Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačního inženýrství

Bakalářská práce

Digitalizace faktur a vytěžení metadat pomocí OCR softwaru

Ondřej Kotala

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Provozně ekonomická fakulta

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Ondřej Kotala

Informatika

Název práce

Digitalizace faktur a vytěžení metadat pomocí OCR softwaru

Název anglicky

Digitization of invoices and extraction of metadata with OCR software

---

Cíle práce

Hlavním cílem je využít OCR systému a vytvořit v něm aplikaci pro digitalizaci a vytěžení předem stanovených faktur, faktury budou naskenovány, zpracovány, digitalizovány a bude v nich vytěžena množina metadat.

Vedlejším cílem je shrnout průběh vývoje aplikace a popsat využité technologie.

Metodika

Teoretická část se bude zabývat analýzou odborné literatury, týkající se vývoje aplikace v prostředí OCR softwaru jménem IBM Datacap. Teorie se bude soustředit jak na hlavní část tohoto softwaru, tak i na návazné systémy, které obsahuje balíček IBM Datacap.

Praktická část bude obsahovat vyvíjení aplikace podle nastudované metodiky. Aplikace bude postupně implementována, a nakonec i otestována, vývoj tedy bude probíhat dle metodiky.

Pomocí logiky si utvoříme obecný závěr. Z dosažených vědomostí během implementace popíšeme, pro koho je vlastně aplikace vytvořena a kdo by jí mohl používat.

Doporučený rozsah práce

40-80 stran

Klíčová slova

datacap, digitalizace, ocr, software

Doporučené zdroje informací

BÍLEK, Jan a Jiří ZUZAŇÁK. Rozpoznávání textu v obraze: Optical Character Recognition. Brno: Vysoké učení technické, Fakulta informačních technologií, 2008.

IBM CORPORATION. IBM Datacap 9.1.8., 2022.

IBM CORPORATION. Implementing Mobile Document Capture with IBM Datacap Software. IBM Redbooks, 2015. ISBN 9780738440903.

PORURAN, Sivakumar a Arif MOHAMMED. Basics OF OPTICAL CHARACTER RECOGNITION SYSTEM. Mauritius: LAP LAMBERT Academic Publishing, 2018. ISBN 9786139931460.

Předběžný termín obhajoby

2022/23 ZS – PEF

Vedoucí práce

doc. Ing. Vojtěch Merunka, Ph.D.

Garantující pracoviště

Katedra informačního inženýrství

Elektronicky schváleno dne 7. 3. 2023

Ing. Martin Pelikán, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 13. 3. 2023

doc. Ing. Tomáš Šubrt, Ph.D.

Děkan

V Praze dne 13. 03. 2023

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Digitalizace faktur a vytěžení metadat pomocí OCR softwaru" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu použitých zdrojů na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 15.03.2023 _____

Poděkování

Rád bych touto cestou poděkoval mému vedoucímu práce, panu doc. Ing. Vojtěchu Merunkovi, Ph.D. za vedení práce a odborné konzultace. Také bych rád poděkoval firmě scanservice a. s., jenž mi schválila využít licenci na produkt IBM Datacap.

Digitalizace faktur a vytěžení metadat pomocí OCR softwaru

Abstrakt

Cílem této práce je, za pomocí produktu IBM Datacap, vytvořit aplikaci, jenž bude automaticky převádět předem stanovené dokumenty z analogového stavu do stavu digitálního. Následně bude aplikace těžit metadata, jejichž rozsah bude taktéž předem stanovený. Následovně bude aplikace implementována a otestována tak, aby byla průkazná její funkcionalita.

Klíčová slova: datacap; software; digitalizace; ocr

Digitization of invoices and extraction of metadata with OCR software

Abstract

The goal of this work is to create an application, using the IBM Datacap product, that will automatically convert predefined documents from analog to digital form. The application will then mine predefined metadata. After that, the application will be implemented and tested, so its funcionality is proven.

Keywords: software; digitization; ocr; datacap

Obsah

1 Úvod ... 13

2 Cíl práce a metodika ... 14

2.1 Cíle práce ... 14

2.2 Metodika ... 14

3 Teoretická východiska ... 15

3.1 Digitalizace dokumentů ... 15

3.1.1 Dopady digitalizace ... 15

3.2 Principy digitalizace ... 16

3.2.1 Skenování ... 16

3.2.1.1 Stolní skener ... 16

3.2.1.2 Listové skener ... 17

3.2.1.3 Ruční skenery ... 17

3.2.1.4 Skenery knih ... 17

3.2.2 Dokument a rozdělení dle struktury ... 17

3.2.2.1 Strukturované dokument ... 17

3.2.2.2 Polostrukturované dokument ... 19

3.2.2.3 Nestrukturované dokument ... 20

3.2.3 Metody rozpoznávání ... 22

3.2.3.1 OCR ... 22

3.2.3.2 ICR ... 23

3.2.3.3 OMR ... 24

3.3 IBM Datacap ... 25

3.3.1 O platformě ... 25

3.3.1.1 Architektura ... 25

3.3.1.2 Integrace ... 25

3.3.1.3 Zabezpečení ... 25

3.3.1.4 Automatizace ... 26

3.3.1.5 Analytika ... 26

3.3.1.6 Personalizace ... 26

3.3.1.7 Uživatelské rozhraní ... 26

3.3.1.8 Škálovatelnost ... 26

3.3.1.9 Umělá inteligence ... 27
3.3.1.10 Specializace ... 27
3.3.1.11 Partnerský program ... 27
3.3.2 Komponenty ... 27
3.3.2.1 Datacap Server ... 27
3.3.2.2 Datacap FastDoc ... 28
3.3.2.3 Datacap Studio ... 28
3.3.2.4 Datacap Rulerunner Server ... 28
3.3.2.5 Datacap Web Server ... 28
3.3.2.6 Datacap wTM ... 28
3.3.2.7 Datacap Report Viewer ... 28
3.3.2.8 Verifikační komponenty ... 28
3.3.3 Hierarchie objektů v programu IBM Datacap Studio ... 29
3.3.4 Struktura workflow v programu IBM Datacap Studio ... 30
4 Praktická část ... 32
4.1 Popis problému ... 32
4.2 Stanovení cílů ... 33
4.3 Návrh řešení aplikace ... 34
4.4 Příprava infrastruktury ... 35
4.4.1 Přehled instalovaných komponent ... 36
4.4.2 Příprava databází ... 37
4.4.3 Instalace IBM Datacap ... 37
4.5 Vývoj aplikace ... 38
4.5.1 Založení aplikace ... 38
4.5.2 Vstupní složka ... 38
4.5.3 Digitalizace ... 39
4.5.4 Korekce obrazu ... 41
4.5.5 Vytěžování ... 42
4.5.5.1 Analytické vytěžování ... 42
4.5.5.2 Vytěžování dle šablon ... 44
4.5.6 Čištění dat ... 45
4.5.7 Formátovaní a validace dat ... 45
4.5.8 Export dat ... 46
4.5.9 Panel pro manuální verifikaci ... 48
4.6 Příprava testovacích dokumentů ... 50

4.7 Testování 50

5 Výsledky a diskuse 51

5.1 Výsledky testování 51

5.2 Zhodnocení 51

6 Závěr 53

7 Seznam použitých zdrojů 54

Přílohy 57

Seznam obrázků a zkratek

Seznam obrázků

Obrázek 1 Strukturovaný dokument – Formulář (8) 18
Obrázek 2 Polostrukturovaný dokument – Faktura (9) 19
Obrázek 3 Nestrukturovaný dokument – Smlouva (10) 21
Obrázek 4 Příklad optického rozpoznávání znaků (13) 22
Obrázek 5 Příklad inteligentního rozpoznávání znaků (15) 23
Obrázek 6 Příklad optického rozpoznávání značek (17) 24
Obrázek 7 Ukázka hierarchie objektů (vlastní zdroj) 30
Obrázek 8 Ukázka struktury workflow (vlastní zdroj) 31
Obrázek 9 Návrh řešení – Diagram (vlastní zdroj) 34
Obrázek 10 Diagram instalovaných komponent produktu IBM Datacap (vlastní zdroj) 36
Obrázek 11 Aplikačního průvodce komponenty IBM Datacap Studio (vlastní zdroj) 38
Obrázek 12 Ukázka upraveného setu pravidel pro import dokumentů (vlastní zdroj) 39
Obrázek 13 Diagram procesu digitalizace (vlastní zdroj) 40
Obrázek 14 Nastavení korekce obrazu (vlastní zdroj) 41
Obrázek 15 Ukázka vytvořených polí v dokumentové hierarchii (vlastní zdroj) 42
Obrázek 16 Ukázka analytického vytěžování (vlastní zdroj) 43
Obrázek 17 Ukázka nastavených zón pro metodu ICR (vlastní zdroj) 44
Obrázek 18 Vytvořený set pravidel, který se stará o čištění částek (vlastní zdroj) 45
Obrázek 19 Validační a formátovací funkce (vlastní zdroj) 46
Obrázek 20 Ukázka funkce pro export dat xml formátu (vlastní zdroj) 47
Obrázek 21 Ukázka polí pro verifikační panel (vlastní zdroj) 49
Obrázek 22 Ukázka zachycené dávky s nevalidními daty v poli (vlastní zdroj) 51

Seznam použitých zkratek

OCR – Optical Character Recognition – Optické rozpoznávání znaků

ICR – Intelligent Character Recognition – Inteligentní rozpoznávání znaků

OMR – Optical Mark Recognition – Optické rozpoznávání značek

ERP systém – Enterprise Resource Planning – Podnikový informační systém

SDK – Software development Kit – Sada vývojových nástrojů

API – Application Programming Interface – Rozhraní pro programování aplikací

FTP – File Transfer Protocol – Protokol pro přenos souborů mezi počítači

AI – Artificial Intelligence – Umělá inteligence

ML – Machine Learning – Strojové učení

NLP – Natural Language Processing – Zpracování přirozeného jazyka

OOTB – Out Of The Box – Nativní funkcionalita připravená od výrobce

1 Úvod

V dnešní digitální době se proces správy dokumentů posunul výrazně kupředu. S pokrokem technologie, byl tradiční způsob manipulace s fyzickými dokumenty nahrazen digitální formou. Digitalizace dokumentů se pro podniky a organizace stala nezbytným procesem pro zefektivnění jejich provozu a snížení nákladů. V minulosti byla digitalizace dokumentů považována za luxus, který si mohly dovolit pouze velké korporace. S nyní dosažitelnou, cenově dostupnou a snadno použitelnou technologií skenování se však digitalizace dokumentů stala přístupnou pro podniky všech velikostí.

Cílem této bakalářské práce je prozkoumat proces digitalizace dokumentů, přiblížit proces rozpoznávání a využít těchto znalostí pro tvorbu vlastní aplikace v rámci OCR softwaru IBM Datacap. Účelem této práce je poskytnout vhled do procesu digitalizace, prozkoumat jeho potenciální důsledky a poukázat na důležitost digitalizace dokumentů v dnešním uspěchaném světě.

2 Cíl práce a metodika

2.1 Cíle práce

Hlavním cílem je využít OCR systému a vytvořit v něm aplikaci pro digitalizaci a vytěžení předem stanovených faktur, faktury budou naskenovány, zpracovány, digitalizovány a bude v nich vytěžena množina metadat.

Vedlejším cílem je shrnout průběh vývoje aplikace a popsat využité technologie.

2.2 Metodika

Teoretická část se bude zabývat analýzou odborné literatury, týkající se digitalizace, vývoje aplikace v prostředí OCR softwaru jménem IBM Datacap a ostatním aspektům důležitým pro zdárnou tvorbu praktické části. Teorie se bude soustředit jak na hlavní část tohoto softwaru, tak i na návazné systémy, které obsahuje balíček IBM Datacap.

Praktická část bude obsahovat vyvíjení aplikace podle nastudované metodiky. Aplikace bude postupně implementována, a nakonec i otestována, vývoj tedy bude probíhat dle metodiky.

Pomocí logiky si utvoříme obecný závěr. Z dosažených vědomostí během implementace popíšeme, pro koho je vlastně aplikace vytvořena a kdo by jí mohl používat.

3 Teoretická východiska

3.1 Digitalizace dokumentů

Digitalizace dokumentů je proces převodu fyzických dokumentů do digitální podoby. Tento proces zahrnuje skenování dokumentů a jejich převod na digitální soubory, které lze ukládat, spravovat a přistupovat k nim elektronicky (1).

Výhody digitalizace dokumentů jsou mnohé. Za prvé, k digitalizovaným dokumentům lze přistupovat odkudkoli a kdykoli, což podnikům usnadňuje spolupráci a sdílení informací mezi svými zaměstnanci. Za druhé, digitalizované dokumenty se snáze hledají, což šetří čas a zvyšuje efektivitu. Za třetí, digitalizované dokumenty zabírají méně fyzického prostoru, což snižuje potřebu úložného prostoru a může vést k úspoře nákladů. A konečně, digitalizované dokumenty mohou být bezpečnější než fyzické dokumenty, protože mohou být šifrovány a chráněny hesly.

