kofax, namenjen skeniranju manjše količine podatkov in ne omogoča zajema metapodatkov.

Kadar potrebujemo produkt, ki omogoča dobro in hitro optično prepoznavanje znakov oziroma OCR, imamo več možnosti. Največkrat uporabljene rešitve so:

- LuraDocument PDF Compressor [12] je produkt podjetja LuraTech, ki omogoča optično prepoznavanje znakov iz več različnih oblik vhodnih datotek (JPG, TIFF ...) in izvoz prepoznane dokumentacije v več izhodnih datotek (PDF, RTF, TXT ...).
- Abbyy Fine Reader [13] je produkt podjetja Abbyy in je namenjen OCR-u manjših količin. Posebnost tega produkta je funkcija, s katero

lahko popravljamo pravilnost prepoznanega besedila, področje prepoznanega besedila itd. Z njim lahko dosežemo najboljšo kakovost OCR-a dokumentarnega gradiva.

- Abbyy Recognition Server [14] je prav tako produkt podjetja Abbyy, vendar je v nasprotju s prejšnjim produktom namenjen veliki količini OCR-a. Njegova prednost je možnost uporabe spletne pošte za pošiljanje in prejemanje dokumentacije, nastavitev avtomatskega uvoza in izvoza dokumentov v in iz določene prednastavljene datoteke in uporaba programa kot servis, kar nam omogoča integracijo v poljubno programsko rešitev.

Vsi zgoraj naštetni produkti so nezahtevni za uporabo in administracijo, vendar imajo tudi manjše število funkcij in so namenjeni le specifičnim operacijam. Kadar naročnik želi kompleksnejšo rešitev za skeniranje in zajem velikega števila metapodatkov, sistemi postanejo zahtevnejši za postavitve in administracijo. Vsi naslednji omenjeni sistemi delujejo na principu strežnik–odjemalec (angl. *server–client*). Postavitev projekta po navadi traja daljše časovno obdobje in zahteva dobro poznavanje rešitve, nemalokrat pa vpliva tudi na obliko poslovnega procesa v podjetju. Vsi sistemi delujejo na principu delovnega toka (angl. *workflow*) in so grajeni modularno. Vsak modul predstavlja posamezno delovno operacijo, kot recimo skeniranje, optično prepoznavanje, vnos in pregledovanje prepoznanih metapodatkov ... Programska oprema, namenjena zahtevnejšim uporabnikom:

- Kofax Capture [15] podjetja Kofax omogoča vse omenjene funkcionalnosti prej omenjenega programa Kofax Express. Dodatno omogoča avtomatsko klasifikacijo dokumentov, avtomatski zajem metapodatkov, povezovanje z različnimi podatkovnimi zbirkami ... Prav tako omogoča prilagajanje vmesnikov glede na naročnikove potrebe in uporabo različnih programskih jezikov za prilagajanje vmesnikov in preverjanje vnesenih vrednosti metapodatkov.
- Abbyy Flexi Capture [16] podjetja Abbyy je podoben programu Kofax

Capture [15]; od njega se razlikuje po uporabniški izkušnji in nekaterih drugih funkcionalnostih.

- eFlow podjetja TIS (TopImageSystems) [17] prav tako omogoča podobne funkcionalnosti kot program Kofax Capture [15]. Poleg tega omogoča več proste uporabe programskih jezikov pri gradnji in povezavi modulov ter preverjanju vnesenih metapodatkov.

V nadaljevanju diplomskega dela bomo prav zaradi manjše omejitve pri uporabi programskih jezikov in s tem povezane večje možnosti proste izdelave svojih modulov in uporabe svojih programskih rešitev uporabljali produkt eFlow podjetja TIS. TIS eFlow je široko uporabljena programska oprema na vseh področjih digitalizacije. Na območju Slovenije je bila programska oprema uporabljena v veliko različnih projektih. Ravno tako je priznana na svetovnem nivoju kar pričajo reference podjetja TIS [31]. Programska oprema TIS eFlow je uporabljena zgolj za prikaz primera digitalizacije prejete pošte in prejetih računov. Namesto programske opreme TIS eFlow bi lahko uporabili katerokoli ostalo programsko opremo za digitalizacijo in zajem metapodatkov.

Poglavje 7

Zakonodaja na področju digitalizacije

Vsa nadaljnje omenjena zakonodaja je veljavna na območju Republike Slovenije in jo je potrebno dosledno upoštevati.

7.1 Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA)

To je ključni zakon elektronskega dokumentnega poslovanja. Izšel je v Uradnem listu RS, št. 30/2006 [1], in je naslednik Zakona o arhivskem gradivu in arhivih (ZAGA), ki je izšel v Uradnem listu RS, št. 20/1997 [18].

ZVDAGA [1] določa zahteve za celotno področje dokumentarnega in arhivskega gradiva, od ustvarjenja do končne hrambe gradiva. V 103 členih določa način, organizacijo, infrastrukturo in izvedbo zajema ter hrambe dokumentarnega gradiva v fizični in elektronski obliki, veljavnost oziroma dokazno vrednost takega gradiva, varstvo arhivskega gradiva in pogoje za njegovo uporabo, naloge arhivov in javne arhivske službe ter s tem povezane storitve in nadzor nad izvajanjem. Podaja tudi določila za varno dolgoročno hrambo dokumentarnega gradiva.

Temeljni člen ZVDAGA, ki omogoča elektronsko dokumentno poslovanje,

je 31. člen, ki se glasi: »Na podlagi zakona se vsaka enota varno hranjenega gradiva v digitalni obliki šteje za enako posamezni enoti izvirnega gradiva, če sta bila zajem in varna hramba opravljena v skladu s pri državnem arhivu potrjenimi notranjimi pravili ter če drug zakon izrecno ne določa drugače.« S tem členom ZVDAGA poenoti verodostojnost fizičnega in elektronskega dokumentarnega gradiva, elektronsko dokumentarno gradivo postane pravno formalno in je enakovredno fizičnemu. Za pravno formalnost elektronske kopije dokumentarnega gradiva pa ZVDAGA podaja veliko zahtev, ki morajo biti izpolnjene. Poleg tega ZVDAGA tudi navaja, da mora imeti podjetje sprejeta notranja pravila, ki jih mora tudi dokazano izvajati, ni pa nujno, da so ta notranja pravila potrjena pri Arhivu Republike Slovenije. Če organizacija nima potrjenih notranjih pravil, se gradivo v digitalni obliki šteje enako izvirnemu, če so izpolnjeni pogoji varne hrambe digitalnega gradiva.

V zakonu srečamo naslednja temeljna načela:

- Načelo ohranjanja dokumentarnega gradiva oziroma uporabnosti njegove vsebine Hramba dokumentarnega gradiva pomeni ohranjanje izvirnega dokumentarnega gradiva ali uporabnosti vsebine tega gradiva.
- Načelo trajnosti Hramba mora zagotoviti trajnost gradiva oziroma trajnost njegove reprodukcije.
- Načelo celovitosti Hramba mora zagotavljati nespremenljivost oziroma izvirnost in integralnost, urejenost dokumentarnega gradiva in dokazljivost njegovega izvora.
- Načelo dostopnosti Gradivo mora biti ves čas hrambe zavarovano in zaščiteno pred nepooblaščenim dostopom. Prav tako mora biti gradivo varno pred izgubo.
- Načelo varstva kulturnega spomenika Arhivsko gradivo se šteje za kulturni spomenik in se mora to upoštevajoč ustrezno obravnavati.

Zakon gradivo razdeli na več sklopov:

1. Dokumentarno gradivo

Dokumentarno gradivo je vsakršno gradivo v izvorni obliki. Gradivo je lahko v fizični ali digitalni izvorni obliki. V obeh primerih mora biti pretvorba¹ zanesljiva. Za zanesljivo pretvorbo se šteje pretvorba:

- ki zagotavlja reprodukcijo vseh bistvenih sestavin vsebine izvirnega gradiva (celovitost),
- ki ohranja uporabnost vsebine izvirnega dokumenta,
- ki zagotavlja avtentičnost gradiva z zajemom ključnih vsebin izvirnega gradiva ali s strogo nadzorovanim dodajanjem vsebin, ki še vedno potrjujejo avtentičnost,
- ki vsebuje primerno število kontrol pravilnosti in kakovosti pretvorbe,
- pri kateri se posebej in jasno ločeno od izvirne vsebine shranijo dodane vsebine ter vse pomembne opombe in podatki glede postopka pretvorbe in glede izvirnega gradiva,
- pri kateri se hrani primeren obseg dokumentacije, s katero se dokazuje, da uporabljene metode in postopki redno zagotavljajo zanesljivo pretvorbo in
- izpolnjuje pogoje, ki jih dodatno predpiše Vlada Republike Slovenije.

Če je pretvorba v digitalno obliko opravljena tako, da dokumentarno gradivo v digitalni obliki zagotavlja enake pogoje uporabnosti kot izvorno dokumentarno gradivo, in je skladno z zakonom urejena hramba digitalnega dokumentarnega gradiva, se lahko izvorno dokumentarno gradivo uniči. Izvornega dokumentarnega gradiva se ne sme uničiti, če ima oznako trajno (T), arhivsko (A) ali je z zakonom kakor koli drugače določeno.

¹Pretvorba: zajem gradiva iz ene oblike v drugo obliko, npr. iz fizične v digitalno obliko.

Oblike zapisa in nosilce zapisa se izbere glede na dolžino hrambe. Če je dolžina dolgoročna, kar pomeni nad pet let, izberemo takšno obliko zapisa in nosilce zapisa, ki zagotavljajo ohranitev vsebine zapisa. Poleg tega izberemo obliko zapisa, pri kateri bo po petih letih mogoče narediti prepis v novejšo obliko. Trenutno je največkrat uporabljena oblika zapisa za dolgoročno hrambo PDF/A.

2. Javno arhivsko gradivo

Javno arhivsko gradivo nastaja pri javnopravnih osebah in je last Republike Slovenije. Javno arhivsko gradivo je pravzaprav odbrano pomembnejše dokumentarno gradivo. Javno arhivsko gradivo hranijo pristojni arhivi ali v dogovoru z Arhivom Republike Slovenije posamezna organizacija sama. Javno arhivsko gradivo mora biti v najmanj 30 letih po nastanku gradiva izročeno v hrambo pristojnemu arhivu.

3. Filmsko arhivsko gradivo

Filmsko arhivsko gradivo je zmontirani originalni slikovni in tonski negativ filma, posnet na filmskem traku, in ena projekcijska kopija istega filma ter filmi, posneti na digitalnih ali analognih nosilcih, ki jih izdelajo slovenski ali tuji producenti oziroma so izdelani v koprodukciji slovenskih in tujih producentov v Republiki Sloveniji ter veljajo za slovenski film. Slovenski producent ali koproducent mora po nastanku filma predložiti kopijo filma pristojnemu arhivu.

4. Zasebno arhivsko

Zasebno arhivsko gradivo je last fizičnih in pravnih oseb zasebnega prava. Zasebno arhivsko gradivo določi pristojni arhiv glede na pregled gradiva. Hranjenje zasebnega arhivskega gradiva se mora izvajati v skladu z navodili arhiva. Zasebno arhivsko gradivo ima vse lastnosti javnega arhivskega gradiva in mora biti hranjeno v skladu z njimi. Pri prodaji zasebnega arhivskega gradiva ima Republika Slovenija predkupno pravico. Veliko zasebnega arhivskega gradiva imajo verske skupno-

sti.

Opravljanje storitev hrambe in spremljevalnih storitev za javnopravne osebe lahko opravlja samo akreditirani ponudnik. Tako morata biti vsa strojna in programska oprema akreditirani s strani Arhiva Republike Slovenije in se voditi v javni evidenci. Poleg tega mora imeti ponudnik potrjena notranja pravila in jih mora dokazljivo izvajati.

V našem diplomskem delu se bomo posvetili predvsem dokumentarnemu gradivu.

7.2 Uredba o varstvu dokumentarnega gradiva (UVDAG)

Uredba UVDAG [19], ki je bila objavljena v Uradnem listu RS, št. 86/2006, prav tako podaja temeljna določila na področju digitalizacije in elektronske hrambe dokumentarnega gradiva. Na področju digitalizacije uredba podaja ključna določila, na katera moramo biti pozorni pred, med in po digitalizaciji dokumentarnega gradiva. 12. člen uredbe podaja postopek zajema izvirnega dokumentarnega gradiva v fizični obliki in pretvorbe v digitalno obliko:

- evidentiranje in upravljanje vseh enot dokumentarnega gradiva,
- primerna klasifikacija zajetega dokumentarnega gradiva,
- pravilen zajem in pretvorba reprodukcije vsebine izbrane enote dokumentarnega gradiva iz fizične v elektronsko obliko,
- samodejna in ročna kontrola pravilnosti zajema in pretvorbe reprodukcije vsebine in metapodatkov,
- varnost in nespremenljivost zajetega in pretvorjenega dokumentarnega gradiva po pravilnem zajemu,

- možnost zajema ter poznejšega popravljanja napak pri zajemu in pre-tvorbi in upravičenega dopolnjevanja metapodatkov samo s strani po-oblaščenih oseb in z zagotavljanjem jasne revizijske sledi takih poprav-kov ali dopolnitev. Kot dopolnitev uredba navaja enotne tehnološke zahteve, ki bodo predstavljene v nadaljevanju dela.

Uredba navaja tudi pogoje za izbiro oblike zapisa za dolgoročno hrambo. Izbrana oblika mora omogočati ohranjanje vsebine gradiva, biti mora splošno priznana, neodvisna od programske ali strojne opreme, omogočati mora pre-tvorbo v kasnejšo splošno priznano obliko, temeljiti mora na odprtem stan-dardu, če ta obstaja, itd. Enotne tehnološke zahteve, različica 2.0 priporočajo več različnih zapisov glede na tip gradiva. Za besedilni tip gradiva se na-vadno uporablja oblika PDF/A, ki jo določa standard ISO 19005 [20], za grafične dokumente pa podajajo več različnih oblik zapisa, kot so TIFF, JPEG, JPEG200 itd. Uredba določa tudi obliko stiskanja (kompresije) do-kumentov, na primer LZW, CCITT group 4 itd. Določeni so tudi pogoji za izbiro nosilca za dolgoročno hrambo. Nosilec zapisa mora biti široko priznan, omogočati mora večje število prepisov na nove nosilce, ohranitev podatkov tudi ob prekinitvi dobave električne energije ali ob vplivu drugih okoljskih pogojev, temeljiti mora na odprtem standardu, če ta obstaja, itd. Poleg tega uredba podaja določila za avtentičnost in celovitost elektronskega do-kumentarnega gradiva. Tema dvema pogojema lahko zadostimo z uporabo elektronskega podpisa ali časovnega žigosanja.

7.3 Zakon o elektronskem poslovanju in elektronskem podpisu (ZEPEP)

Zakon o elektronskem poslovanju in elektronskem podpisu [21] izenači elek-tronski podpis dokumentarnega gradiva z lastnoročnim podpisom. Tako je elektronsko podpisano elektronsko dokumentarno gradivo po 5. členu tega zakona pravnoformalni dokument. Poleg tega se podatkov v elektronski

obliki ne sme odreči veljavnosti ali dokazne vrednosti samo zato, ker so v elektronski obliki, kar zagotavlja 4. člen zakona.

7.4 Enotne tehnološke zahteve (ETZ)

Enotne tehnološke zahteve, različica 2.0 (v nadaljevanju ETZ 2.0 ali ETZ) [22] so bile objavljene na spletnih straneh Arhiva Republike Slovenije 6. aprila 2011. Različica 2.0 nadgrajuje in zamenjuje prejšnjo različico 1.0, ki je bila objavljena 1. decembra 2006. ETZ podrobneje opredelijo poslovne, organizacijske in tehnološke pogoje za izpolnjevanje ZVDAGA. Vsebinsko so razdeljene na tri sklope:

1. uvodna poglavja,
2. enotne tehnološke zahteve za upravljanje in hrambo gradiva v digitalni obliki,
3. dodatne tehnološke zahteve za ponudnike, strojno in programsko opremo ter storitve.

Enotne tehnološke zahteve so močno povezane s standardom oziroma priporočilom Model Requirements for the Management of Electronic Records (MOREQ) [23], ki ga je odobrila Evropska komisija. MOREQ podaja dobre prakse na področju upravljanja elektronskih zapisov.

Poleg omenjenih zakonov, uvedb in predpisov obstaja še veliko drugih področnih predpisov, ki urejajo vsako posamezno področje. To so:

- ISO 15489 Information and documentation – Records Management,
- Zakon o tajnih podatkih,
- Zakon o varstvu osebnih podatkov,
- Zakon o splošnem upravnem postopku,
- Zakon o dostopu do informacij javnega značaja,

- Uredba o pogojih za elektronsko poslovanje in elektronsko podpisovanje.

Poglavje 8

Digitalizacija v praksi

Zaradi pozitivnih lastnosti digitalizacije se zanjo odloča vse več podjetij. Trend v digitalizaciji je pretvorba vsega fizičnega dokumentarnega gradiva v elektronsko obliko. Tako se veliko podjetij odloča za digitalizacijo vsega dokumentarnega gradiva. V nadaljevanju diplomske naloge se bomo posvetili predvsem dvema trenutno najbolj aktualnima področjema. To sta digitalizacija prejete oziroma vhodne pošte ter digitalizacija prejetih oziroma vhodnih računov. V obeh primerih govorimo o pretvorbi iz fizične v elektronsko dokumentarno obliko. Kadar organizacija prejme elektronsko obliko dokumenta proces digitalizacije ni potreben. Za pravilnost elektronskega arhiviranja takšne dokumentacije navadno poskrbi DMS sam. V kolikor je potrebno elektronsko prispelim dokumentom dodati metapodatke, se le ti navadno dodajo v DMS. V redkih primerih pa se lahko uporabi tudi sistem za digitalizacijo in zajem metapodatkov. V tem primeru se elektronski dokumenti uvozijo v sistem za digitalizacijo, celoten proces razen skeniranja pa je identičen digitalizaciji fizičnega dokumentarnega gradiva.

8.1 Vhodna pošta

Vhodna oziroma prejeta pošta predstavlja večji del dokumentarnega gradiva, ki prispe v podjetje, zato ima obvladovanje vhodne pošte velik pomen

za poslovanje podjetja. Vhodna pošta predvsem v večjih podjetjih po navadi prispe v oddelek, kjer pošto sprejmejo pooblašcene osebe, jo razvrstijo in signirajo ter razdelijo po podjetju. Vhodna pošta predstavlja veliko različnih tipov dokumentarnega gradiva, ki se razlikuje po vsebini, obliki, količini in prejemniku. Pravila razvrščanja vhodne pošte glede na tip najdemo v klasifikacijskem načrtu. Vsako podjetje ima svoj klasifikacijski načrt, po katerem se vhodna pošta razvrsti. Prav tako ima vsako podjetje svoj signirni načrt, na podlagi katerega se vhodna pošta dodeli prejemnikom. Poleg tega je treba vso vhodno pošto evidentirati. Za evidentiranje vhodne pošte se navadno uporablja t. i. knjiga vhodne pošte. Ta mora po standardih ISO 15489 [24] vsebovati vsaj:

- identifikacijski znak,
- naslov ali kratek opis,
- datum in čas vpisa,
- avtorja, pošiljatelja in prejemnika.

Javnopravne organizacije morajo glede na Uredbo o upravnem poslovanju [25] na prejeto pošto odtisniti stampiljko, na kateri je označeno:

- ime organa, ki je dokument prejel,
- datum prejema, v primeru javnega naročila še čas prejema,
- številko, ki je sestavljena iz klasifikacijskega znaka (številka zadeve, leto, številka dokumenta),
- signirni znak, ki je sestavljen iz organizacijske enote (delovnega mesta),
- število prilog,
- število vrednotnic.

Pooblaščen oseba mora torej za prejem pošte dobro poznati vsebino vhodne pošte. V vložišču se odpre vsa pošta razen:

- tajne pošte,
- pošte namenjene razpisu/natečaju,
- posebej označene pošte za osebno vročanje,
- napačno naslovljene ali
- nevarne pošte.

Če so vsi zgoraj omenjeni postopki ročni, navadno trajajo dolgo časa. Vhodna pošta tudi kasneje ostane v fizični obliki, kar ima veliko slabosti v primerjavi z digitalno obliko.