Digitalizace dokumentů však přináší i svá úskalí. Tento proces může být časově náročný a vyžaduje značné zdroje, zejména pro podniky s velkými objemy dokumentů. Kromě toho může být proces digitalizace složitý a chyby v procesu mohou vést ke ztrátě nebo poškození dokumentů. Potřeba řádných plánů zálohování a obnovy po havárii je zásadní, aby se zajistilo, že se dokumenty neztratí v případě selhání systému.

3.1.1 Dopady digitalizace

Potenciální dopad digitalizace dokumentů je obrovský a může ovlivnit různá odvětví v jiném měřítku. Například ve zdravotnictví může digitalizace dokumentů vést ke zlepšení péče o pacienty, protože k digitalizovaným lékařským záznamům mohou zdravotníci snadno přistupovat (2). V právním odvětví může digitalizace dokumentů vést k rychlejšímu vyhledávání dokumentů, což může zlepšit efektivitu právních procesů. Ve vzdělávacím průmyslu může digitalizace dokumentů vést ke zlepšení výsledků studentů, protože k digitalizovaným vzdělávacím zdrojům mají studenti i učitelé snadný přístup.

3.2 Principy digitalizace

Podle nastudované publikace (3), je zřejmé, že před zahájením procesu digitalizace je důležité proces pečlivě naplánovat. To zahrnuje identifikaci dokumentů, které je třeba digitalizovat, stanovení požadavků na skenování a rozhodnutí o formátu digitalizovaných dokumentů. Proces plánování pomůže zajistit, aby digitalizované dokumenty odpovídaly svému tištěnému originálu.

3.2.1 Skenování

Teorie procesu skenování dokumentu zahrnuje převod fyzického dokumentu do digitálního formátu, který lze ukládat, upravovat a sdílet elektronicky. Skenování dokumentu zahrnuje použití skeneru k zachycení obrazu dokumentu a jeho uložení jako digitálního souboru. Naskenovaný dokument pak může být uložen v počítači nebo v cloudu, takže je snadno přístupný komukoli s patřičnými oprávněními.

Pro skenování dokumentů je na trhu k dispozici několik typů skenerů a nejlepší typ skeneru, závisí na konkrétních potřebách subjektu a jeho preferencích (4) (5) (6). Mezi nejběžnější typy skenerů používaných pro skenování dokumentů patří:

- Stolní skenery
- Listové skenery
- Ruční skenery
- Skenery knih

3.2.1.1 Stolní skener

Ploché skenery jsou nejběžnějším typem skenerů používaným pro digitalizaci dokumentů. Jsou univerzální a mohou skenovat širokou škálu dokumentů, včetně fotografií, knih a dalších položek, které nelze podávat podavačem dokumentů. Ploché skenery jsou nejlepší pro skenování malého množství dokumentů a jsou ideální pro subjekty, kteří potřebují skenovat různé velikosti a typy dokumentů.

V praktické části této práce, budeme využívat technologie stolního skeneru.

3.2.1.2 Listové skener

Listové skenery jsou navrženy pro rychlé a efektivní skenování velkého množství dokumentů. Jsou ideální pro subjekty, kteří potřebují pravidelně skenovat velké objemy dokumentů. Listové skenery mohou skenovat obě strany dokumentu současně a mohou skenovat dokumenty různých velikostí.

3.2.1.3 Ruční skenery

Přenosné skenery jsou kompaktní a lehké, takže jsou ideální pro subjekty, kteří mají omezený prostor pro skener a potřebují skenovat malá množství dokumentů. Přenosné skenery jsou obvykle archové a mohou skenovat obě strany dokumentu současně.

3.2.1.4 Skenery knih

Knižní skenery jsou speciálně navrženy pro digitalizaci vázaných dokumentů, jako jsou knihy a časopisy. Jsou ideální například pro knihovny a muzea. Knižní skenery obvykle používají kolébku ve tvaru písmene “V“, která drží knihu na místě, a mohou skenovat obě strany stránky současně.

3.2.2 Dokument a rozdělení dle struktury

Pro úspěšnou digitalizaci a následnou fázi těžení je důležité si dokumenty rozdělit. Nejčastěji se setkáváme s rozdělením podle dokumentové struktury. Základními typy jsou strukturované, polostrukturované a nestrukturované dokumenty (3).

3.2.2.1 Strukturované dokument

Takovéto dokumenty jsou pro digitalizaci a vytěžování úplně nejlepší. Jedná se totiž o typ, kde všechny informace mají svou přesnou pozici a formát. Díky tomu můžeme nastavit vytěžování s velice vysokou úspěšností. Nejčastějšími příklady jsou formuláře, daňové přiznání a žádanky (7).

Než začnete vyplňovat tiskopis, přečtěte si, prosím, pokyny.

01 Finančnímu úřadu pro / Specializovanému finančnímu úřadu

02 Územnímu pracovišti v, ve, pro

03 Daňové identifikační číslo

C Z

04 Rodné číslo (identifikační číslo)

otisk podacího razítka finančního úřadu

05 Daňové přiznání[1])

řádné opravné dodatečné

Důvody pro podání dodatečného daňového přiznání zjištěny dne

06 Počet příloh

07 Kód rozlišení typu přiznání / datum

/

PŘIZNÁNÍ

k dani silniční za kalendářní rok

podle zákona č. 16/1993 Sb., o dani silniční, ve znění pozdějších předpisů

I. ODDÍL – Údaje o poplatníkovi

Fyzická osoba

08 Příjmení

09 Jméno(-a)

10 Tituly*)

Právnická osoba

11 Název právnické osoby

12 Adresa místa pobytu fyzické osoby / sídla právnické osoby

a) obec

b) PSČ

c) ulice/část obce

d) číslo popisné/orientační

/

13 stát

14 Kontaktní údaje*)

a) telefon

b) e-mail

c) identifikátor datové schránky

25 5407 MFin 5407 - vzor č. 19

[1]) Označte křížkem odpovídající variantu.
*) Označené údaje jsou nepovinné.

Obrázek 1 Strukturovaný dokument – Formulář (8)

3.2.2.2 Polostrukturované dokument

U tohoto typu dokumentů víme, že obsahuje určitou množinu informací, ale nevíme, kde přesně jsou tyto informace obsaženy. Formát faktury je do určité míry dodržován, ale stejná data na různých typech polostrukturovaných dokumentů, mohou být rozdílně pojmenována a rozmístěna (7).

Například u faktur se nejčastěji setkáme se základním údajem jako je číslo faktury. Toto pole může být na jednom layoutu pojmenováno jako “číslo faktury“ na druhém zase jako “faktura č:“. Pro nastavování vytěžovacích postupů a regulárních výrazu, to vytváří určité komplikace.

Dobrým příkladem takovéhoto typu dokumentů jsou právě faktury a objednávky.

COMPANY LOGO

Faktura 183-20118

DAŇOVÝ DOKLAD

DODAVATEL

Martin Pajer
Nákladní 103
101 00 Praha

IČO 87654321
DIČ CZ1212121218

ODBĚRATEL

Firma s.r.o.
Pajerova 123
150 00 Praha

IČO 45126489

Bankovní účet 1234/1234
Variabilní symbol 18320118
Způsob platby Převodem

Datum vystavení 20. 12. 2018
Datum splatnosti 03. 01. 2019
Datum zdan. plnění 20. 12. 2018

Fakturujeme Vám následující položky

			DPH	CENA ZA MJ	CELKEM BEZ DPH
10	hod	Malování zdi	21 %	550,00 Kč	5 500,00 Kč
2	hod	Štukování	21 %	550,00 Kč	1 100,00 Kč

QR Platba

SAZBA	ZÁKLAD	DPH
21 %	6 600,00 Kč	1 386,00 Kč

7 986,00 Kč

Obrázek 2 Polostrukturovaný dokument – Faktura (9)

3.2.2.3 Nestrukturované dokument

Nestrukturované dokumenty, jsou pro vytěžování ty vůbec nejsložitější. Jedná se jednak o časovou náročnost, z důvodů komplikovaného nastavování aplikace pro vytěžení, tak i třízení jednotlivých dokumentů do svých podkategorií. Příčinou těchto komplikací je jejich nestálý a neurčitý formát, taktéž informace, které netušíme, kde a v jakém množství se na tomto typu vyskytují (7).

Klasickým příkladem tohoto typu jsou dopisy, smlouvy a články.

Smlouva o výpůjčce nemovitosti

Smluvní strany:

Jaromíra Čaková
r. č. 505412/112
trvale bytem Klamavá 17, PSČ 252 26, Třebotov
(dále jen „**vypůjčitelka**") na straně jedné

a

Petra Částková
r. č. 765120/3111
trvale bytem Pražská 19, PSČ 252 19, Rudná u Prahy
(dále jen „**půjčitelka**") na straně druhé

dohromady též jako smluvní strany

uzavírají podle § 2193 a násl. zákona č. 89/2012 Sb., občanský zákoník, tuto

smlouvu o výpůjčce

I.

1.1 Půjčitelka prohlašuje, že je výlučnou vlastnicí nemovitostí, a to pozemku p. č. St. 55, jehož součástí je objekt k bydlení č. p. 77, pozemku p. č. 362/3 o výměře 773 m² (zahrada) a pozemku p. č. 455/4 o výměře 678 m² (trvalý travní porost), vše v katastrálním území Roblín, obec Roblín, okres Praha-západ (dále jen „**Nemovitosti**").

II.

2.1 Půjčitelka uzavřela s vypůjčitelkou dne 1. 2. 2019 smlouvu o smlouvě budoucí kupní, na jejímž základě se smluvní strany dohodly na uzavření kupní smlouvy, jejímž obsahem bude úplatný převod vlastnického práva k nemovitostem z půjčitelky na vypůjčitelku (dále jen "**Kupní smlouva**"). Kupní smlouva by měla být za podmínek uvedených ve smlouvě o smlouvě budoucí mezi smluvními stranami uzavřena nejdéle do 14. 12. 2019.

2.2 Vypůjčitelka má však zájem Nemovitosti užívat ještě před uzavřením Kupní smlouvy. Půjčitelka a vypůjčitelka se s ohledem na své dosavadní vztahy a dohody výslovně dohodly, že půjčitelka přenechává po dobu dále stanovenou vypůjčitelce Nemovitosti k bezplatnému užívání.

III.

3.1 Půjčitelka se zavazuje přenechat vypůjčitelce Nemovitosti k bezplatnému užívání ode dne podpisu této smlouvy do dne, kdy se vypůjčitelka stane vlastníkem

Obrázek 3 Nestrukturovaný dokument – Smlouva (10)

3.2.3 Metody rozpoznávání

Rozpoznávání můžeme chápat jako proces, při kterém počítač, za využití optimálně zvolené metody, nejdříve analyzuje strukturu dokumentu, poté postupně rozdělí stránku do bloků, které se dále člení na obrázkové bloky, textové bloky a tabulky. Jednotlivé řádky textu jsou poté rozvrženy do slov a dále na znaky. Jakmile program skončí, s rozdělováním textu na znaky, pokračuje další fází a to porovnáváním (11).

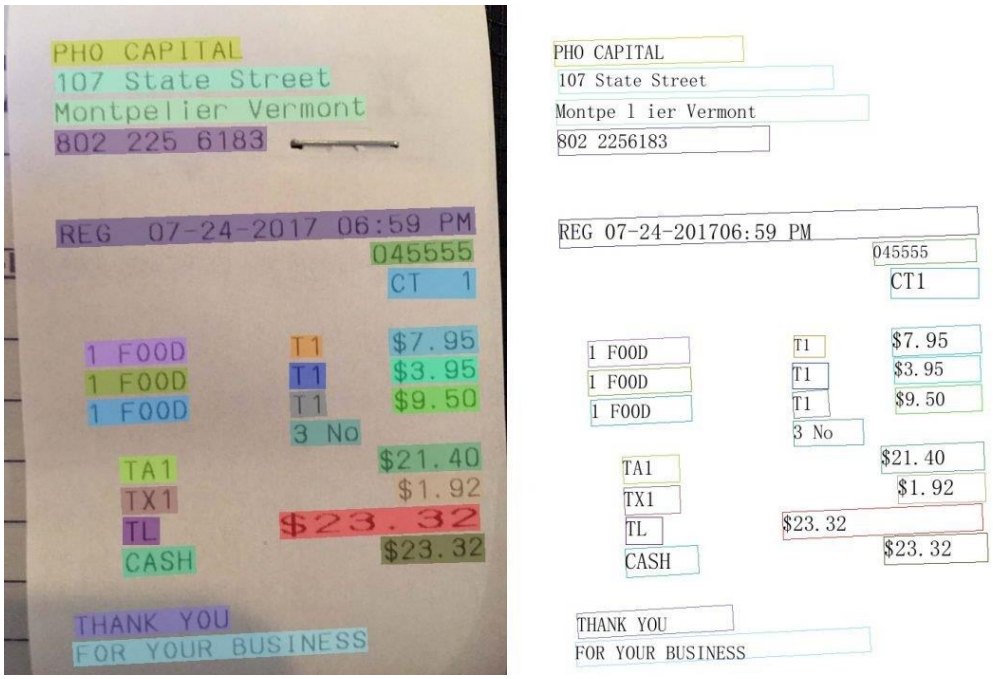
Dle využití a očekáváného výsledku se dají tyto metody dále rozdělit. Dělíme je na:

1) OCR – Optické rozpoznávání znaků – strojově psaný text
2) ICR – Inteligentní rozpoznávání znaků – ručně psaný text
3) OMR – Optické rozpoznávání značek – zaškrtávací pole

3.2.3.1 OCR

Při následném porovnávání se uvažuje v potaz hlavně strojově psaný text. Funguje na bázi uložených šablon a vzorů. Hlavní algoritmus poté porovnává jednotlivé znaky postupně a hledá podobnosti ve vzorech. Pokud podobnost najde, zobrazí nám již rozpoznaný znak.

Díky své jednoduchosti a hardwarové nenáročnosti, dokáže zpracovávat značné objemy textu za velice krátký čas (12).



Obrázek 4 Příklad optického rozpoznávání znaků (13)

3.2.3.2 ICR

Algoritmus analyzuje text v mnoha aspektech a hledá podobnosti v atributech každého znaku, mezi tyto atributy můžeme zařadit zejména křivky, průsečíky a úrovně tloušťky jednotlivých čar znaku.

Všechny výsledky jednotlivých atributů se nakonec zkombinují a vyhodnotí tak, aby nám zobrazený znak, měl co nejvyšší úroveň jistoty.

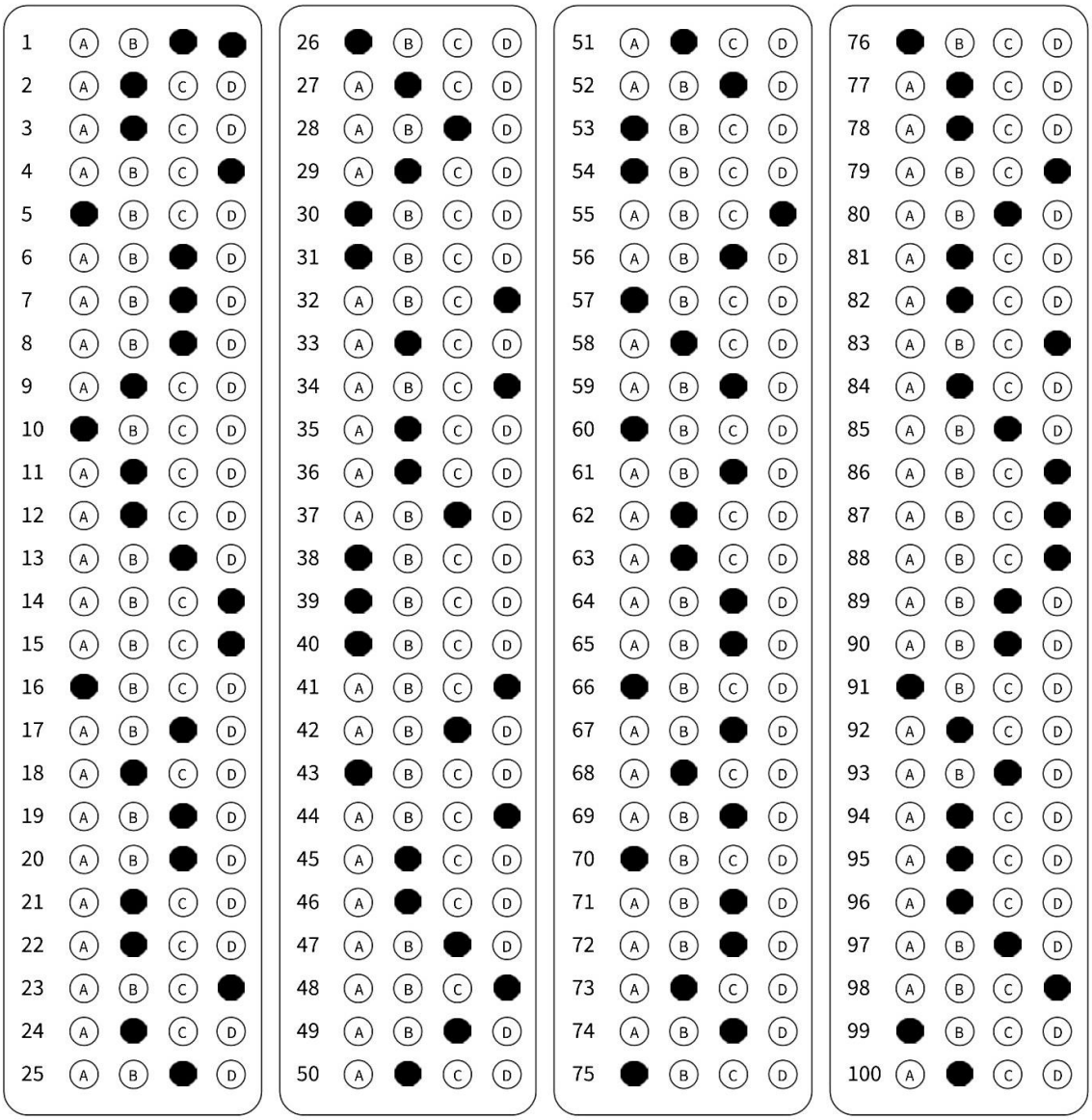
Na rozdíl od klasického OCR, využívá ICR daleko pokročilejších metod a strojového učení, právě díky těmto pokročilejším postupům a výkonnějším algoritmům se využívá pro převedení ručně psaného textu do digitální podoby (14).

Obrázek 5 Příklad inteligentního rozpoznávání znaků (15)

3.2.3.3 OMR

Technologie OMR se ve většině případů, využívá na rozpoznávání zaškrtávacích polí u formulářů. Její hlavní výhodou je rychlost a vysoká přesnost.

Optické rozpoznávání značek nevyužívá šablon, vzorů ani složitých algoritmů, ale pro svou funkčnost využívá speciálně navržených boxů, tzv „checkboxů“, v nichž kontroluje podíl černé a bíle barvy (16).



Obrázek 6 Příklad optického rozpoznávání značek (17)

3.3 IBM Datacap

IBM Datacap je kompletní softwarové řešení pro digitalizaci, vytěžování a zpracování dokumentů na úrovni podniku (18). Pomocí IBM Datacap mohou organizace shromažďovat a digitalizovat různé typy dokumentů, jako jsou faktury, smlouvy, objednávky a další.

Využití tohoto softwaru, najdeme nejen v organizacích různé velikosti, ale i v různých typech odvětví, hlavní výhody použití jsou flexibilita, škálovatelnost a bezpečnost. Tyto výhody pak vedou ke zvýšení provozní efektivity a snížení chybovosti.

Software zvládne extrahovaná data z dokumentů následně vložit do backendových systémů. Extrakce a digitalizace dat funguje za pomocí pokročilých technologií rozpoznávání, včetně optického rozpoznávání znaků (OCR), inteligentního rozpoznávání znaků (ICR) a optického rozpoznávání značek (OMR).

Kromě toho má IBM Datacap nástroje, jako je ověřování dat, klasifikace dokumentů a přesměrování, které zaručují, že data jsou správná, komplexní a budou doručena na správné místo.

3.3.1 O platformě

3.3.1.1 Architektura

Modulární design IBM Datacap umožňuje uživatelům vybrat komponenty, které potřebují k vytvoření řešení pro zachycení a zpracování dokumentů.

Software je k dispozici jak pro lokální, tak i jako cloudové řešení (19).

3.3.1.2 Integrace

IBM Datacap se zvládá připojit k řadě backendových systémů, včetně databází, systémů pro správu obsahu a systémů ERP.

E-mail, FTP a webové služby jsou jen některé ze vstupních a výstupních formátů, které software podporuje.

3.3.1.3 Zabezpečení

Program má silné bezpečnostní funkce, jako je bezpečné ověřování uživatelů, řízení přístupu na základě rolí a šifrování dat při přenosu i v klidu.

Program navíc usnadňuje dodržování ochrany o osobních údajích, tzn. GDPR.

3.3.1.4 Automatizace

Automatizace postupů zpracování dokumentů je možná díky schopnostem tohoto programu. Aby se snížila potřeba zásahu uživatele, program může například automaticky klasifikovat tištěné dokumenty na základě jejich obsahu a přeposílat je na správné místo.

3.3.1.5 Analytika

IBM Datacap má nástroje pro vytváření reportů a analýz, které zákazníkům i podnikovým subjektům umožňují sledovat pracovní postupy zpracování dokumentů, a tím pádem najít i místa, kde se mohou zlepšit.

Kromě jiných parametrů může software vytvářet zprávy o množství zpracovaných dokumentů, dobách zpracování a chybovosti.

3.3.1.6 Personalizace

Program se dodává se sadou pro vývoj softwaru (SDK), která uživatelům umožňuje upravit program tak, aby vyhovoval jejich vlastním unikátním požadavkům.

3.3.1.7 Uživatelské rozhraní

IBM Datacap nabízí uživatelsky přívětivé rozhraní, které uživatelům usnadňuje nastavení a řízení operací při zpracování a validaci dokumentů.

Program má rozhraní drag and drop pro vytváření šablon na následný záchyt dokumentů a webový administrační panel pro ovládání systému, nastavení systému, uživatelů a zabezpečení.

3.3.1.8 Škálovatelnost

Je navržen tak, aby rostl a vyhovoval potřebám velkých podniků s náročnými potřebami při zpracování velkoobjemových dokumentů.

Schopnost softwaru zpracovat obrovské množství dokumentů, při zachování vysoké úrovně výkonu a dostupnosti, je umožněna jeho podporou distribuovaného zpracování, vyvažování zátěže a převzetí služeb při selhání.

3.3.1.9 Umělá inteligence

Celé řešení obsahuje i pokročilé funkce AI, včetně strojového učení (ML) a zpracování přirozeného jazyka (NLP), které mohou zvýšit přesnost a efektivitu zpracování dokumentů. Software může například využívat ML k učení se z komentářů uživatelů a postupně zvyšovat přesnost rozpoznávání, stejně jako NLP k extrakci důležitých informací z nestrukturovaných textů, jako jsou smlouvy, dopisy nebo právní dohody.

3.3.1.10 Specializace

Nabízí řešení specializovaná na požadavky konkrétních odvětví, jako jsou zdravotnictví, bankovnictví, pojišťovnictví a státní správa.

Takovéto systémy se dodávají již s hotovými šablonami pracovních postupů, kritérií ověřování dat a šablonami pro zachycení dokumentů, které jsou nastaveny a přizpůsobeny požadavkům podniku.

3.3.1.11 Partnerský program

Systémoví integrátoři, prodejci a další technologičtí partneři tvoří značnou část partnerského ekosystému platformy, který může produktu nabídnout více služeb a podpory.

Tito partneři mohou pomáhat podnikům s nastavením, přizpůsobením a trvalou podporou IBM Datacap. Mezi tyto partnery spadá i firma scanservice a.s., díky které, tato bakalářská práce mohla vzniknout.

3.3.2 Komponenty

3.3.2.1 Datacap Server

Jedná se o serverovou komponentu, jenž spravuje jednotlivé úlohy dle pořadí a dávky mezi stanicemi a uživateli. Také poskytuje přístupy k frontám, dávkám a databázím (20).

Pro komunikaci mezi dalšími komponenty využívá Datacap Server protokolu Datacap socket. Mezi databázemi využívá Microsoft Object Linking and Embedding.

3.3.2.2 Datacap FastDoc

Hlavním úkolem FastDocu je indexování a skenování. Dá se také použít jako komponenta pro rychlý vývoj algoritmicky nenáročných aplikací (21).

3.3.2.3 Datacap Studio

Datacap Studio je vývojovým a testovacím prostředím pro tvorbu algoritmicky složitějších a funkcionálně rozšířených aplikací (21).

3.3.2.4 Datacap Rulerunner Server

Služba, jenž se při správném nakonfigurování stará o automaticky průchod celé dávky s dokumenty danou aplikací. Průchodem aplikací rozumíme úlohy, jenž nevyžadují interakci s uživatelem. Typicky se jedná o rozpoznávání a export (22).

3.3.2.5 Datacap Web Server

Datacap Web Server hostí Datacap webové aplikace pomocí Microsoft IIS (23).

3.3.2.6 Datacap wTM

Webová služba podporující protokoly HTTP a HTTPS, založená na systému Microsoft Windows, nebo Microsoft IIS (23).

3.3.2.7 Datacap Report Viewer

Komponenta pro reportování veškerých aktivit Datacapu, načítá statistiky ohledně využití, chybovosti a délce běhu, které následně ukládá do databáze. Zprávy o stavu zobrazuje v reálném čase (24).

3.3.2.8 Verifikační komponenty

Pokud jde o ověřování a verifikaci vytěžených dat, IBM Datacap Content Navigator i IBM Datacap Desktop obsahují podobné ověřovací komponenty. Tyto komponenty jsou navrženy tak, aby zajistily, že zachycená data budou přesná a úplná, bez ohledu na použité rozhraní.

Mezi těmito dvěma rozhraními však existují určité rozdíly, které mohou ovlivnit proces ověřování. IBM Datacap Content Navigator je webové rozhraní, které uživatelům umožňuje přístup a správu obsahu odkudkoli s aktivním připojením k internetu. To se hodí zejména v situacích, kdy jednotliví pracovníci, nebo týmy, pracují na obsahu vzdáleně. Naproti tomu IBM Datacap Desktop je aplikace pro Windows, která se instaluje lokálně do počítače uživatele (25) (26).