8.1.1 Digitalizacija vhodne pošte

Digitalizacijo vhodne pošte lahko izvajamo na več različnih načinov. Celoten postopek lahko opravlja naročnik sam s svojo strojno in programsko opremo. Navadno sta strojna in programska oprema za digitalizacijo vhodne pošte dragi in težavni za postavitve in vzdrževanje, zato si manjša podjetja tega ne morejo privoščiti. Poleg tega mora podjetje samo poskrbeti za vse postopke digitalizacije, kar prinaša dodatne stroške poslovanja. Druga možnost je prepustitev celotnega postopka digitalizacije vhodne pošte zunanjemu partnerju oziroma izvajalcu. Podjetje izvajalcu digitalizacije zaupa vso vhodno pošto na dnevnem nivoju, izvajalec pa jo naročniku vrne v digitalni obliki. Pri tem postopku naročnik nima stroškov z nakupom strojne in programske opreme. Naročnik pravzaprav plača najem storitve, ki je odvisna od količine vhodne pošte in zahtev naročnika za zajem. Med naročnikom in izvajalcem mora obstajati tesno zaupanje, izvajalec pa mora dobro poznati naročnikovo poslovanje, da lahko pravilno razvrsti in dodeli vhodno pošto. Tretji način je kombinacija prvih dveh. Naročnik uporablja le del storitev, ki jih potrebuje za izvajanje digitalizacije s pomočjo izvajalca. Naročnik po navadi najame programsko opremo ali zgolj njen modul za zajem metapodatkov, OCR in pripravo potrebnih datotek za kasnejši uvoz v naročnikov

sistem DMS. Naročnik sam zagotovi potrebno osebje in drugo programsko in strojno opremo za digitalizacijo.

8.1.2 Zahteve za metapodatke

Področje metapodatkov je zelo pomemben del digitalizacije vhodne pošte. S pomočjo metapodatkov bomo kasneje iskali digitalizirane dokumente. Če bomo uporabljali premalo metapodatkov, bomo imeli težave pri iskanju elektronskega dokumentarnega gradiva. Zahteve za metapodatke se usklajujejo skupaj z naročnikom glede na tip posamezne vhodne pošte.

Minimalne zahteve za besedilne dokumente glede na 2. del ETZ [22] določajo naslednje metapodatke:

- enolična identifikacijska oznaka (lahko je uporabljena črna koda),
- naslov ali kratka oznaka vsebine,
- datum (prejetja, nastanka,,
- avtor oziroma pošiljatelj,
- naslovnik.

Minimalne zahteve za metapodatke precej dobro pokrijejo področje vhodne pošte, dodatni metapodatki so potrebni le v nekaterih posameznih primerih. Tak primer so javna naročila. Javnih naročil se sicer ne odpira, ovojnica javnih naročil pa se kljub temu digitalizira. Pri javnih naročilih je poleg minimalnih zahtev za metapodatke treba dodati tudi datum in čas dospelja ponudbe za javno naročilo.

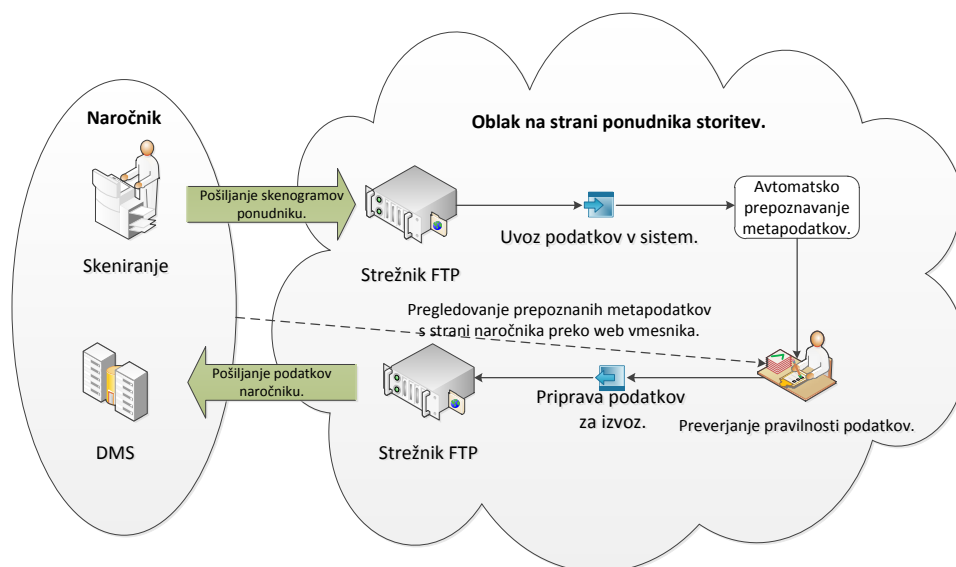
Poleg zahtev moramo določiti tudi obliko oziroma tip vseh metapodatkov in možnost njihovega preverjanja pravilnosti vsebine. Sistem TIS eFlow uporablja naslednje tipe podatkov:

- besedilo,
- celo število,

- število z decimalno vejico,
- datum in
- čas.

Pravilnost vsebine metapodatkov lahko preverjamo na veliko različnih načinov. Naštejmo nekaj najpogostejših oblik preverjanja:

- Dolžina besedila. Poštna številka v Sloveniji je lahko dolga najmanj in največ štiri mesta.
- Obvezen podatek.
- Preverjanje začetka, konca ali vmesnega dela besedila. Preverjamo lahko, na kakšen način se začne in konča besedilo oziroma ali vsebuje določene znake.
- Preverjanje s podatkovnimi bazami. Če je mogoče, celotno vsebino ali le del vsebine preverimo s podatkovnimi bazami. Z uporabo podatkovnih baz lahko pravilnost podatkov najbolj omejimo na točno določene vrednosti, ki so v podatkovnih bazah.
- Preverjanje z vnaprej določenimi vrednostmi, ki jih lahko naštejemo v TIS eFlow. Primer so lahko merske enote za težo katere lahko uporabniki samo izbirajo.
- Preverjanje z regularnimi izrazi (angl. *regular expressions*), ki jih lahko uporabljamo znotraj poljubnega programskega jezika.
- Zaporedna preverjanja, temelječa na več različnih, že vnesenih podatkih.
- Zapolnjevanje podatkov na podlagi že vnesenih, kot na primer zapolnitev imena pošte na podlagi poštne številke. Navadno so polja, ki se avtomatsko zapolnijo, namenjena le branju in jih ne moremo popravljati.



Slika 8.1: Postopek digitalizacije z najemom oblaka storitev.

- Vsa druga preverjanja, ki jih lahko napišemo v programskih jezikih.

8.1.3 Skeniranje pri naročniku in uporaba najete storitve

Naročnik sam poskenira pošto s svojo opremo. Že skenirano pošto v digitalni obliki pošlje ponudniku storitev, ki uporabi sistem za zajem metapodatkov. Naročnik lahko dostopa do tega sistema in pregleda ter vnese vse potrebne metapodatke. Po končanem vnosu se elektronski podatki avtomatsko prene-sejo v naročnikov sistem DMS. Celoten postopek prikazuje Slika 8.1.

Tak primer najema storitev bi lahko označili kot sistem v oblaku, kjer naročnik dostopa do zunanje najete programske opreme, sam pa zagotovi vse potrebno osebje in izpeljavo procesov, ki niso vključeni v oblak.

8.1.3.1 Skeniranje

Za skeniranje poskrbi naročnik sam, tako s strojno kot s programsko opremo ter delovno silo. Če želimo vhodno pošto pregledovati v digitalni obliki, jo moramo predhodno pretvoriti iz fizične v digitalno obliko. Za pretvorbo iz fizične v digitalno obliko navadno uporabljamo proces skeniranja. Zaradi različnih oblik vhodne pošte moramo poskrbeti za ustrezno podporo predvsem strojne opreme. Za vhodno pošto po navadi uporabljamo dva tipa skenerjev. Prvi je skener z avtomatskim polnjenjem dokumentov oziroma ADF (Automatic Document Feeder). Skener je namenjen pretvorbi velikega števila papirnih dokumentov v elektronsko obliko. Pred pretvorbo je treba dokumentacijo urediti. Pri urejanju moramo paziti, da liste med sabo ločimo in jih zložimo tako, da je skeniranje najlažje. Drugi tip skenerja je ploskovni skener. Ploskovni skener je namenjen dokumentaciji, katere listov ne smemo ločiti, skeniranju paketov in debelejših dokumentacij.

Pri skeniranju lahko popravljamo oziroma izboljšujemo slike do mere, da izboljšava ne vpliva na vsebino dokumenta. Lahko torej uporabimo nekatere funkcije, kot so avtomatsko obrezovanje dokumenta, obračanje dokumenta glede na pisavo in robove, odstranjevanje ozadja in grafik na njem itd. Sliko dokumenta brez ali z minimalno izgubo pomembnih informacij dokumenta lahko bistveno izboljšamo z uporabo produkta VRS (Virtual Re-Scan) podjetja Kofax.

Glede na tip dokumentacije ločimo tudi različne tipe izhodnih datotek skeniranja. Upoštevajoč enotne tehnološke zahteve se za izhodno obliko prejete pošte uporablja oblika PDF/A, ki jo določa standard ISO 19005. PDF/A je oblika PDF-a (Portable Document Format), zasnovana za digitalno arhiviranje dokumentov [26]. Je standard za dolgotrajno hrambo dokumentarnega gradiva in omogoča omejene funkcionalnosti, s čimer poskrbi za ohranitev vsebine dokumenta za daljše časovno obdobje. Na kratko, PDF/A omogoča preverjanje, ali je bil elektronski dokument v času od njegovega nastanka pa do danes spremenjen ali je enak izvirnemu fizičnemu dokumentu. Ta standard mora biti podprt tudi z notranjimi pravili, ki morajo biti doka-

zljivo izvajana. Trenutno se uporablja verzija PDF/A-2. Oblika PDF/A zagotavlja reproduciranje dokumentov na popolnoma enak način v prihodnosti. Za zagotovitev tega pogoja morajo dokumenti PDF/A vsebovati celotno vsebino. Vse informacije za prikaz vsebine morajo biti vsebovane v datoteki, zato PDF/A ne podpira povezave na zunanje datoteke. Druge lastnosti standarda PDF/A so:

- avdio in video vsebine niso dovoljene,
- JavaScript in izvedljive datoteke niso dovoljene,
- kriptiranje ni dovoljeno,
- povezava na vsebino zunanjih datotek ni dovoljena,
- stiskanje LZW ni dovoljeno, stiskanje JPEG2000 je dovoljeno le v verziji PDF/A-2,
- vse pisave morajo biti vgrajene,
- transparentni objekti in plasti v verziji PDF/A-1 niso dovoljeni, v verziji PDF/A-2 pa so,
- vgrajene datoteke so prepovedane v verziji 1, verzija 2 pa ponuja to možnost.

Glede na tip dokumentacije določimo tudi ločljivost skeniranja. Splošno pravilo, ki se je uveljavilo za skeniranje besedilne in mešane dokumentacije, je 300 DPI (Dots Per Inch – pik na palec) [27]. V svetovnem merilu točno določen standard, katera ločljivost naj se uporabi pri skeniranju, ne obstaja. V zadnjem času zaradi možnosti popravljanja skenirane slike proizvajalci programov OCR priporočajo skeniranje z ločljivostjo 200 DPI. Tako se kakovost OCR-a sicer nekoliko poslabša, vendar zmanjšamo velikost skenirane datoteke. Kadar prejmemo vhodno pošto, ki vsebuje pomembne detajle, kot je na primer grafična dokumentacija, starejša ali slabše ohranjena dokumentacija ipd., lahko uporabimo višjo ločljivost, na primer 400, 600 ali 1200 DPI,



Slika 8.2: Prikaz uporabe boljše in slabše ločljivosti skeniranja.

vendar se moramo zavedati, da to pomeni večjo velikost skenirane datoteke. Slika 8.2 prikazuje slabo ločljivost na desni strani in dobro ločljivost na levi strani.