Dalším rozdílem je, že produkt IBM Datacap Desktop obsahuje rozšířené funkcionality, které nejsou dostupné v produktu IBM Datacap Content Navigator. IBM Datacap Desktop například obsahuje rozhraní drag and drop pro vytváření a úpravu profilů zachycování metadat a další užitečné nástroje pro konfiguraci, rozšíření a správu prostředí.

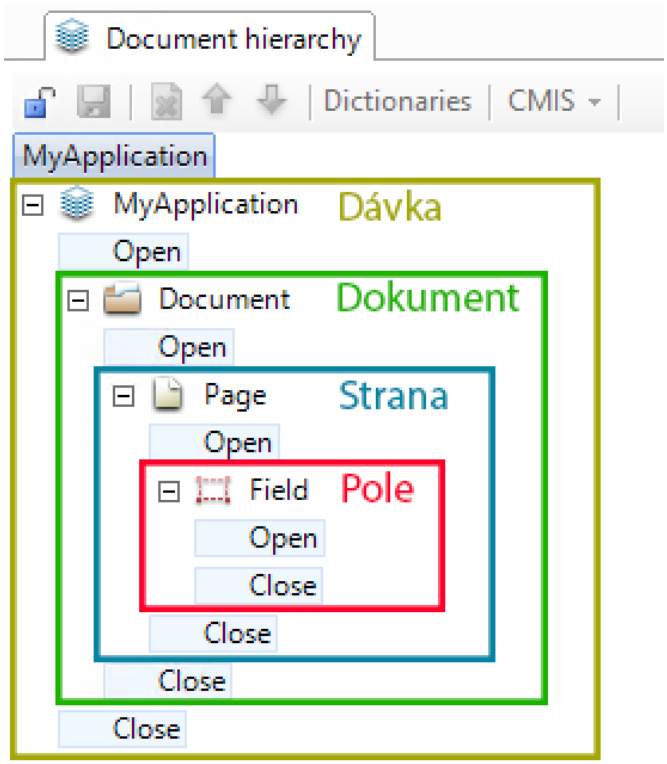
Celkově lze říct, že IBM Datacap Content Navigator i IBM Datacap Desktop poskytují podobné ověřovací komponenty pro zajištění přesnosti a úplnosti zachycených dat. Volba mezi těmito dvěma rozhraními může záviset na faktorech, jako jsou uživatelské preference, potřeba vzdáleného přístupu, nebo speciální funkce požadované pro daný proces verifikace.

3.3.3 Hierarchie objektů v programu IBM Datacap Studio

Hierarchie dokumentů se dá rozdělit na dávku, která obsahuje dokumenty, dokument je složen ze stran a strany obsahují pole informací (3).

Struktura objektů:

- Batch – Dávka
- Document – Dokument
- Page – Strana
- Field – Pole



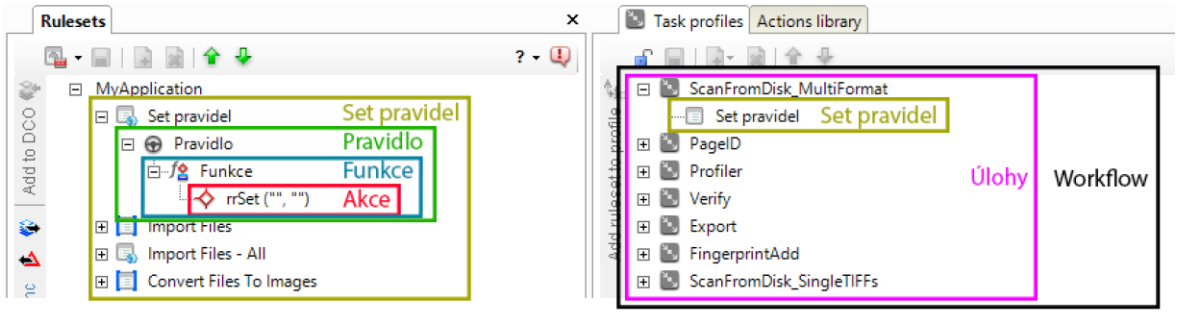
Obrázek 7 Ukázka hierarchie objektů (vlastní zdroj)

3.3.4 Struktura workflow v programu IBM Datacap Studio

Workflow, obsahuje množinu úkolů, úkoly obsahují sety pravidel, sety pravidel jsou složeny z jednotlivých pravidel, pravidla obsahují funkce a funkce jsou složeny z akcí (3).

Struktura tedy vypadá následovně:

- Workflow – Pracovní postup
- Task – Úkol
- Ruleset – Set pravidel
- Rule – Pravidlo
- Function – Funkce
- Action – Akce



Obrázek 8 Ukázka struktury workflow (vlastní zdroj)

4 Praktická část

Cílem praktické části je vytvořit všestrannou aplikaci pro účely osoby blízké (dále jen paní V.). V řešení budou uváženy dokumenty se strojovým textem tak i ručně psané.

Autor práce nejdříve naskenuje dokumenty pomocí domácího skeneru, takto naskenované dokumenty budou poté odeslány do vstupní složky aplikace. Aplikace bude fungovat za pomocí naprogramované workflow.

Ze vstupní složky se dokumenty nahrají do programu, kde se provede digitalizace, korekce obrazu, vytěžení určených metadat, porovnání míry konfidence dat a dle výsledků aplikace rozhodne, zdali je nutné data verifikovat manuálně, či nikoli. Následuje fáze čistících a formátovacích pravidel. Konečným krokem bude export metadat ve zvoleném formátu.

4.1 Popis problému

Zásadním problémem je, že paní V. posílá a skladuje všechny faktury potřebné pro své podnikání ve fyzické podobě, tzn. tištěná papírová forma. Hlavním důvodem rozhodnutí se o digitalizaci je, že množství dokumentů, které musí zařadit do účetnictví a dalších návazných systémů už jenom narůstá a práce s nimi se tak stává daleko časově a finančně náročnější.

Hlavním zájmem paní V. je množina osmi metadat, které z faktur ručně opisuje a zadává do návazného systému manuálně. Jmenovitě se jedná o:

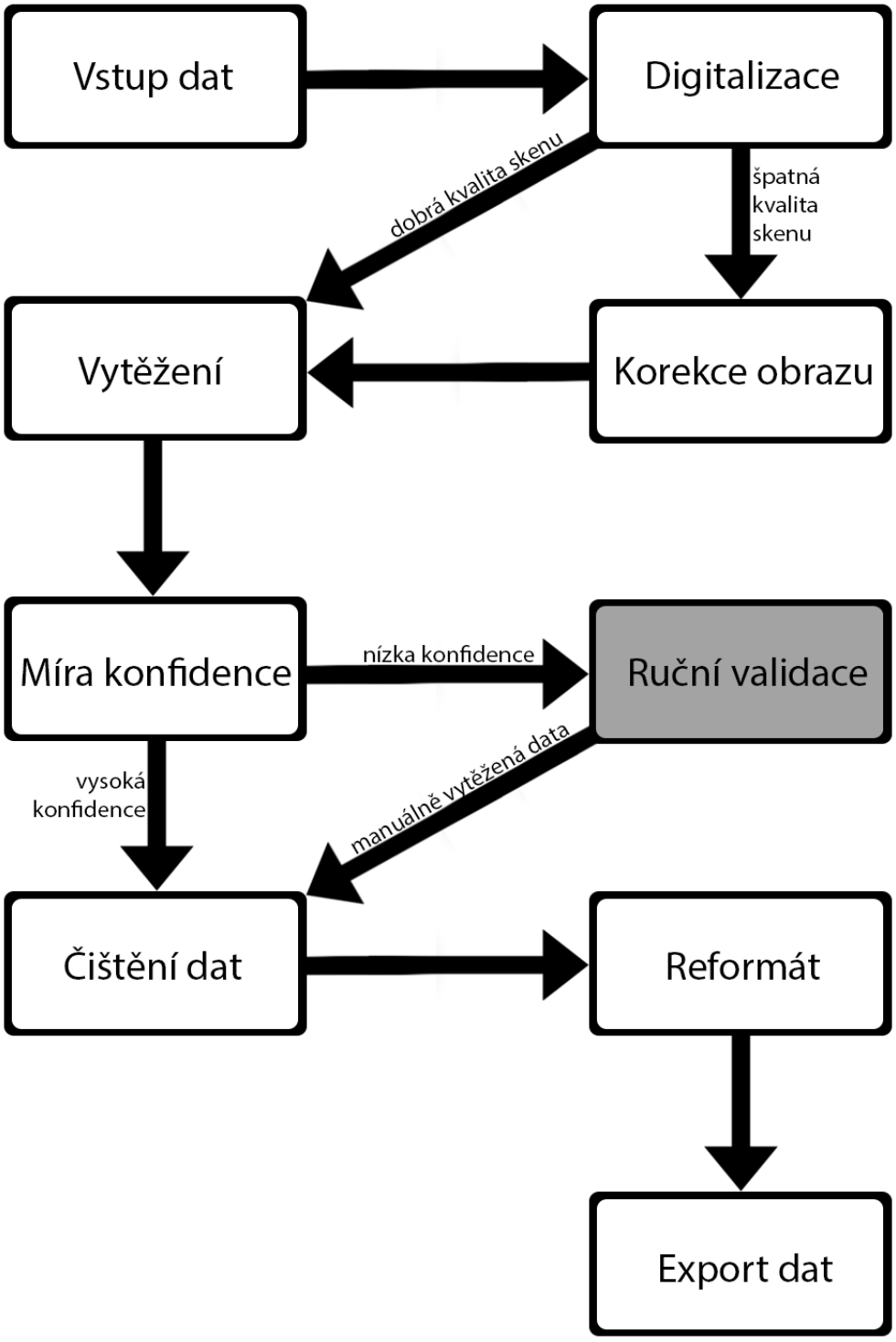
- Jméno dodavatele
- Číslo faktury
- Variabilní symbol
- Datum fakturace
- Adresa dodavatele
- Daň
- Cena (částka bez daně)
- Částka s daní

4.2 Stanovení cílů

Cílem je vymyslet řešení, které bude splňovat tyto požadavky, systematicky jde o:

- Sběr a digitalizaci dokumentů ze vstupní složky
- Korekce obrazu u dokumentů zapříčiní, že následná automatická fáze těžení bude mít daleko vyšší přesnost
- Proces vytěžování bude hledat pouze námi zvolenou množinu metadat
- Na metadatech budou použita formátovací a čistící pravidla
- Aplikace bude pracovat automaticky od vstupu až po export dat, bez zásahu lidské ruky, pouze pokud systém uzná za vhodné, dokumenty nám předá k verifikaci
- Řešení bude obsahovat i verifikační panel pro jednoduchou kontrolu vytěžených metadat

4.3 Návrh řešení aplikace



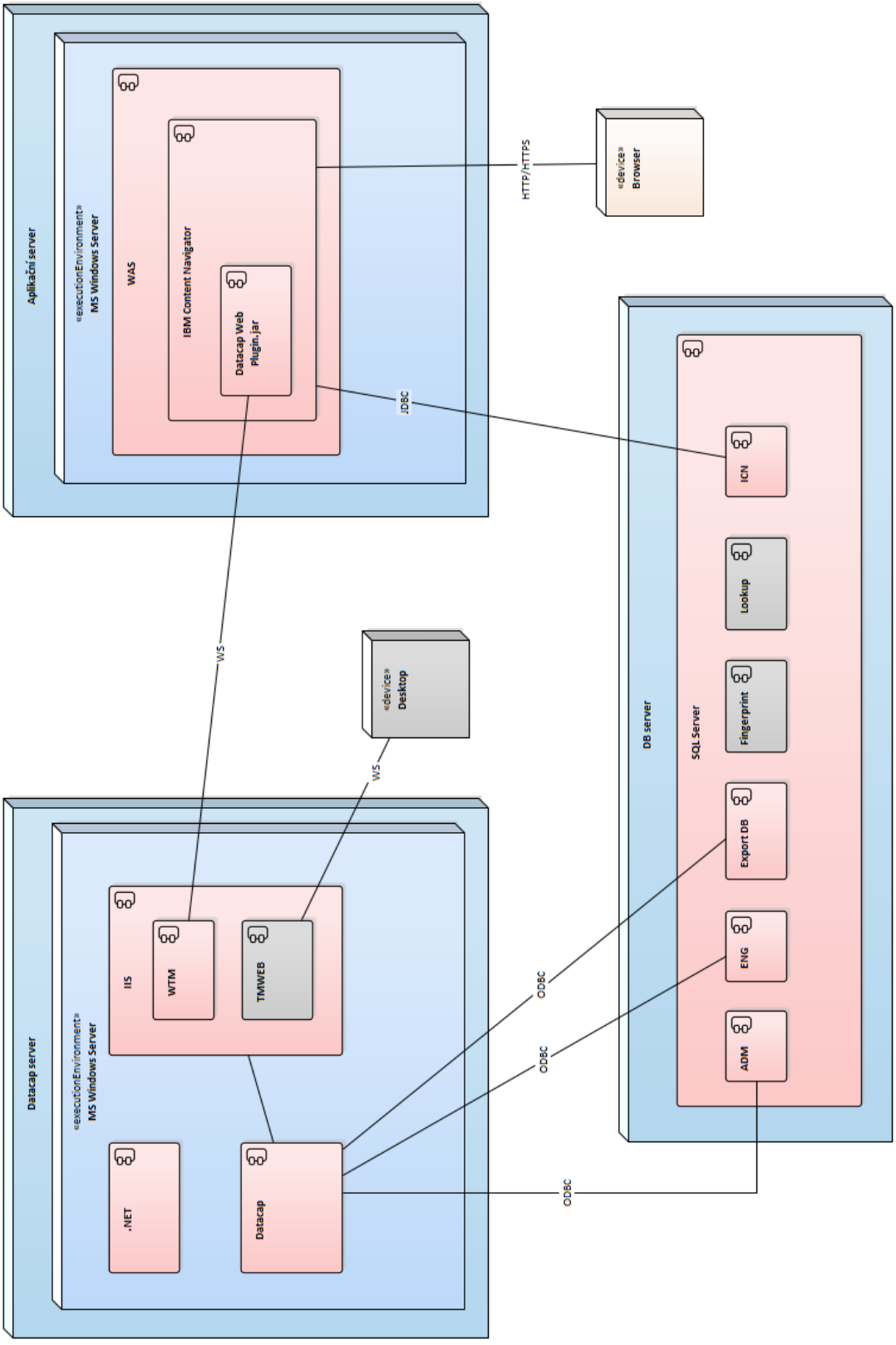
Obrázek 9 Návrh řešení – Diagram (vlastní zdroj)

4.4 Příprava infrastruktury

Nejvhodnější variantou pro přípravu infrastruktury je stejná konfigurace testovacího i produkčního prostředí. V případě že databáze testovacího prostředí budou na jednom serveru, může být výkon a rychlost testovacího serveru značně odlišný než u produkčního (27).