Naročnik mora za ločevanje dokumentov poskrbeti sam. Pri velikih količinah skeniranja za ločevanje navadno uporabljamo črtne kode. Črtno kodo nalepimo na vsako prvo stran novega dokumenta. Namesto lepljenja črtnih kod se lahko ustvarijo šampiljke s QR (Quick Response) črtnimi kodami. Prednost kod QR je njihova dobra avtomatska prepoznavna v primeru slabe kakovosti kode. Črtno kodo lahko uporabimo le za ločevanje dokumentov (negovoreča črna koda), lahko pa vsebuje še dodatne podatke (govoreča črna koda). V primeru govorečih črtnih kod lahko dokumente opremimo z različnimi kodami, ki jih avtomatsko razvrstijo po sektorjih ali oddelkih znotraj podjetja.

8.1.3.2 Pošiljanje skenogramov preko FTP-ja

Po končanem skeniranju je treba podatke prenesti k ponudniku ostalih storitev. Za prenos skenogramov lahko uporabimo različne načine. Skenograme bi lahko prenašali daljše časovno obdobje s pomočjo prenosnega medija, kot je prenosni disk, vendar vhodno pošto potrebujemo takoj po prejemu, zato

mora biti čim hitreje pripravljena na elektronsko uporabo. Kadar želimo prenašati podatke v realnem času med dvema uporabnikoma, moramo uporabiti internetno povezavo. Ob dobri internetni povezavi lahko pošljemo veliko dokumentov v zelo kratkem času. Poznamo več različnih možnosti pošiljanja podatkov preko interneta. Podatke bi lahko pošiljali preko spletne pošte. Proces, ki bi vseboval pošiljanje podatkov preko spletne pošte, je preprosto vpeljati, vendar zahteva dodatno delo uporabnikov. Uporabnik mora vsakič, ko želi prenesti podatke k ponudniku storitev, odpreti dodaten program, pripeti podatke kot priponko in jih poslati. To dodatno delo po nepotrebnem obremenjuje zaposlenega.

Za pošiljanje podatkov preko interneta se priporoča uporabo FTP-ja (File Transfer Protokol), ki je široko uporabljen protokol. Pri vseh boljših komercialnih aplikacijah za skeniranje po navadi lahko integriramo pošiljanje podatkov preko FTP-ja v samo aplikacijo. Na ta način uporabnik aplikacije ne čuti nobene spremembe in ga ne obremenjujemo s pošiljanjem podatkov. Tako lahko proces digitalizacije dokumentarnega gradiva popolnoma avtomatiziramo.

Ker se podatki pošiljajo preko spleta, ki je množično uporabljen, je treba poskrbeti za varno pošiljanje. Vhodna pošta vsebuje veliko občutljivih in za poslovanje pomembnih podatkov; v primeru razkritja bi lahko to pripeljalo do izgube ugleda podjetja ali izgube velikega finančnega dela. Za pošiljanje podatkov preko spleta zato uporabljamo več različnih mehanizmov za zagotavljanje varnosti in integritete podatkov. Protokol FTP temelji na načinu strežnik–odjemalec ali pošiljatelj. FTP izvorno ni varen prenos podatkov, saj ne uporablja nobenega mehanizma, ki bi podatke kakor koli zavaroval. Za varovanje podatkov med prenosom uporabljamo več načinov. Najpogostejša sta enkripcija podatkov in podpisovanje podatkov. Enkripcija podatkov pomeni spreminjanje oziroma zakrivanje vrednosti podatkov, tako da jih napadalec ne more prebrati. Podpisovanje podatkov pa se uporablja za dokazovanje izvora podatkov in preverjanje integritete podatkov med prenosom. Kombinacija obeh metod omogoča visoko varnost prenosa podatkov. Prenos

podatkov mora biti tako močno zaščiten, da v primeru, ko napadalec med prenosom pridobi podatke, iz njih ne more pridobiti vsebine.

Za uporabo varnega prenosa FTP poznamo več različnih verzij protokola [28]:

- FTPS ali FTP–SSL: FTPS je pravzaprav FTP, ki ima dodatno kriptirano sejo in kriptiran prenos podatkov. Kriptiranje se izvede preko protokola SSL (Secure Sockets Layers)/TLS (Transport Layer Security). Pri FTPS-ju se lahko sami odločamo, ali bomo kriptirali samo nadzor ali tudi prenos podatkov. Če so podatki že predhodno kriptirani, jih ni treba kriptirati še enkrat in jih lahko prenašamo kot čisto besedilo. Tipično SSL/TLS uporabljata izmenjevanje ključev za ugotavljanje identitete, simetrično kriptiranje za kriptiranje podatkov ter digitalni izveček za ugotavljanje integritete sporočila.
- SFTP (Secure File Transfer Protocol): kot nadomestek uporablja kriptiranje seje oziroma nadzora in podatkovni prenos. Seja in podatkovni prenos sta združena v enem kanalu, medtem ko FTP uporablja dva ločena kanala. SFTP sloni na Secure Shell (SSH) protokolu. Za uporabo SFTP-ja ne moremo uporabljati standardnega strežnika FTP ali odjemalca.

Kot smo že omenili, podatkov ni treba dodatno kriptirati, če so bili predhodno že kriptirani. Tako lahko naročnik sam izbere način prenosa in enkripcije. Če se naročnik odloči sam kriptirati podatke, ima na voljo veliko brezplačnih programov, ki omogočajo kriptiranje. Eden izmed njih je GnuPG (Gnu Privacy Guard) [29], ki omogoča kriptiranje in podpisovanje podatkov. Program je skladen s standardom RFC (Request For Comments) 4880 in je alternativa standardu PGP (Pretty Good Privacy). GPG uporablja metodo javnega in zasebnega ključa ter asimetričnega kriptiranja. Prav tako lahko uporablja simetrično kriptiranje in digitalno podpisovanje.

Poleg skenogramov mora naročnik zagotoviti tudi podatke, namenjene avtomatskemu zajemu in preverjanju pravilnega vnosa metapodatkov. Po-

datkovne zbirke morajo biti poslane skladno z dogovorjeno obliko tipa datoteke. Standard za pošiljanje podatkovnih datotek glede na ETZ 2.0 je oblika datotek XML (Extensible Markup Language). Uporabljati moramo tudi dogovorjeno kodiranje znakov. Zaradi velikega števila različnih znakov in posebnosti slovenščine se priporoča uporaba kodiranja znakov Base64; to ni občutljivo na posebne znake v posamezni slovnici. Kodiranje Base64 zamenja vse znake v znake iz kodne tabele ASCII. Na ta način lahko pošiljamo različne znake, ki ne nastopajo v kodni tabeli ASCII. Večina podjetij, predvsem večjih, ima datkovne zbirke partnerjev, ki se dnevno dopolnjujejo. Zaradi različnosti vhodne pošte in manjšega števila metapodatkov je veliko podatkov nepreverljivih s podatkovnimi zbirkami. Preverjamo lahko naslednje elemente:

- pošta in poštna številka pošiljatelja,
- naziv pošiljatelja, če je to pravna oseba in
- naslovnik oziroma prejemnik.

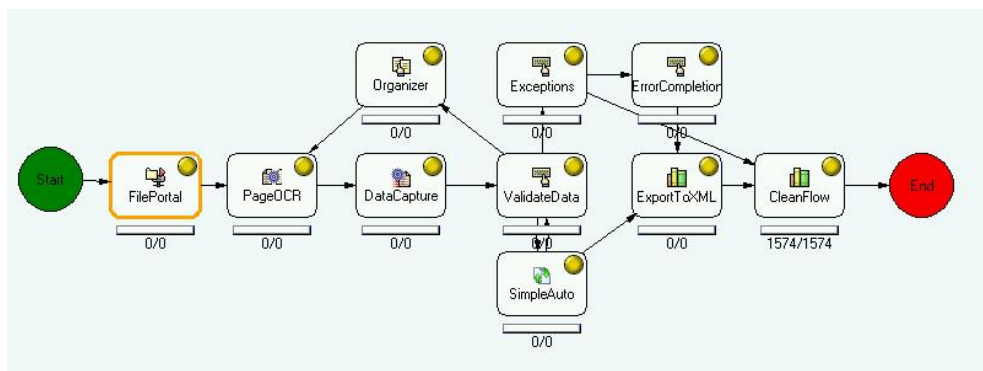
Več o tipu, obliki in zahtevah za metapodatke je opisano v poglavjih o avtomatskem zajemu metapodatkov in preverjanju metapodatkov.

8.1.3.3 Storitve v oblaku

Od tega koraka naprej vse storitve potekajo na strani ponudnika storitev. Za lažjo ponazoritev bomo uporabljali spodnjo sliko 8.3 aplikacije eFlow podjetja TIS.

Priporočeno je združeno pošiljanje vseh podatkovnih zbirk in skenogramov. To pomeni, da naročnik v oblak pošlje eno datoteko, ki vsebuje vse potrebne informacije. To navadno dosežemo s stiskanjem datotek in map. Prednost tega je samo ena datoteka, katere velikost še dodatno zmanjšamo.

Kot smo že omenili, aplikacije za zajem podatkov in digitalizacijo delujejo na način delovnega toka (angl. *workflow*). Ko en modul aplikacije konča z izvajanjem operacij, preda podatke naslednjemu, ki izvaja drugačne



Slika 8.3: Primer modularne zgradbe aplikacije TIS eFlow.

operacije. Tok dokumentov med moduli je na Sliki 8.3 predstavljen z usmerjenimi puščicami. Navadno delovni tok poteka z leve proti desni strani, le izjemoma se podatki vrnejo v predhodni modul. V celotnem toku aplikacije se skenogrami nahajajo v obliki zbirk (angl. *collection*). Zbirka je lahko enostranska, v primeru enostranskih dokumentov, ali večstranska, v primeru večstranskih dokumentov. Zbirka lahko vsebuje enega ali več dokumentov. Torej ima ena zbirka eno ali več strani dokumentarnega gradiva ne glede na število dokumentov.