Vzhledem k praktickým a kapacitním důvodům, nebude v této práci brán na tvorbu odlišného databázového serveru zřetel.

4.4.1 Přehled instalovaných komponent



Obrázek 10 Diagram instalovaných komponent produktu IBM Datacap (vlastní zdroj)

- Server pro Datacap
    - Microsoft Windows Server 2016
    - Microsoft.NET 3.5 a 4.6
    - Microsoft IIS + Datacap wTM služby
    - IBM Datacap 9.1.8
- Aplikační server
    - Microsoft Windows Server 2016
    - WAS 9.0.0
    - Java Database Connectivity 4.1
    - Oracle Java ME SDK 8.3
    - IBM Content Navigator 3.0.9
    - Microsoft SQL Server Standard Edition 2017

4.4.2 Příprava databází

Po nainstalování Microsoft SQL Server byly na serverovém prostředí autorem vytvořeny celkem čtyři samostatné databáze.

- ADM – Slouží pro Datacap
- ENG – Slouží pro Datacap
- Export – Stěžejní pro export výsledků z Datacapu
- ICN – Databáze důležitá pro správnou funkčnost IBM Content Navigatoru

Jednotlivé struktury, uživatele a oprávnění k přístupu, si Datacap vytvořil sám. Při tvorbě aplikace stačilo využít funkcionality předem implementovaného aplikačního průvodce.

4.4.3 Instalace IBM Datacap

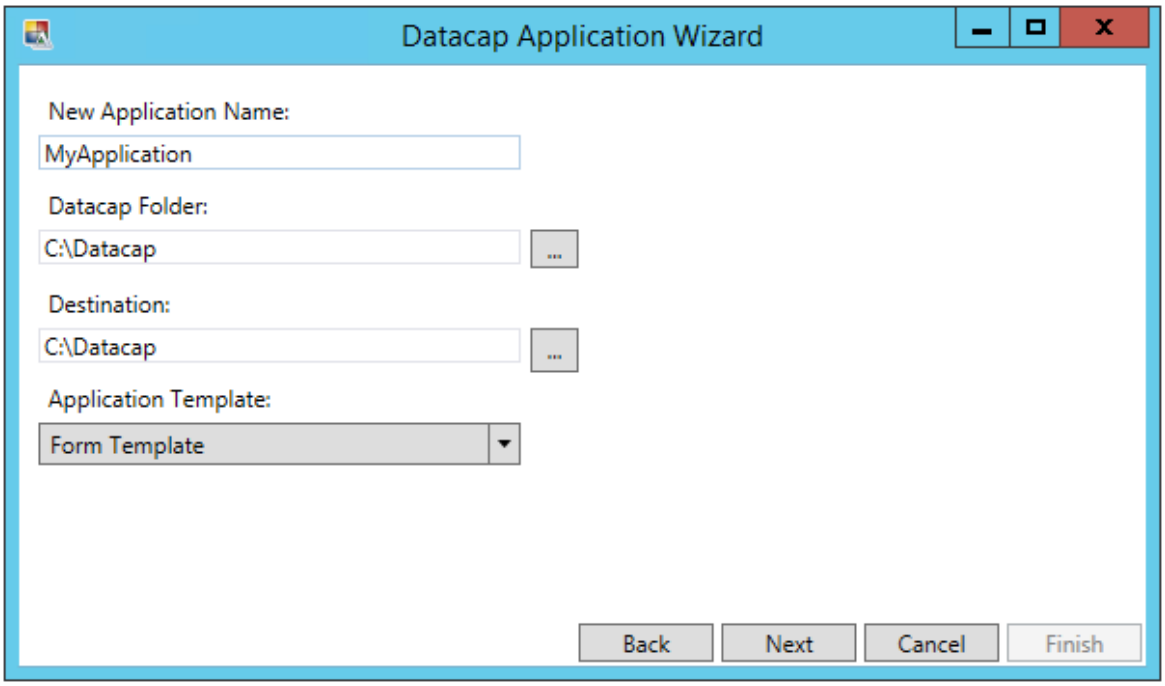
Po přípravě databáze, už zbývalo jenom nainstalovat samotný Datacap, to autor provedl tak, že instalační balíček stáhnul z webu IBM, následně spustil jako administrátor a instalaci nechal proběhnout.

Jako volitelný bylo autorem uvedeno, že při instalaci vybral pouze ty komponenty, jenž bude ve svém řešení využívat, a to z důvodu omezeného místa na serverovém úložišti.

4.5 Vývoj aplikace

4.5.1 Založení aplikace

Pro vytvoření nové aplikace, autor využil funkcionality průvodce Datacap Application Wizard, jenž je součástí komponenty IBM Datacap Studio.



Obrázek 11 Aplikačního průvodce komponenty IBM Datacap Studio (vlastní zdroj)

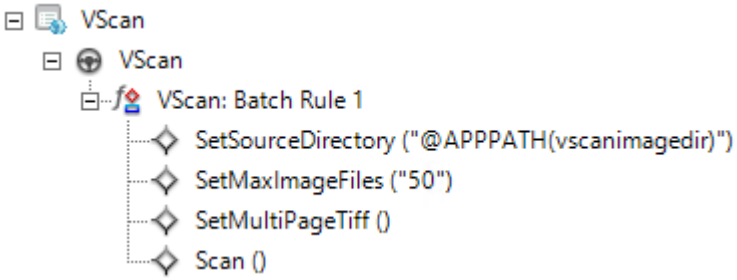
Takto vytvořená aplikace, je připravena pro implementaci řešení a nastavení workflow.

4.5.2 Vstupní složka

Dokumenty aplikace přijímá již v naskenované formě. Proto nám v této práci odpadla nutnost nastavování a propojování domácího skeneru se serverovým prostředím.

Našim požadavkům, nejvíce vyhovoval již předem připravený set pravidel “VScan“, který byl použit a upraven tak, aby kontroloval složku “Input“ v základním adresáři aplikace a následně z ní posouval dokumenty formátu tiff dál do workflow.

Využité OOTB akce v setu pravidel jsou SetSourceDirectory, SetMaxImageFiles a Scan. Časová náročnost návrhu, úpravy a implementace byla jeden den.

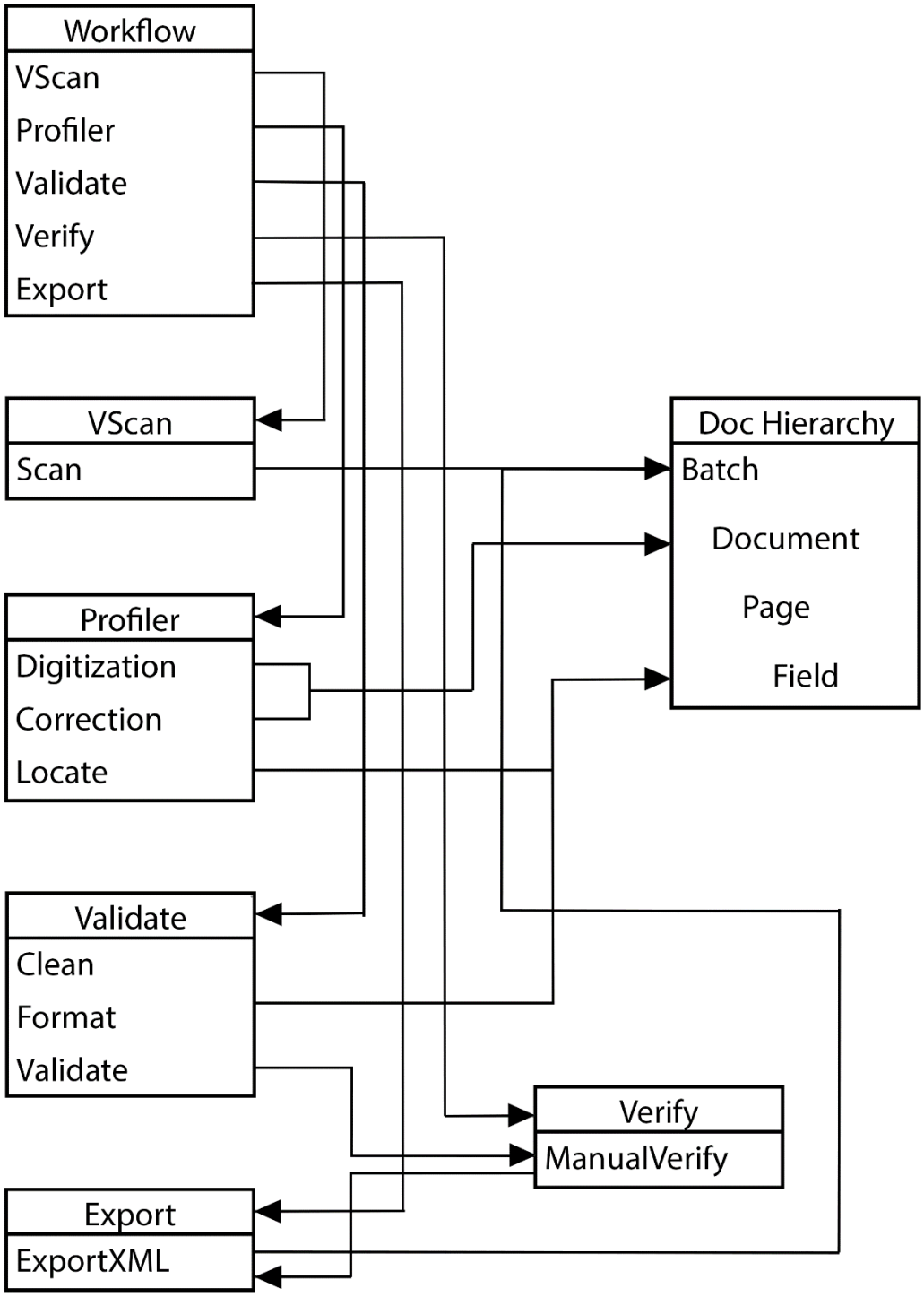


Obrázek 12 Ukázka upraveného setu pravidel pro import dokumentů (vlastní zdroj)

4.5.3 Digitalizace

V procesu digitalizace není potřeba manuálně vybírat metodu rozpoznávání. Datacap si sám načte obsah dokumentu a dle něj vybere nejlepší možnou metodu pro co nejpřesnější rozpoznání. V případě potřeby lze metodu v záložce "properties", pro jednotlivá pole manuálně změnit (přehled metod rozpoznávání byl popsán v teoretické části na str.22).

Vzhledem ke skutečnosti, že řešení je implementováno především pro polostrukturované dokumenty, ve velké části tvořené hlavně strojově psaným textem, tak aplikace automaticky vybrala rozpoznávání dle metody OCR. Na některých polích, kde ale byl očekáváný vstup ručně psaného textu, jsme změnili metodu rozpoznávání na metodu ICR.

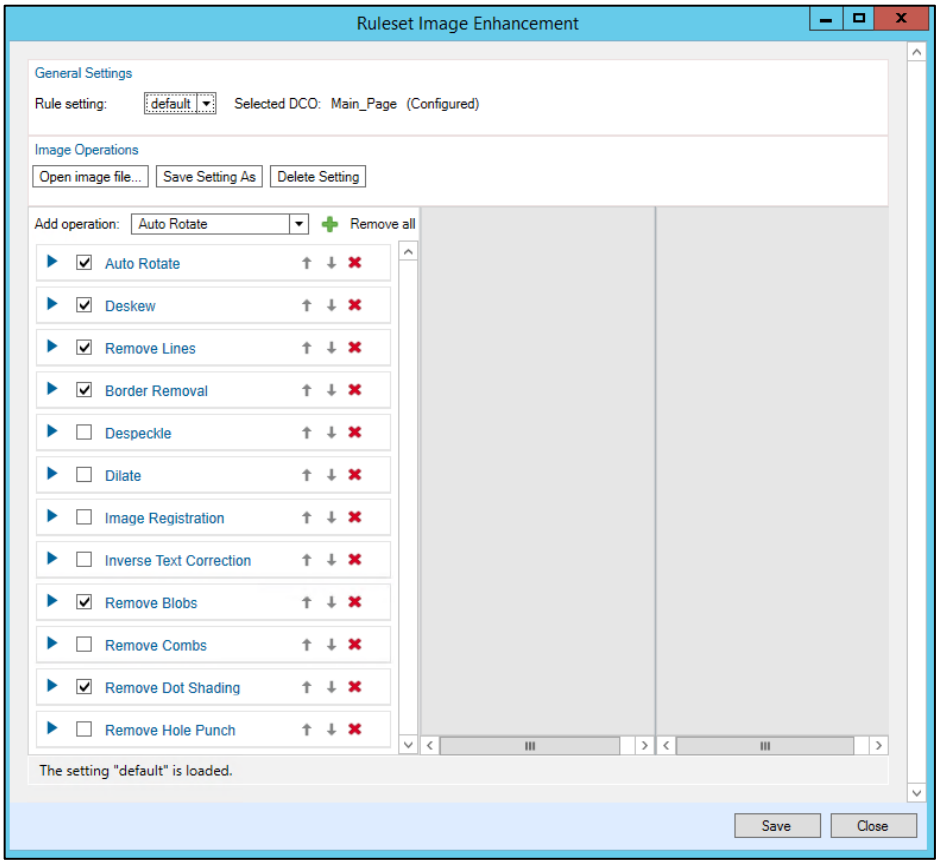


Obrázek 13 Diagram procesu digitalizace (vlastní zdroj)

4.5.4 Korekce obrazu

Pro krok korekce obrazu autor zvolil a nastavil set pravidel “Image Enhancement“, který umožňuje odstranit vodící linky a nežádoucí efekty, jako jsou například skvrny, zrnitost, rozmazanost nebo pootočení jednotlivých stránek v dokumentu při špatném procesu skenování.

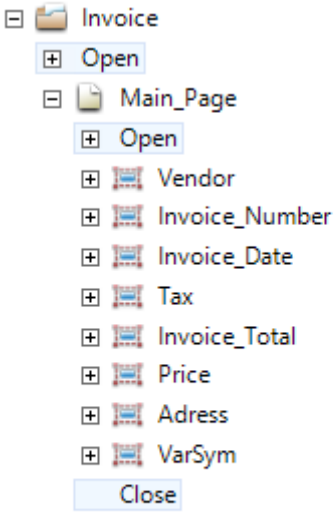
Tento set pravidel využívá souboru předem implementovaných akcí, které se však musí upravit, aby byla zajištěna stoprocentní funkcionalita. Časová náročnost návrhu, úpravy a testování byla jeden den.



Obrázek 14 Nastavení korekce obrazu (vlastní zdroj)

4.5.5 Vytěžování

Autor nejdříve vytvořil strukturu osmi polí do dokumentové hierarchie (Dokumentová hierarchie popsána na str.29). Na takto vytvořené pole se poté přiřadily sety pravidel pro analytické těžení. Níže uvedená pole nám taktéž reprezentují zvolená metadata, které z faktur extrahujeme.



Obrázek 15 Ukázka vytvořených polí v dokumentové hierarchii (vlastní zdroj)

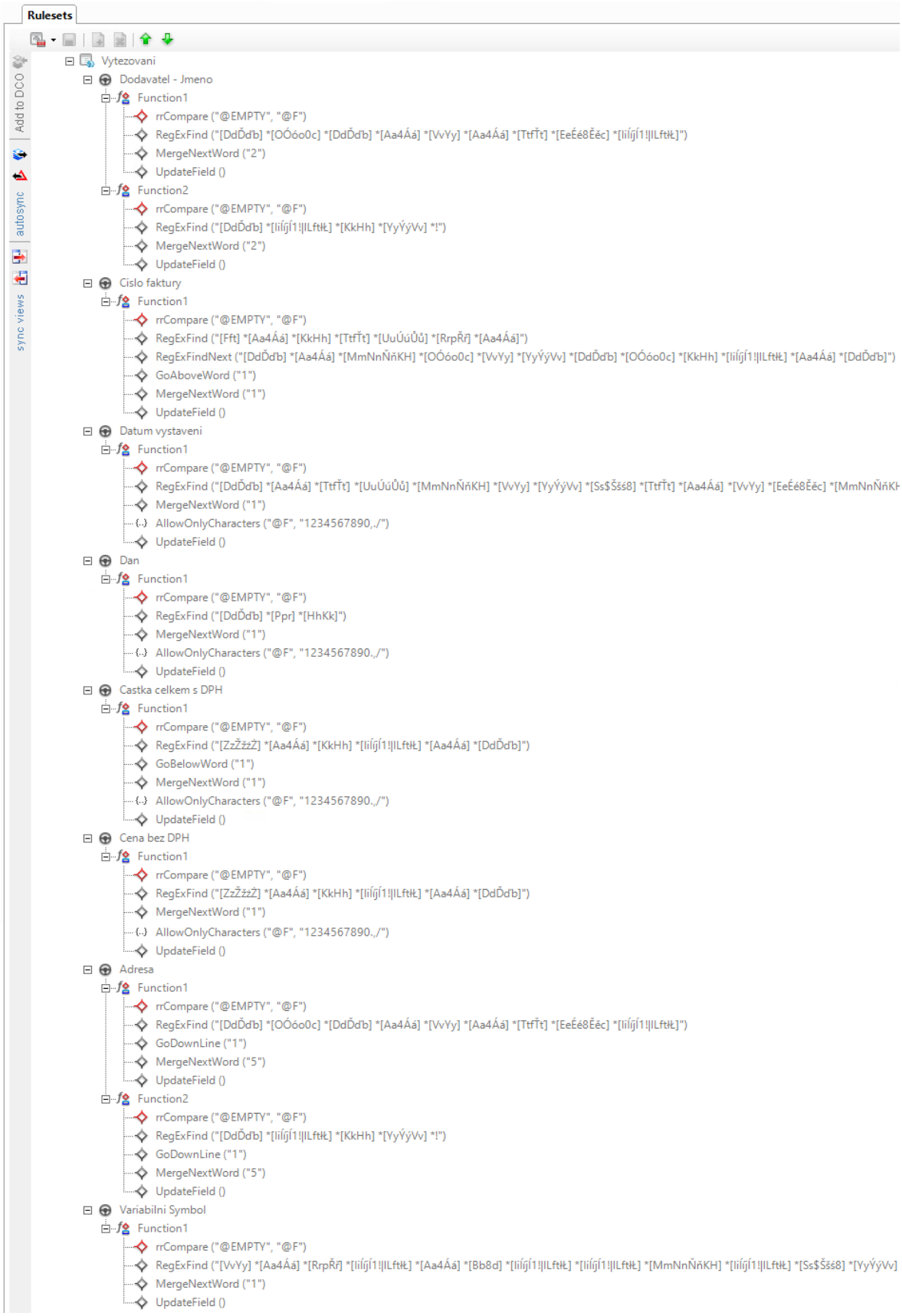
4.5.5.1 Analytické vytěžování

U analytického vytěžování se data získávají užitím akcí. Finální sety pravidel obsahující vytěžovací akce, se následně navázaly na pole, které si je samostatně zavolají a vykonají.

Na výběr bylo z obrovského množství knihoven plných akcí. Autor se rozhodnul pro zvolení takových akcí, aby aplikace zůstala co nejvíce výkonově nenáročná, ale zachovala si svou plnou funkcionalitu.

Ve finálním řešení analytického těžení bylo využito především akce “RegExFind“, která funguje na principu vyhledávání v textu, za pomocí námi určeného regulárního výrazu.

Časová náročnost návrhu, úprav, testování a implementace byla čtrnáct dní.



Obrázek 16 Ukázka analytického vytěžování (vlastní zdroj)

4.5.5.2 Vytěžování dle šablon

Další možností vytěžování dat z faktur je využitím šablon. Pro demonstraci tohoto způsobu autor využil ručně psaných faktur, které jsou součástí testovacích dokumentů a vytvořil pro všechna dílčí metadata, na nichž je aktivována metoda ICR, zóny zájmu. Časová náročnost návrhu, implementace a testování byla jede jeden den.

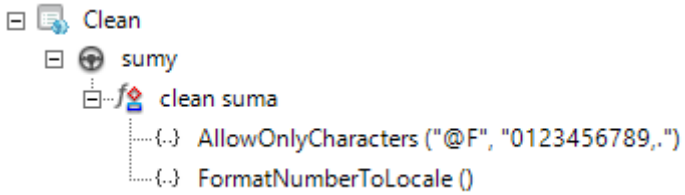
Obrázek 17 Ukázka nastavených zón pro metodu ICR (vlastní zdroj)

4.5.6 Čištění dat

Pokud by aplikace, při některé z faktur, náhodou vytěžila i znaky navíc, je to pro následný export a konzistenci dat problém. Proto je dobré, aby aplikace uměla automaticky kontrolovat a čistit data od zbytečných znaků. Takto čistá data jsou následně připravena na přeformátování, či na konečný export.

Při vytváření pravidla pro čištění dat bral autor v potaz především pole s částkami, jelikož hlavně částky obsahují nejen čísla, ale i měnu, např. Kč, EUR, USD. Při vytěžování by tedy mohla být extrahována suma, i s měnou.

Tomu se autor vyvaroval právě díky vytvoření vlastního pravidla čištění, jenž funguje na bázi načtení extrahované informace a omezení jejího znakového rozsahu.



Obrázek 18 Vytvořený set pravidel, který se stará o čištění částek (vlastní zdroj)

Časová náročnost návrhu, úpravy a implementace byla jeden den.

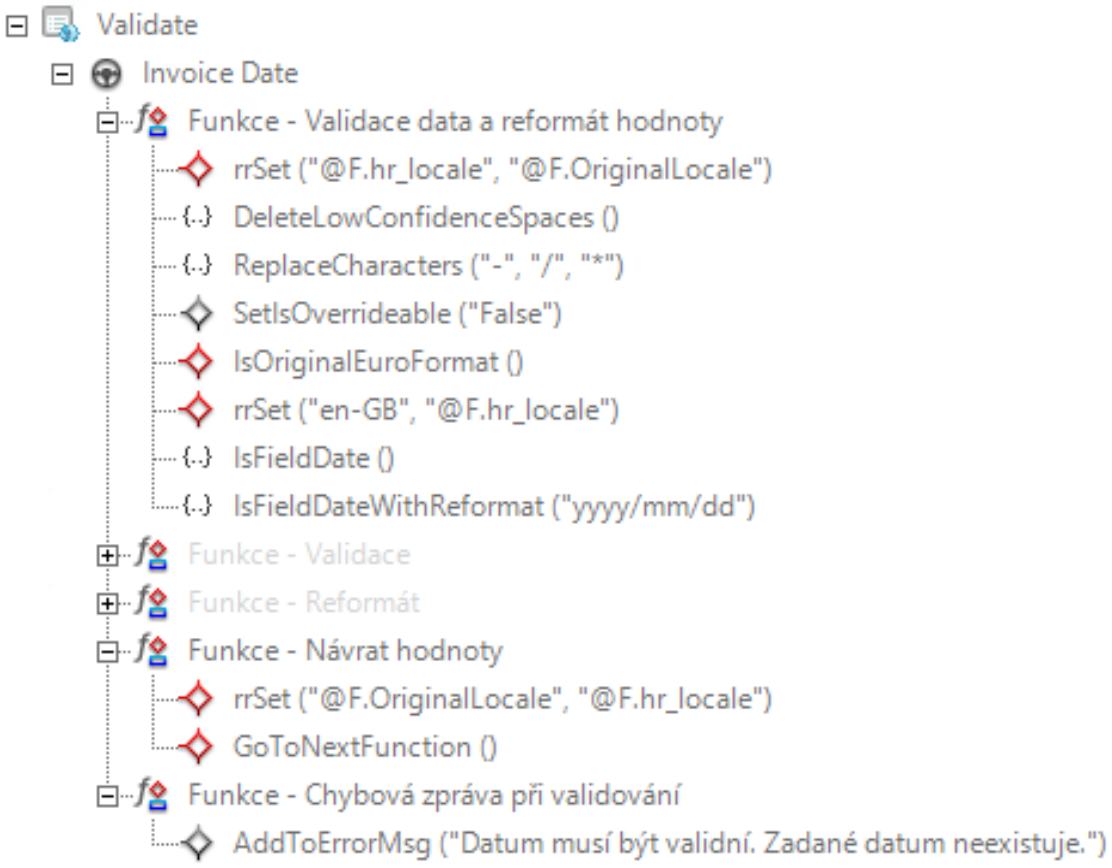
4.5.7 Formátovaní a validace dat

Validování a formátování dat je taktéž nezbytnou součástí řešení. Autor si dal za úkol vytvořit pravidlo, které bude načítat a validovat pole s datumy.