Opis modulov in osnovnih pojmov na zgornji sliki 8.3:

- File Portal je modul, ki skrbi za uvoz skenogramov v aplikacijo.
- PageOCR je modul, ki opravi prepoznavo znakov na celotnem dokumentu. Brez tega modula v nadaljnjih korakih ne moremo avtomatsko zajemati metapodatkov.
- DataCapture je modul, namenjen avtomatskemu zajemu metapodatkov glede na predhodne parametre. Modul lahko zajema podatke na vnaprej določenih mestih ali poljubno glede na ključne besede. V primeru vhodne pošte bi lahko bila ključna beseda »Pošiljatelj«, saj se navadno poleg te besede nahaja naziv in naslov pošiljatelja.
- ValidateData je modul, kjer uporabniki preverijo in dopolnijo avtomat-

sko zajete metapodatke.

- Collection Organizer je modul, ki je namenjen operacijam nad zbirkami. V tem modulu lahko zbirke združujemo, razdružujemo, jim dodajamo ali brišemo strani, jih obračamo itd.
- SimpleAuto je modul, ki nam omogoča prosto programiranje njegovih funkcij in v osnovi ne vsebuje nobene funkcionalnosti.
- Exceptions je modul za popravljanje napak in anomalij, ki jih vnašalci med pregledom in vnosom metapodatkov ne morejo odpraviti zaradi omejitve sistema ali pomanjkanja znanja. Ta modul je namenjen vodji vnašalcev, ki pozna celoten postopek in celotno dokumentacijo. V tem modulu lahko zbirka čaka dalj časa, dokler se ne odpravijo vse anomalije.
- ErrorCompletion je modul, namenjen ponovnemu pregledu zbirk, popravljenih v modulu Exceptions. V tem modulu lahko vnašalci ponovno pregledajo popravljene zbirke in si ogledajo rezultate popravkov; te bodo lahko upoštevali v naslednjih primerih enakih ali podobnih zbirk.
- Modul Export skrbi za izvoz zbirk in metapodatkov iz sistema v dogovorjeni podatkovni obliki. Upoštevajoč ETZ 2.0 je oblika izvoženih skenogramov tipa PDF ali PDF/A, oblika izvoženih metapodatkov pa XML. Glede na dogovor z naročnikom se lahko tipi datotek spremenijo.
- Modul CleanFlow se uporablja za brisanje zbirk in vseh podatkov o posamezni zbirki iz aplikacije. Dokler je zbirka v aplikaciji, se jo lahko ročno prenese v posamezen modul in se jo naprej obdeluje.

Uporabnik sistema eFlow uporablja samo module ValidateData, Exceptions, ErrorCompletion in CollectionOrganizer. Vsi drugi moduli tečejo v ozadju kot procesi in jih uporabnik ne uporablja. Za dostop do modulov morajo uporabniki imeti povezavo do sistema eFlow. Za vzpostavitev povezave med naročnikom in ponudnikom se navadno uporablja VPN (Virtual

Private Network). Za uporabo morajo imeti uporabniki nameščenega klienta sistema eFlow, ki komunicira s strežnikom sistema eFlow. Omenjena rešitev je težja za implementacijo, zato se v današnjem času uporablja spletne različice modulov. V takem primeru potrebujejo uporabniki le spletni dostop do strežnika eFlow, vse drugo pa se odvija v spletnem pregledovalniku. Na ta način dosežemo popoln oblak in preprostost uporabe.

8.1.3.4 Podrobnejši opis modulov sistema TIS eFlow

V nadaljevanju so opisani pomembnejši moduli in njihove funkcionalnosti.

8.1.3.4.1 Modul FilePortal

Z modulom FilePortal uvozimo skenograme v sistem eFlow. V tem modulu se avtomatsko ustvarijo zbirke, če niso bile ustvarjene že med samim skeniranjem. Poleg uvoza skenogramov lahko v tem modulu dodatno izvedemo uvoz podatkovnih zbirk. Podatkovne zbirke navadno uvažamo v podatkovne baze, kot so na primer Oracle Database, IBM DB2, Microsoft SQL ipd. Uvoz podatkov v podatkovne baze je priporočljiv zaradi lažje in hitrejši uporabe podatkovnih baz v naslednjih postopkih. Primer podatkov, ki jih pošlje naročnik v obliki XML:

```
<?xml version="1.0" encoding="utf-8"?>
<data>
  <PostData>
    <Number id="1">1000</Number>
    <Name id="1">Ljubljana</Name>
  </PostData>
  <Customer>
    <Name id="1">Podjetje d.o.o.</Name>
    <Address id="1">Polžja ulica 10, 1000 Ljubljana</Address>
  </Customer>
</data>
```

8.1.3.4.2 Modul DataCapture

Z modulom DataCapture avtomatsko zajemamo podatke iz zbirk. Avtomatski zajem podatkov poteka na podlagi OCR-a. Vedeti moramo, katere podatke bomo zajeli in katere od njih je smiselno zajeti avtomatsko. V primeru vhodne pošte se dokumenti med sabo zelo razlikujejo, po navadi si niti dva dopisa nista enaka ali zelo podobna po obliki in vsebini. Zaradi tega je težko določiti statično mesto, kjer naj bodo metapodatki prepoznani, ali določiti ključne besede, ki označujejo metapodatke. V zelo redkih primerih pošte, ki vsebuje obrazce, lahko uporabimo avtomatsko prepoznavanje. Tak primer so poštne povratnice, ki imajo vse enako obliko. Pri dokumentih s statično obliko lahko določimo tudi statične lokacije prepoznavanja metapodatkov. Aplikacija TIS eFlow uporablja EFI (Empty Form Information) za določitev statičnih polj. EFI je enak vsem prejetim dokumentom, vendar ne vsebuje dodatne vsebine. Primer EFI-ja z določenimi polji za prepoznavo prikazuje Slika 8.4. Prikazana je vročilnica, ki jo izpolnimo ob prejemu osebnega pisma. Vsak rdeči kvadrat ponazarja statično lokacijo posameznega metapodatka.

8.1.3.4.3 Modul ValidateData

V modulu ValidateData preverimo pravilnost avtomatsko prepoznanih podatkov in jih v primeru nepravilnosti popravimo oziroma v primeru neprepoznanih podatkov dopolnimo. Kot smo že omenili, avtomatsko prepoznavanje podatkov ni 100

Večina validacij se implementira z uporabo programskih jezikov. Poskrbeti je treba za povezave na podatkovne baze, preverjanje podatkov s tabelami, uporabljati različne regularne izraze itd. Spodnji del kode prikazuje preverjanje pravilnosti vnesenega datuma. Kot lahko razberemo iz kode, se datum preverja na več različnih načinov. Tako lahko uporabnik vnese datum v poljubni obliki, koda pa ga preoblikuje v želeni format. V primeru napačne oblike datuma koda vrne obvestilo: »Napačna oblika datuma!«

The diagram shows a tax receipt form with the following fields highlighted by red boxes:

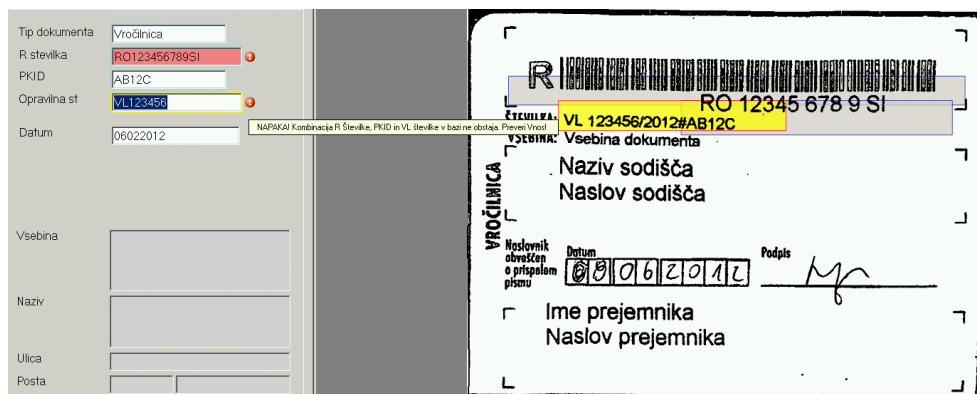
- A large box at the top for the invoice number (ŠTEVILKA).
- A box below it for the VAT number (VSEBINA).
- A box for the date (Datum) and signature (Podpis).

Slika 8.4: Primer EFI, prikazan na vročilnici.

```

if (Requirements.Normalization.NormalizeDate.Supported)
{
    DateTime fieldValue;
    if (!DateTime.TryParse(Globals.oFieldData.Contents, out fieldValue))
    {
        if (DateTime.TryParseExact(Globals.oFieldData.Contents, "ddMMyyyy",
            CultureInfo.InvariantCulture, DateTimeStyles.None, out fieldValue))
        {
            Globals.oFieldData.Contents =
                fieldValue.ToString(Requirements.Normalization.NormalizeDate.DateFormat);
        }
        else
        {
            if (DateTime.TryParseExact(Globals.oFieldData.Contents, "yyyyMMdd",
                CultureInfo.InvariantCulture, DateTimeStyles.None, out fieldValue))
            {
                Globals.oFieldData.Contents =
                    fieldValue.ToString(
                        Requirements.Normalization.NormalizeDate.DateFormat);
            }
            else
            {
                throw new Exception("Napačna oblika datuma!");
            }
        }
    }
}

```



Slika 8.5: Validacija vročilnice v sistemu TIS eFlow.

```

    }
  }
  else
  {
    Globals.oFieldData.Contents =
      fieldValue.ToString(Requirements.Normalization.NormalizeDate.DateFormat);
  }
}

```

Z vnaprej določenimi podatki, ki jih ročno vpišemo v projekt, ali z uporabo povezave na podatkovne baze se lahko preverjajo le statične vrednosti. Primer Slike 8.4 prikazuje pregledovanje in vnos podatkov v sistemu TIS eFlow.

Validacijo podatkov lahko opravlja naročnik sam ali zaposleni pri ponudniku storitev. V primeru, ko validacijo opravljajo zaposleni pri ponudniku storitev, morajo le-ti dobro poznati poslovanje in strukturo organizacije, za katero se validacija vhodne pošte opravlja.

Modul za validacijo mora biti oblikovan tako, da sta uporabniku omogočena nezapleten vnos in pregledovanje. To pomeni, da si morajo polja logično slediti v pravilnem vrstnem redu, slika, s katere se preverja podatke, mora biti dovolj velika, da prikazuje čim več podatkov hkrati z lahko berljivo velikostjo pisave. Obveščanje o napakah mora biti jasno in lahko opazljivo. Vse funkcionalnosti modula morajo biti dosegljive s kombinacijo tipk na tipkovnici,

da je delo s tipkovnico čim hitrejš.