Kontrola datumu funguje na úrovni porovnání hodnoty pole s dostupným číselníkem, výstupem je tedy “pravda“, nebo “nepravda“.

Pokud se funkce vrátí jako “pravda“, tak datum dále putuje ve funkcích na formátové pravidlo, jenž transformuje hodnoty na jednotný formát.

Časová náročnost návrhu, úpravy a implementace byla šest dnů.



Obrázek 19 Validační a formátovací funkce (vlastní zdroj)

4.5.8 Export dat

Do této fáze vstupují data vytěžená, očištěná a ve formátu co nejjednotnějším. Posledním krokem na pořadí byla implementace exportní funkce takto zpracovaných dat.

Exportní funkce a její pravidla byly vytvořeny dle dokumentové hierarchie a data budou uložena ve formátu xml.

Časová náročnosti návrhu, úpravy a implementace byla tři dny.

Systematicky se bavíme o hierarchii:

- Oddíl dávky
    - Název dávky
    - Výpis dokumentů v dávce
    - Oddíl dokumentu

- Název dokumentu
- Výpis stran v dokumentu
- Oddíl stran
    - Typ strany
    - Výpis jednotlivých polí neboli metadat
    - Oddíl pole
        - Název pole
        - Vytěžený obsah pole

Obrázek 20 Ukázka funkce pro export dat xml formátu (vlastní zdroj)

4.5.9 Panel pro manuální verifikaci

Proces vytváření panelu pro manuální validaci dat probíhal pomocí internetového prohlížeče Google Chrome, přes adresu “srv-datacap05:9080/navigator/“. Celková správa komponenty IBM Content Navigator probíhala právě přes internetový prohlížeč.

Po přihlášení do administrace ICN autor vytvořil novou pracovní plochu. Následně vepsal název panelu, do kolonky “ID“ vložil název aplikace z komponenty Datacap Studio a v možnosti “Úložiště“ vybral ze seznamu stejnojmennou aplikaci. Poté už stačilo jenom uložit změny a nová pracovní plocha aplikace byla úspěšně vytvořena.

K následnému nastavení verifikačního panelu bylo využito administrativní konzole ICN. Konzole obsahuje řadu nástrojů a šablon pro jednoduchou implementaci a propojení polí mezi ICN a Datacap Studiem.

Časová náročnosti návrhu, úpravy a implementace byla tři dny.

Obrázek 21 Ukázka polí pro verifikační panel (vlastní zdroj)

4.6 Příprava testovacích dokumentů

Dokumenty pro test aplikace autor získal za pomocí internetových stránek firmy fakturoid.cz, zabývající se tvorbou a správou digitálních faktur. Tato firma nabízí na svých stránkách vzorové faktury ke stažení zdarma. Autor tedy využil všech dostupných šablon, které firma poskytuje a použil je pro přípravu testovacích dokumentů své aplikace. Šablony byly upraveny tak, aby mezi těženými metadaty vznikla určitá variabilita.

Variabilitou metadat myslíme pozměnění jmen a adres dodavatele, změnu variabilních symbolů, částek a data vystavení. Tímto docílíme daleko přirozenějších výsledků a komplexnějšího testování. U vybraných šablon byla metadata psaná strojově odstraněna a následně byla vepsána ručně, tak aby se otestovala i metoda na inteligentní rozpoznávání ručně psaného textu.

Testovací dokumenty byly vytištěny a opětovně naskenovány domácím skenerem, a to z důvodu simulace vstupu faktur do aplikace přes digitalizační médium.

4.7 Testování

Prvotní proces testování proběhl tak, že všechny připravené dokumenty, u nichž bylo jisté, že vyplněná data jsou validní, byly nakopírovány a odeslány do autorem implementované vstupní složky v několika verzích. Datacap si je odtud sám nahrál do aplikace.

Před provedením druhého testu, jsme deaktivovali část řešení, jenž se stará o export dat z aplikace, tím docílíme zachycení dávek s dokumenty přímo ve verifikačním panelu. Tento test je vhodný pro zobrazení dokumentů, které procházejí sice bezchybně, ale nemůžeme u nich zkontrolovat, jak tato data vypadají těsně před exportem.

K třetímu testu autor nahrál ke zpracování dva dokumenty, které byly schválně upraveny tak, aby pole “datum vystavení“ nebylo validní. Autor uvedl datum “29. února 2018“, ale k takovému datu v uvedeném roce nedošlo. Testováním chtěl tedy autor docílit ověření validačního modulu své aplikace.

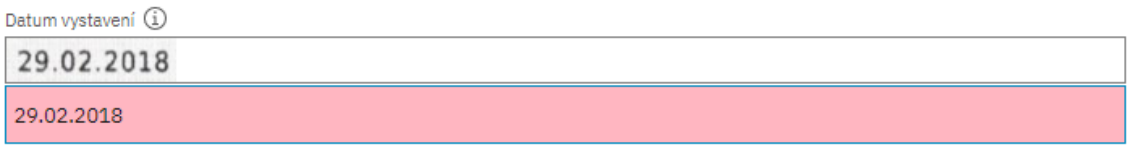
5 Výsledky a diskuse

5.1 Výsledky testování

U prvního testu prošly dávky s dokumenty od vstupu dat, přes digitalizaci, korekci obrazu, vytěžování až k čistícím a formátovým pravidlům. Dávka se všemi dokumenty a metadaty automaticky došla až k exportnímu setu pravidel a vyexportovala se do požadovaného souboru ve formátu xml. Po prozkoumání exportního souboru, nám bylo jasné že struktura byla vytvořena exaktně, tak jak jsme vyžadovali a vytěžená data, jež byla extrahována, jsou taktéž pravdivá a přesná. Obsah vyexportovaného xml souboru je přiložen v příloze A.

U testu č. 2, jsme zkoumali, jak vypadají správně zpracovaná a vytěžená metadata, která by v normálních podmínkách Datacap rovnou vyexportoval. Dokumenty byly opět vloženy do vstupní složky, odkud si je nahrála aplikace. Za krátko už byly doklady zpracovány a odeslány do verifikačního panelu, kde jsme mohli posoudit vzhled dokumentů a stav jednotlivých metadat před exportem. Výsledkem byla stoprocentní úspěšnost vytěžování na námi zvolených testovacích dokladech. Výsledky jsou přiloženy v příloze B.

U posledního testu jsme se soustředili hlavně na funkčnost validace. Dokumenty prošly až k modulu s formátovými pravidly, kde se dávka s dokumenty automaticky odeslala k manuální verifikaci. Jak můžeme na níže uvedené ukázce vidět, validační modul správně zachytil takto nevalidní data a vyžadoval po nás manuální opravu. Výsledky tohoto testu jsou přiloženy v příloze C.



Obrázek 22 Ukázka zachycené dávky s nevalidními daty v poli (vlastní zdroj)

5.2 Zhodnocení

Aplikace splňuje všechny požadavky, které byly stanoveny při vytváření návrhu. Testování modulů probíhalo dle dokumentace k produktu (28). Aplikace funguje automaticky od

vstupu až po export dat, tato funkčnost byla rovněž otestována s kladným výsledkem. Zaměření této aplikace je především pro vytěžování definované množiny metadat u polostrukturovaných dokumentů, zejména faktur.

6 Závěr

V teoretické části práce byly prozkoumány a popsány všechny potřebné prerekvizity pro přiblížení principu digitalizace a metod rozpoznávání obrazu. Dále se teoretická část věnovala popisu platformy IBM Datacap a shrnutí funkcionalit jednotlivých komponent. Tyto teoretické znalosti byly následně využity pro tvorbu řešení v praktické části této bakalářské práce.

V praktické části byla provedena analýza stanoveného problému a následná definice klíčových cílů, které byly stěžejní pro správnou funkčnost řešení. Při implementaci byl na tyto cíle brán zřetel. Jako platforma, na které se bude aplikace vyvíjet a následně fungovat, byl zvolen produkt IBM Datacap ve verzi 9.1.8. Tento produkt byl nainstalován na serverovém prostředí spolu, se všemi nepostradatelnými subsystémy. Na platformě byla založena aplikace, do které byly postupně implementovány jednotlivě navržené úkoly, pravidla a akce. Jednalo se především o vstup dat, digitalizaci a korekci obrazu. Dalším implementovaným úkolem aplikace bylo vytěžování, které autor rozdělil mezi analytické a šablonové vytěžování. Analytické vytěžování bylo vytvořeno na principu mapování slov okolo námi hledané informace. Mapování bylo provedeno za pomocí regulárního výrazu a následná izolovaná informace byla poté uložena do hodnoty pole. Šablonové vytěžování zase porovnávalo šablonu faktury, uloženou v databázi, s fakturou nově příchozí. Pokud aplikace vyhodnotila že dokumenty jsou obdobné, provedla hledání metadat na námi přesně zvolených zónách. Takto vytěžené informace byly očištěny od nadbytečných znaků a reformátovány na jednotné formáty. Zpracovaná data následně aplikace vyexportovala do xml souboru.

V testovací fázi bylo ověřeno, že aplikace umí zdigitalizovat a zpracovat jak strojově, tak i ručně psaný text. Na dokumentech zvládne vytěžit stanovená metadata exaktně a nad rámec původních cílů byla přidána funkcionalita čištění, formátování a exportování dat.

Cíle stanovené v zadání práce byly splněny, jako řešení otestovány a jsou plně funkční.

7 Seznam použitých zdrojů

1) ABDUL MALAK, Haissam. What is Document Digitization? Why Is It Important?. *The ECM Consultant* [online]. [cit. 2023-03-15]. Dostupné z: https://theecmconsultant.com/what-is-document-digitization/
2) Elektronické zdravotnictví. *Digitální česko* [online]. [cit. 2023-03-15]. Dostupné z: https://www.digitalni-cesko.eu/udalost/elektronicke-zdravotnictvi
3) CHEN, Whei-Jen, Ben ANTIN, Kevin BOWE, Ben DAVIES, Jan DEN HARTOG, Daniel OUIMET a Tom STUART. *Implementing Document Imaging and Capture Solutions with IBM Datacap* [online]. 2. IBM, 2015 [cit. 2023-03-15]. ISBN 9780738440903. Dostupné z: https://www.redbooks.ibm.com/abstracts/sg247969.html
4) Typy skenerů a jejich využití. *PREMO* [online]. [cit. 2023-03-15]. Dostupné z: https://www.premocz.eu/typy-skeneru-a-jejich-vyuziti
5) Katalog skenerů. *Scanservice a.s.* [online]. [cit. 2023-03-15]. Dostupné z: https://scanservice.cz/scs-hardware/katalog-skeneru/
6) Knižní skenery. *Scanservice a.s.* [online]. [cit. 2023-03-15]. Dostupné z: https://scanservice.cz/scs-hardware/knizni-skenery/
7) POLANSKÝ, Petr. Vybíráte vytěžovací nástroj? 3 typy dokumentů dle struktury. *EXON s.r.o.* [online]. [cit. 2023-03-15]. Dostupné z: https://www.exon.cz/cs/blog/vytezovani-dat-dle-struktury-dokumentu
8) Přiznání k dani silniční – Finanční správa ČR (FS ČR). *Businessinfo* [online]. [cit. 2023-03-15]. Dostupné z: https://storage.googleapis.com/businessinfo_cz/2021/12/37568ca4-zzform-fs-priznani-dan-silnicni-5407-19.pdf
9) Vzor faktury – Solaris. *Fakturoid* [online]. [cit. 2023-03-15]. Dostupné z: https://www.fakturoid.cz/images/screenshots/invoice-template/fa-solaris.pdf
10) Smlouva o výpůjčce nemovitosti – vzor smlouvy podle nového občanského zákoníku ke stažení. *Vzorovedokumenty* [online]. [cit. 2023-03-15]. Dostupné z: https://www.vzorovedokumenty.cz/pictures/329292/preview01.JPG?1606384190
11) What are OMR, OCR, and ICR? A Helpful Guide. *Remark* [online]. [cit. 2023-03-15]. Dostupné z: https://remarksoftware.com/blog/2021/07/what-are-omr-ocr-and-icr-a-helpful-guide/

12) What Is Optical Character Recognition (OCR)?. *IBM* [online]. [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/cloud/blog/optical-character-recognition
13) PaddleOCR: Awesome multilingual OCR toolkits. *Reddit* [online]. [cit. 2023-03-15]. Dostupné z: https://miro.medium.com/v2/resize:fit:720/format:webp/1*bW-1EEIFuKz6_yjjyvgU0A.jpeg
14) Co je ICR?. *EXON s.r.o.* [online]. [cit. 2023-03-15]. Dostupné z: https://www.exon.cz/cs/blog/co-je-icr
15) BLOOD, Brad. What is Intelligent Character Recognition (ICR)? ICR vs OCR. *Bisok* [online]. 2021 [cit. 2023-03-15]. Dostupné z: https://blog.bisok.com/hs-fs/hubfs/blog-images/icr-example.gif?width=550&name=icr-example.gif
16) Optical Mark Recognition. *GdPicture* [online]. [cit. 2023-03-15]. Dostupné z: https://www.gdpicture.com/guides/gdpicture/Optical%20Mark%20Recognition.html
17) Remark Office OMR Software. *SimpleOCR* [online]. [cit. 2023-03-15]. Dostupné z: https://i.stack.imgur.com/YX8SZ.jpg
18) Datacap functional overview. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=overview-datacap-functional
19) Planning your system architecture. *IBM* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.6?topic=system-planning-your-architecture
20) Datacap Server. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=components-datacap-server
21) Datacap Clients. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=components-datacap-clients
22) Datacap Rulerunner Server. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=components-datacap-rulerunner-server
23) Datacap Web Server and Web Services. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=components-datacap-web-server-web-services
24) Datacap Report Viewer. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=components-datacap-report-viewer