8.1.3.4.4 Modul Exceptions

Modul Exceptions (napake) služi odpravljanju logičnih napak, nastalih med pregledovanjem v postaji ValidateData, ki so bile zaznane s strani pregledovalcev, vendar zaradi različnih neskladij niso bile odpravljene. V primeru neskladij, ki jih pregledovalec sam ne more odpraviti, lahko v modulu ValidateData označi polje kot neveljavno. Zbirka z neveljavnimi polji gre v modul Exceptions. Razlogov za neveljavnost podatkov je lahko več:

- vneseni podatek ne obstaja v bazi, vendar se podatek preverja z bazo,
- podatek v zbirki je drugačne oblike od pričakovanega,
- podatek v zbirki manjka, vendar je obvezen,
- vse druge logične napake, ki jih pregledovalci opazijo med pregledovanjem podatkov.

Napake navadno odpravlja ena sama oseba, ki najboljše pozna vsebino dokumentov in poslovanje celotnega podjetja. Ta oseba je po navadi vpletena tudi v postavitve sistema. Pogosto mora za odpravo ene logične napake sodelovati več različnih služb podjetja. Primer logične napake je pošta novega pošiljatelja, katerega še ni v bazi partnerjev. Pregledovalec v modulu ValidateData označi polje s podatkom o nazivu pošiljatelja kot neveljavno. Oseba, ki odpravlja napake v modulu Exceptions, obvesti zadolžene za dodajanje pošiljateljev o novem pošiljatelju. Ko je pošiljatelj dodan v bazo, se zbirka dokončno popravi in dopolni ter sprostí za naslednji modul.

8.1.3.4.5 Modul Export

Modul Export služi pripravi zbirk na izvoz iz sistema TIS eFlow ter izvozu pripravljenih zbirk. Izvoz podatkov in dokumentov se implementira z uporabo programskih jezikov. Izvoz metapodatkov mora biti skladen z

naročnikovim dogovorom. Glede na priporočilo se uporablja izvoz v obliki XML s kodiranjem znakov base64. Metapodatki se v kasnejši fazi po navadi avtomatsko uvozijo v naročnikov dokumentni sistem in se jih kasneje uporabi kot ključne besede za iskanje dokumentov. Spodnji primer prikazuje obliko metapodatkov po izvozu iz sistema TIS eFlow.

```
<?xml version="1.0" standalone="yes"?>
<import>
  <Document Scheme="VhodnaPosta" Pool="Posta/2012">
    <Attributes>
      <NamedDV Name="Id_Poste">
        <Value Type="TEXT" Encoding="base64">MjAxMjAwMDIONTI=</Value>
      </NamedDV>
      <NamedDV Name="NaslovAliKratkaOznaka">
        <Value Type="TEXT" Encoding="base64">
          UG9nb2RiYSBvIHNRbGVuaXR2aSBwYXJ0bmVyc3R2YS4=
        </Value>
      </NamedDV>
      <NamedDV Name="Daum">
        <Value Type="TEXT" Encoding="base64">MjAxMjA1MDQ=</Value>
      </NamedDV>
      <NamedDV Name="Posiljatelj">
        <Value Type="TEXT" Encoding="base64">
          UGFydG51ciBkLm8uby4sIExvZ292YSAxNCwgMTAwMCBManVibGphbmE=
        </Value>
      </NamedDV>
      <NamedDV Name="Naslovnik">
        <Value Type="TEXT" Encoding="base64">SmFuZXogTm92YWs=</Value>
      </NamedDV>
      <BlobContent>
        <Reference Type="RELATIVEFILE" Link="2012_05_04_037\11000000519_1.tif"/>
        <Attributes/>
      </BlobContent>
      <BlobContent>
        <Reference Type="RELATIVEFILE" Link="2012_05_04_037\11000000519_2.tif"/>
        <Attributes/>
      </BlobContent>
    </Attributes>
  </Document>
</import>
```

Poleg metapodatkov se izvozijo tudi pripadajoči elektronski dokumenti. Priporočilo za izvoz dokumentov besedilnega tipa je PDF oziroma PDF/A.

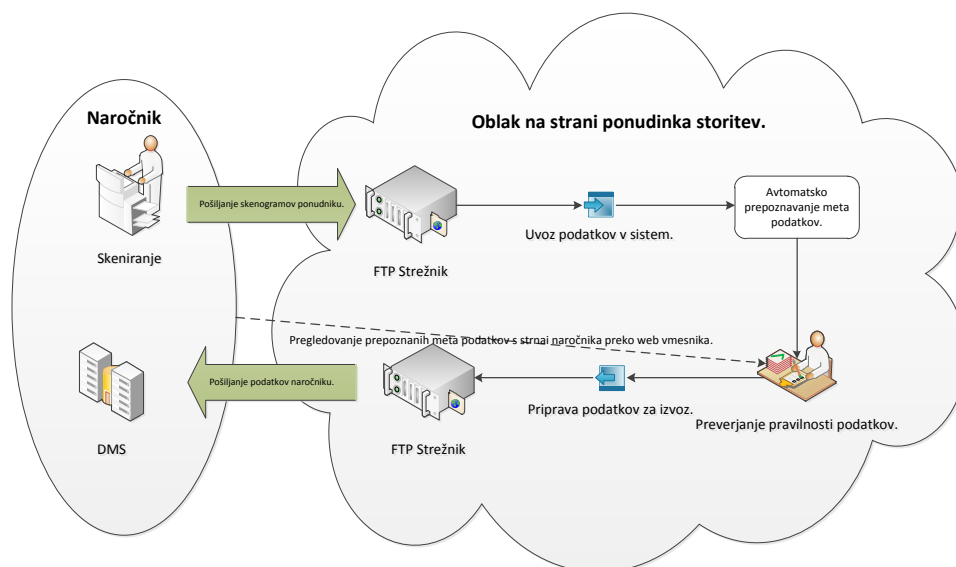
Glede na želje naročnikov je lahko dodatno uporabljena funkcija prepoznavanja znakov celotnega dokumenta. Izvoz podatkov zaradi hitrosti po navadi poteka lokalno na strežniku TIS eFlow. Po končanem izvozu se vsi podatki v celoti prenesejo k naročniku z uporabo varne povezave FTP.

8.1.4 Preusmeritev pošiljanja pošte k izvajalcu digitalizacije

V prvem primeru je naročnik aktivno vpleten v digitalizacijo vhodne pošte. Drugi primer, ko naročnik ni vpleten v digitalizacijo, zaradi različnih dejavnikov še ni zaživel v polnem obsegu. V drugem primeru naročnik preusmeri vso vhodno pošto na izvajalca digitalizacije. Izvajalec poskrbi za celoten postopek, od skeniranja, preverjanja podatkov do izvoza in pošiljanja podatkov naročniku. S pošto službo se lahko dogovorimo o preusmeritvi vse vhodne pošte na drug naslov in pooblastimo drugo osebo, v našem primeru izvajalca digitalizacije, za prejem vhodne pošte.

Omenjeni način pošiljanja vhodne pošte še ni dobro zaživel in vprašanje je, ali bo sploh kdaj zaživel v polni meri. Največja težava je zaupanje med naročnikom in izvajalcem, saj so bistvene informacije za poslovanje podjetja odvisne ne samo od naročnika, temveč tudi od izvajalca. Predstavljajte si, da vso vašo pošto zaupate tretji osebi, ki jo za vas ureja, in je od tega odvisna vaša poslovna kakovost ter prihodnost. Izvajalec mora odlično poznati organizacijo, saj mora vedeti, komu posamezna vhodna pošta pripada, katero vhodno pošto lahko odpira in katere ne, kam jo signirati in klasificirati itd. Vse to včasih težko ugotovijo že naročnikovi zaposleni, ki izvajajo te operacije in poznajo organizacijo, kaj šele izvajalec digitalizacije. V najboljšem primeru bi naročnik na koncu prejel le digitalno verzijo vhodne pošte, originali pa bi se hranili pri izvajalcu. Tega ni mogoče doseči, saj osebne pošte, javnih naročil itd. izvajalec ne sme odpreti. Tako vhodno pošto je treba poslati nazaj naročniku, kar pa zopet podaljša čas prejema vhodne pošte.

Prednosti omenjenega procesa so predvsem v nevpletenosti naročnika v proces digitalizacije. Naročnik ne potrebuje strojne in programske opreme



Slika 8.6: Model digitalizacije, kjer naročnik ni vpleten v postopek.

ter osebja, namenjenega skeniranju in pregledovanju vhodne pošte. Vse to opravi izvajalec, kar prikazuje tudi Slika 8.6.

8.2 Vhodni računi

Vhodni ali prejeti računi predstavljajo drugi največji del prejetega dokumentarnega gradiva posameznega podjetja. Obvladovanje prejetih računov je velikega pomena, saj so le-ti eden večjih izdatkov podjetja. Z dobrim obvladovanjem prejetih računov lahko posamezno podjetje veliko privarčuje in pridobi konkurenčno prednost. Vsako podjetje ima svoj način vodenja računov. Manjša podjetja to zaupajo zunanjim izvajalcem, medtem ko imajo večja podjetja za to usposobljeni kader. Vsi računi morajo biti pravilno zavedeni v knjigi prejetih računov in tudi knjiženi. Navadno večja podjetja

za upravljanje z računi uporabljajo različne programe, namenjene lažjemu obvladovanju računov.

Pri prejetih računih moramo biti zelo pazljivi pri vseh elementih digitalizacije. Predstavljajte si, da bi se zmotili le v eni številki pri vsoti računa. Tako bi bil račun lahko veliko višji od pravilne vsote. Napak ne dovoljujejo niti različni računovodski zakoni in pravilniki. Torej mora biti digitalizacija prejetih računov opravljena brez napak.

Glede na to, da je postopek digitalizacije vhodnih računov zelo podoben digitalizaciji vhodne pošte, bomo v nadaljevanju opisali samo razlike med postopkoma. Največja razlika med postopkoma je avtomatski zajem metapodatkov in preverjanje pravilnosti podatkov. Vsak račun je navadno razdeljen na tri logične dele:

- glava računa, kjer so navedeni pošiljatelj, datum, številka računa,
- postavke računa, kjer so navedene vse postavke računa, kot so ime, število artiklov, cena artiklov, skupna cena artiklov itd. Obstajajo tudi računi brez postavk,
- noga računa, kjer so navedeni: skupni znesek, davek in znesek davka, skupni popust računa itd.

Račune lahko razdelimo na dve obliki glede na naročilo. Račun ima lahko naročilo ali pa ga ne vsebuje. V primeru, ko račun vsebuje naročilo, lahko podatke na računu preverjamo z oddanim naročilom. Pri tem predpostavljamo, da ima podjetje urejeno zbirko naročil. Na ta način lahko preprosto preverjamo razliko med naročenim in zaračunanim blagom.

Poleg računov poznamo tudi dobavnice, ki so zelo podobne računom, vendar ne vsebujejo skupnih zneskov. Z digitalizacijo dobavnic lahko preprosto primerjamo naročeno in dobavljeno blago.

Digitalizacija vhodnih računov vedno poteka v sodelovanju naročnik–izvajalec, saj vhodni računi že pred samo digitalizacijo zahtevajo natančno obdelavo.

8.2.1 Zahteve za metapodatke

Vhodni računi v nasprotju z vhodno pošto vsebujejo veliko več podatkov, ki jih je treba med procesom digitalizacije zajeti. ETZ ne opredeljuje natančnejših minimalnih zahtev za račune. Računi so glede na ETZ obravnavani kot besedilni dokumenti in imajo enake minimalne zahteve kakor vhodna pošta, vendar v poslovnem svetu minimalne zahteve za metapodatke v primeru računov niso zadovoljive. Skoraj vsak podatek na računu je v naslednjih korakih obvladovanja računov lahko pomemben. Metapodatke, pridobljene z digitalizacijo, lahko preprosto preverjamo z obstoječimi sistemi za vodenje računov in preprosto ugotavljamo neskladja.

Glede na razdelitev računa lahko določimo naslednje metapodatke:

- Metapodatki glave računa:
 - črtna koda, ki je navadno enolična identifikacijska oznaka,
 - prejemnik in/ali šifra prejemnika,
 - številka računa,
 - datum računa,
 - valuta računa oziroma rok plačila računa,
 - sklic,
 - pošiljatelj, ki ga razdelimo na več delov:
 - * ID pošiljatelja,
 - * naziv pošiljatelja,
 - * naslov pošiljatelja,
 - * davčna številka,
 - * transakcijski račun pošiljatelja,
 - * IBAN – mednarodna številka bančnega računa in
 - * BIC – bančna identifikacijska koda.
- Metapodatki postavk računa:

- številka naročila,
 - številka dobavnice,
 - količina,
 - enota mere,
 - cena na enoto,
 - popust,
 - skupna cena na postavko,
 - številka artikla in
 - naziv artikla.
- Metapodatki noge računa:
 - skupna neto cena,
 - skupen popust,
 - davek oziroma davčna stopnja,
 - znesek davka,
 - skupna bruto cena in
 - denarna valuta računa.

V primeru vhodnih računov preverjamo tudi pravilnost vsebine metapodatkov. Način preverjanja veljavnosti je opisan v poglavju o vhodni pošti. Če v primeru vhodne pošte nismo imeli veliko možnosti uporabe podatkovnih zbirk, imamo v primeru vhodnih računov veliko možnosti preverjanja podatkov s podatkovnimi zbirkami. Veliko podjetij ima podatkovne zbirke, ki vsebujejo skoraj vse podatke računa razen zneskov, številke računa, datumov računa ipd.

Sistem TIS eFlow za avtomatsko prepoznavo podatkov zahteva točno določene zbirke podatkov. Za prepoznavo uporablja sedem različnih datotek oblike CSV (Comma-Separated Values). Metapodatki teh datotek se uvozijo v podatkovno bazo. Nujno potrebne datoteke so:

- baza prejemnikov računov,
- baza dobaviteljev,
- baza bančnih podatkov dobaviteljev,
- baza povezav med bazo dobaviteljev in bazo bančnih podatkov dobaviteljev,
- baza vseh postavk računov glede na naročilo,
- baza povezav med naročilom in dobaviteljem. Naročila so v bazi postavk računov,
- baza povezav med dobavnico in postavkami računa.

8.2.2 Klasifikacija dokumentov v sistemu TIS eFlow

Klasifikacija je tesno povezana z avtomatskim zajemom metapodatkov. Boljša klasifikacija računov pomeni večji odstotek avtomatsko zajetih metapodatkov in manj ročnega pregledovanja.

V primeru vhodne pošte smo govorili o klasifikaciji dokumentov, kot jo imamo definirano v klasifikacijskem načrtu. Vhodni računi imajo glede na klasifikacijski načrt isti klasifikacijski znak, na primer vhodni račun. Klasifikacija v sistemu TIS eFlow je sicer konceptualno podobna, vendar je precej natančnejša. Pri prejetih računih srečujemo samo en tip dokumentacije, v podrobnostih pa se računi med sabo precej razlikujejo.

Klasifikacija dokumentov navadno poteka glede na ključne besede in/ali obliko dokumenta. Po navadi ima vsako podjetje specifično obliko računa, ki je enaka vsem računom in se ne spreminja pogosto. Za avtomatski zajem metapodatkov moramo definirati čim več takšnih oblik računov. V tem primeru lahko uporabimo Paretovo načelo oziroma pravilo 20/80, ki pravi, da 20 % vzrokov povzroči 80 % posledic. Če to pravilo preoblikujemo za našo tematiko, lahko rečemo, da 20 % najpogostejših dobaviteljev, ki nam

izstavljajo račune, pošlje 80 % vseh prejetih računov. Da bi bilo pravilo bolj natančno, ga moramo nekoliko prilagoditi v 30/70.

Klasifikacija glede na obliko poteka tako, da sistemu vnaprej podamo posamezne račune. Sistem za nove račune v postopku prepoznavanja sam izbira primerne oblike računov glede na podane. Za pripravo predlog se v sistemu TIS eFlow uporablja dodaten modul za učenje med pregledovanjem podatkov. V tem modulu glede na obstoječi dokument označimo lokacije metapodatkov. Ko bomo naslednjič prejeli dokument istega dobavitelja, bo lahko sistem avtomatsko prepoznal označene podatke.

Klasifikacijo glede na ključne besede izvajamo tako, da v sistem vnašamo besede, ki so enolični identifikator računa. Največkrat so ključne besede besedne zveze, ki opisujejo pošiljatelja, kot so ime pošiljatelja, davčna številka itd.

Klasifikacijo dokumentov izvajamo šele, ko imamo dobro urejene ključne besede, ki opisujejo posamezen metapodatek. Ko sistem metapodatka, ki je opisan s ključnimi besedami, ne prepozna samodejno, uporabimo klasifikacijo. Zavedati se moramo, da z uporabo klasifikacije določimo oziroma narišemo statično polje, kot na primer v primeru vročilnic na Sliki 8.4. Ko se pojavi nov račun z drugačno obliko, moramo ponovno uporabiti modul za učenje oblike.

8.2.3 Skeniranje in pošiljanje preko FTP-ja

Digitalizacija vhodnih računov se strukturno ne razlikuje od digitalizacije vhodne pošte. Postopek digitalizacije je od začetka do konca popolnoma enak; predstavlja ga Slika 1. Digitalizacija se razlikuje samo v nekaterih modulih, kar bomo predstavili v nadaljevanju.

8.2.4 Podrobnejši opis modulov sistema TIS eFlow

Zaradi podobnosti z digitalizacijo vhodne pošte bomo v nadaljevanju predstavili samo razlike v posameznih modulih glede na digitalizacijo vhodne

pošte.

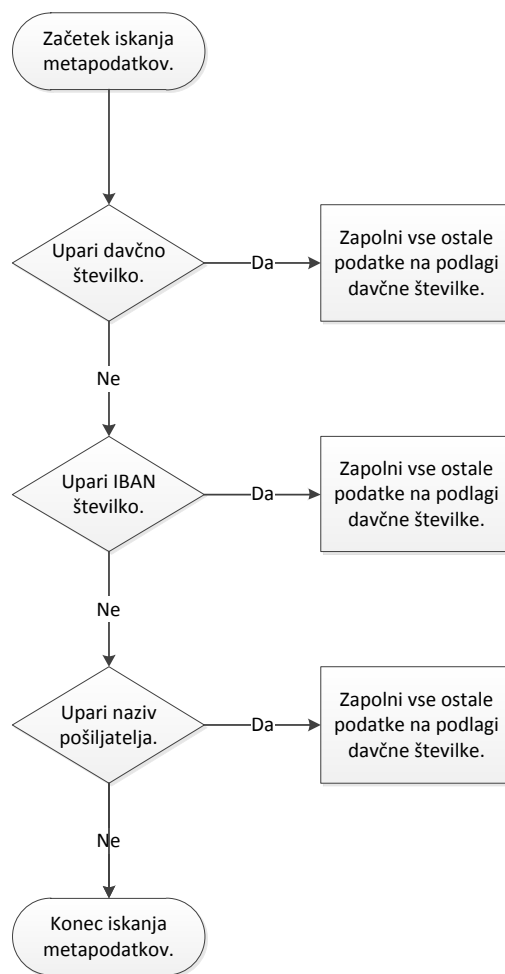
8.2.4.1 Modul DataCapture

Modul DataCapture se oblikovno ne razlikuje od modula, ki se uporablja za vhodno pošto, pomembna pa je razlika avtomatskega prepoznavanja metapodatkov. Metapodatki se avtomatsko prepoznavaajo glede na predhodno klasifikacijo, vendar to ni dovolj za prepoznavanje metapodatkov. Če bi za iskanje metapodatkov uporabljali samo klasifikacijo, bi dobili premalo samodejno prebranih metapodatkov. Poleg klasifikacije moramo definirati tudi ključne besede, ki opisujejo posamezen metapodatek. Te so za dober avtomatski zajem podatkov najpomembnejše. Za vsak metapodatek moramo definirati besedne zveze, ki ga opisujejo. Poleg tega moramo podati lokacijo metapodatka glede na najdeno ključno besedo. Tako sistem TIS eFlow avtomatsko zajema podatke, ki ne obstajajo v bazah. Primeri takih podatkov so številka računa, datum računa, datum valute itd.

Drugi način iskanja podatkov je s pomočjo podatkovnih baz. Kot smo že omenili, sistem zahteva sedem datotek, ki jih porabi kot predloge za samodejno prepoznavanje podatkov. V ozadju sistem za odločanje uporablja utežene vsote za izbiro pravih rezultatov. Kot primer vzemimo iskanje pošiljatelja računa. Recimo, da sistem prepozna davčno številko, IBAN in naziv pošiljatelja. Pomembnost metapodatkov pada v tem vrstnem redu. V bazi dobaviteljev in bančnih dobaviteljev imamo vse podatke pošiljateljev. Sistem poskuša najprej najti najpomembnejše metapodatke, potem pa na podlagi tega iz baze črpati še vse druge podatke. V našem primeru je najpomembnejši podatek davčna številka. Sistem poskuša na posameznem dokumentu najti davčno številko, ki obstaja v bazi. Če jo najde, avtomatsko zapolni druga dva podatka iz baze. Če davčne številke ne najde, nadaljuje z drugim najpomembnejšim podatkom, v našem primeru je to IBAN. Če najde številko IBAN v bazi, avtomatsko zapolni vse druge podatke. Ta postopek ponavlja, vse dokler ne najde ujemajočega zadetka v bazi ali mu zmanjka kandidatov za metapodatke. Algoritem na Sliki 8.7 prikazuje iskanje in

zapolnjevanje metapodatkov v povezavi s podatkovno bazo. Kot lahko opazimo, vse ključne besede iščemo v podatkovnih bazah.

Prikazani način iskanja metapodatkov je preprost, kadar sistem išče več podatkov, kot so na primer postavke računov, naenkrat, pa je iskanje precej bolj zapleteno. Sistem poleg obstoječih podatkovnih baz uporablja tudi logična ujemanja med metapodatki. Za iskanje pravih postavk upošteva tudi pravilne preračune med vsemi zneski na računu. Sistem v ozadju poskuša najti tako kombinacijo podatkov, ki sovpada z vsemi zneski na računu, kot so: skupna vsota, davek, posamezna vrednost postavke ... Natančnega opisa, kako sistem išče več med sabo povezanih metapodatkov, ne moremo podati, saj je to poslovna skrivnost podjetja TIS.



Slika 8.7: Algoritem iskanja in zapolnjevanja metapodatkov s pomočjo podatkovnih baz.

8.2.4.2 Modul Validation

Validacija prejetih računov je oblikovno enaka validaciji vhodne pošte. Bistvena razlika je število preverjanj pravilnosti podatkov. Metapodatki računov morajo biti popolnoma pravilni, za kar moramo poskrbeti z zadostnim številom preverjanj vnesenih podatkov. Veliko pravilnosti lahko preverjamo v povezavi s podatkovnimi bazami, pravilnost vnosa zneskov pa lahko preverjamo s preračuni različnih podatkov. Primer je lahko bruto vrednost računa, ki jo ponazarja enačba (8.1).

$$\textit{brutovrednost} = \textit{netovrednost} + \textit{netovrednost} * \textit{davek} \quad (8.1)$$

Seveda predpostavljamo, da je preverjanje neto vrednosti seštevek vseh predhodnih postavk računa.

Poglavje 9

Sklepne ugotovitve

Ob predhodnem poznavanju programske opreme za digitalizacijo in zajem metapodatkov ter poznavanju problematike digitalizacije v Sloveniji in svetu, smo izdelali dve rešitvi, kateri bosta v prihodnje verjetno najpopularnejši na področju digitalizacije. Z uporabo obstoječe programske opreme TIS eFlow, smo izdelali rešitvi digitalizacije prejete pošte in prejetih računov kot storitev v oblaku. Oba modela smo zasnovali na podlagi preteklih izkušenj in pričakovanj razvoja programske opreme za digitalizacijo v prihodnosti. Za izdelavo obeh postopkov smo morali upoštevati vso obstoječo zakonodajo na področju Slovenije, katero smo tudi podrobneje povzeli v diplomski nalogi. Za vsako rešitev opisujemo celoten postopek digitalizacije dokumentarnega gradiva. Podali smo natančne opise vseh zahtev in postopkov za skeniranje dokumentov, pošiljanje podatkov, prepoznavanje in izpolnjevanje metapodatkov ter uporabo priporočenih standardov na področju digitalizacije dokumentarnega gradiva. Dodatno smo vsako rešitev razčlenili na dva primera. V prvem primeru je v digitalizacijo dokumentarnega gradiva naročnik neposredno vpleten medtem, ko v drugem primeru naročnik prejme le končni rezultat digitalizacije.

Pri postopku obvladovanja poslovne digitalizacije prejete pošte in prejetih računov smo spoznali, da je postopek veliko bolj zapleten kakor obvladovanje poslovne dokumentacije v fizični obliki. Upoštevati moramo veliko

zakonskih predpisov in priporočil, katerih se moramo dosledno držati. Na trgu obstaja veliko različne strojne in programske opreme, ki nam ponujata različne možnosti glede na potrebe naročnika. To programsko opremo je treba implementirati in vzdrževati. Implementacija programske opreme za digitalizacijo v veliko primerih zahteva odlično strokovno usposobljenost in znanje. Zavedati se moramo, da pri digitalizaciji prejete pošte in prejetih računov delamo z vsebinsko zelo občutljivimi dokumenti. Vsako razkritje ali izguba pomembnejših informacij lahko privedeta do velike poslovne izgube in padca ugleda organizacije. Pomemben dejavnik pri vpeljavi digitalizacije je osebje, ki izvaja digitalizacijo. Osebje je treba ustrezno izobraziti in motivirati za uporabo novega postopka.

Opisani postopki digitalizacije prejete pošte in prejetih računov so skladni z zakonskimi predpisi in se lahko uporabljajo v vsaki organizaciji javnopravne ali zasebne oblike. Oba izdelana primera lahko služita kot dobri praksi na področju digitalizacije prejete pošte in prejetih računov v oblaku storitev. Izdelava opisa procesa digitalizacije zajema dobre prakse na področjih, ki so bila že predhodno izdelana. Digitalizacija vsega poslovnega gradiva podjetju prinese veliko konkurenčno prednost. Velike organizacije si bodo tudi v prihodnje same kupovale celoten sistem za izvajanje digitalizacije, medtem ko je za mala podjetja priporočljivejši sistem najema programske opreme. V diplomski nalogi je predstavljen koncept oblaka oziroma najema storitev, ki je primeren za vsa podjetja in sledi zadnjim trendom informacijske dobe. V prihodnosti lahko pričakujemo še boljšo zakonsko podprtost digitalizacije in seveda izboljšano programsko in strojno opremo za izvajanje digitalizacije. Zanimivo bi bilo še podrobneje raziskati in dodatno podpreti način digitalizacije, ki bi jo v celoti izvajal izvajalec sam. Trenutno je digitalizacija poslovnega gradiva v Sloveniji in svetu v velikem razmahu. Vsa podjetja si želijo digitalizirati svoje poslovno gradivo in ga uporabljati v elektronski obliki. Kljub svetli prihodnosti pa v daljšem časovnem obdobju pričakujemo počasen upad digitalizacije poslovnega gradiva. Z boljšim razvojem informacijskih sistemov za neposredno elektronsko poslovanje med podjetji in boljšo

zakonsko podporo neposrednega elektronskega poslovanja bo digitalizacija poslovnega gradiva upadla, kljub temu pa bo še vedno v veliki meri prisotna na vsakem našem koraku.

Literatura

- [1] (2012) Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA).

Dostopno na:

<http://www.uradni-list.si/1/content?id=72425>

- [2] (2012) Definicija arhivskega gradiva. Dostopno na:

http://sl.wikipedia.org/wiki/Arhivsko_gradivo

- [3] (2012) Projekt Europeana 1914 – 1918. Dostopno na:

<http://www.europeana1914-1918.eu/sl>

- [4] (2012) Europeana. Dostopno na:

<http://www.europeana.eu/portal/>

- [5] (2012) Google Books. Dostopno na:

<http://www.google.com/googlebooks/library.html>

- [6] (2012) Zgodovina OCR. Dostopno na:

http://en.wikipedia.org/wiki/Optical_character_recognition#History

- [7] (2012) Kofax VRS. Dostopno na:

<http://www.kofax.com/vrs-virtualrescan/>

- [8] (2012) Kofax Express. Dostopno na:

<http://www.kofax.com/document-scanning-software/>

- [9] (2012) Kodak Capture Pro. Dostopno na:
http://graphics.kodak.com/docimaging/GB/en/products/software/-document_scanning_software/capture_pro_software/index.htm
- [10] (2012) Abbyy Scan Station. Dostopno na:
http://www.abbyy.com/scan_station/
- [11] (2012) Kofax Desktop. Dostopno na:
http://www.kofax.com/support/Desktop/2.0/2.0_downloads.asp
- [12] (2012) Lura Document PDF Compressor. Dostopno na:
<http://www.luratech.com/en/home/products/software-and-solutions-for-document-processing/document-and-data-conversion-software/luradocument-pdf-compressor.html>
- [13] (2012) Abbyy Fine Reader. Dostopno na:
<http://finereader.abbyy.com/>
- [14] (2012) Abbyy Recognition Server. Dostopno na:
http://www.abbyy.com/recognition_server/
- [15] (2012) Kofax Capture. Dostopno na:
<http://www.kofax.com/document-capture-software/>
- [16] (2012) Abbyy Flexi Capture. Dostopno na:
http://www.abbyy.com/data_capture_software/
- [17] (2012) TIS eFlow. Dostopno na:
<http://www.topimagesystems.com/solutions/overview/eflow-overview>
- [18] (2012) Zakon o arhivskem gradivu in arhivih (ZAGA). Dostopno na:
http://zakonodaja.gov.si/rpsi/r02/predpis_ZAKO482.html
- [19] (2012) Uredba o varstvu dokumentarnega in arhivskega gradiva (UV-DAG). Dostopno na:
http://zakonodaja.gov.si/rpsi/r02/predpis_URED3462.html

- [20] (2012) ISO 19005. Dostopno na:
http://www.iso.org/iso/catalogue_detail?csnumber=38920
- [21] (2012) Zakon o elektronskem poslovanju in elektronskem podpisu (ZEP). Dostopno na:
http://zakonodaja.gov.si/rpsi/r03/predpis_ZAKO1973.html
- [22] (2012) Enotne tehnološke zahteve 2.0 (ETZ v 2.0). Dostopno na:
http://www.arhiv.gov.si/si/zakonodaja_in_dokumenti/-predpisi_s_podrocja_arhivske_dejavnosti_v_sloveniji/
- [23] (2012) MOREQ. Dostopno na:
<http://www.moreq2.eu/moreq2>
- [24] (2012) ISO 15489. Dostopno na:
http://www.iso.org/iso/catalogue_detail?csnumber=31908
- [25] (2012) Uredba o upravnem poslovanju. Dostopno na:
http://zakonodaja.gov.si/rpsi/r02/predpis_URED3602.html
- [26] (2012) PDF / A. Dostopno na:
<http://en.wikipedia.org/wiki/PDF/A>
- [27] (2012) Standard ločljivosti skeniranja. Dostopno na:
<http://scansnapcommunity.com/tips-tricks/1652-why-is-ocr-at-300-dpi-a-standard/>
- [28] (2012) Opis FTP, SFTP, FTPS. Dostopno na:
http://en.wikipedia.org/wiki/File_Transfer_Protocol
- [29] (2012) GnuPG. Dostopno na:
<http://www.gnupg.org/>
- [30] Jure Longyka, Digitalizacija avdiovizualnega arhiva. Diplomsko naloga. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2002.

- [31] (2012) Reference podjetja TIS. Dostopno na:
<http://www.topimagesystems.com/solutions/overview/success-stories>