25) IBM Content Navigator overview. IBM. *IBM.com* [online]. 2012 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/content-navigator/2.0.3?topic=overview-content-navigator
26) Running tasks with Datacap Desktop. IBM. *IBM.com* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8?topic=applications-running-tasks-datacap-desktop
27) BOWE, Kevin. *IBM FileNet Capture and IBM Datacap* [online]. IBM, 2015 [cit. 2023-03-15]. ISBN 9780738454320. Dostupné z: https://www.redbooks.ibm.com/redpapers/pdfs/redp5236.pdf
28) IBM Datacap 9.1.8. *IBM* [online]. 2019 [cit. 2023-03-15]. Dostupné z: https://www.ibm.com/docs/en/datacap/9.1.8

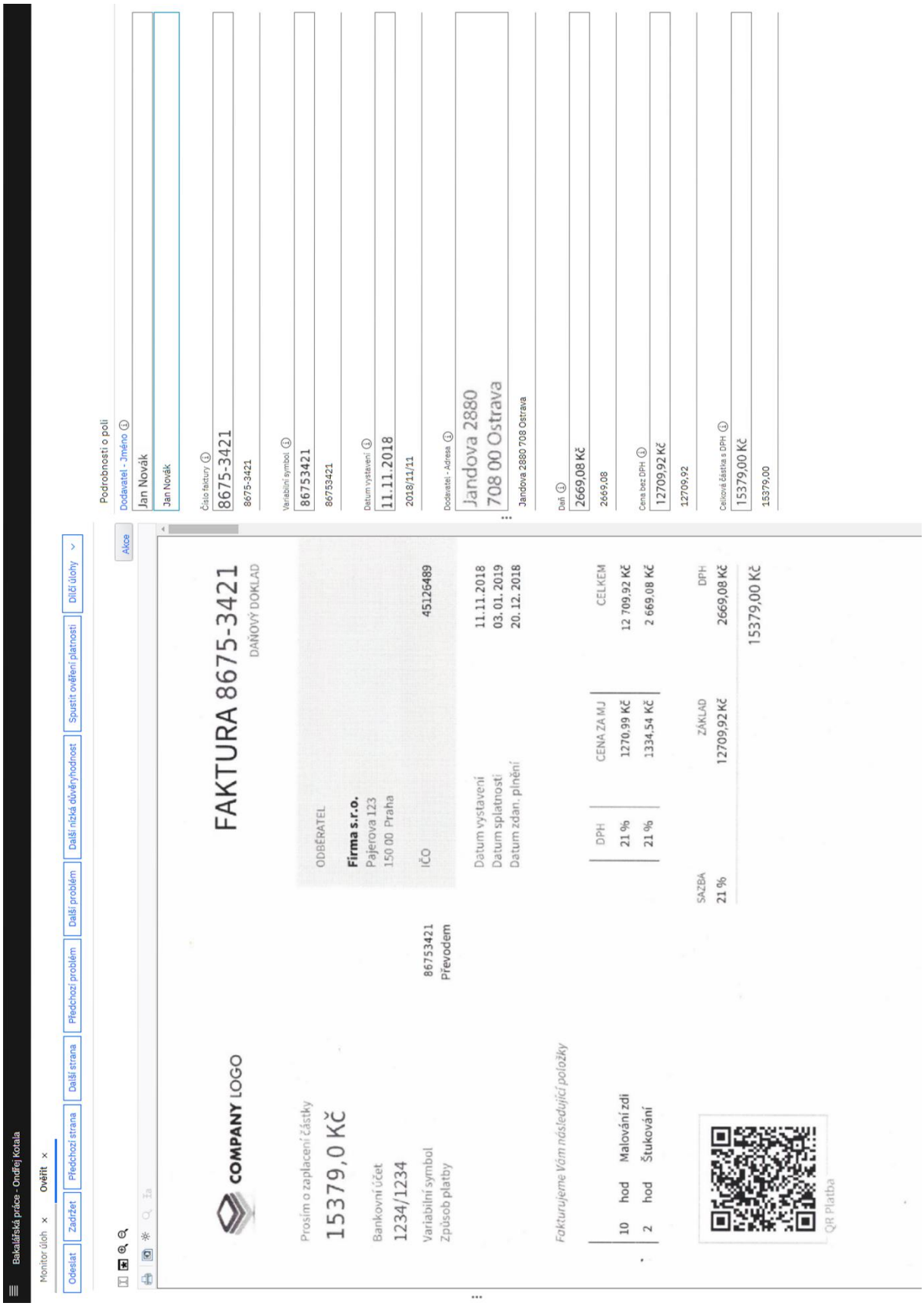
Přílohy

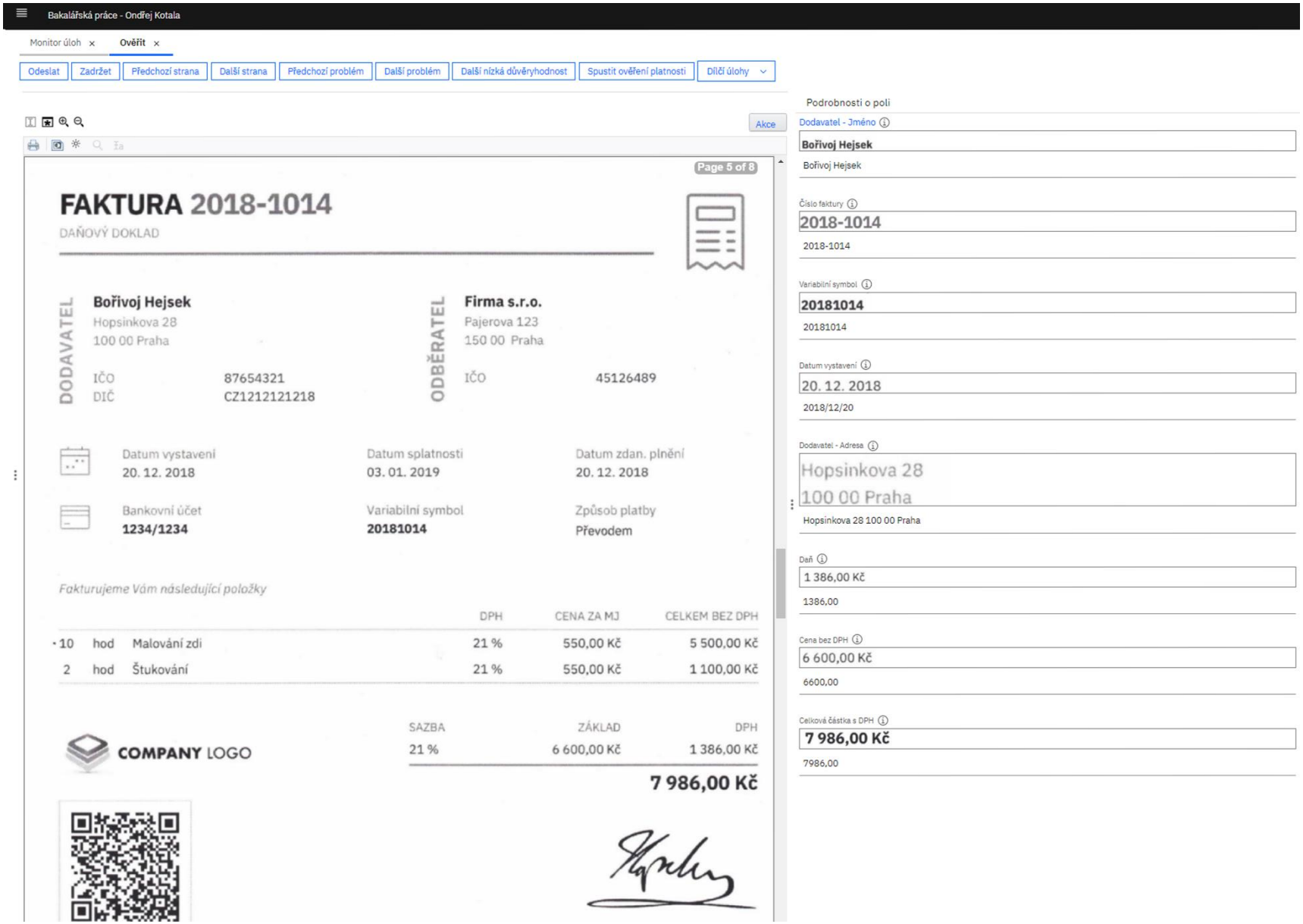
Příloha A Obsah vyexportovaného xml souboru z prvního testu

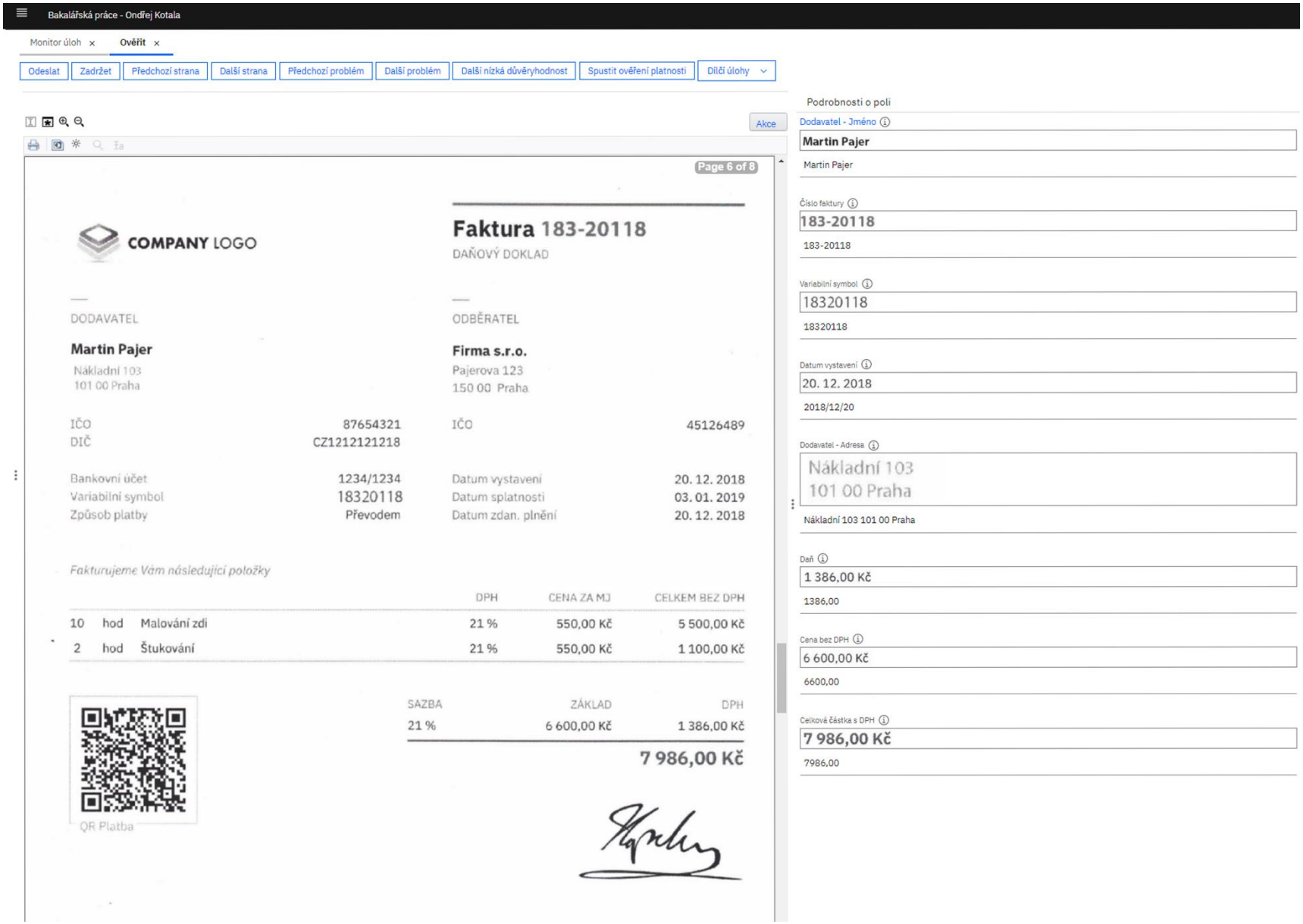
```
<?xml version="1.0" encoding="UTF-8"?>
<Batch Id="20230315.000000">
 <Document Type="Invoice" Id="20230315.000000.01">
  <Page Type="Main_Page">
   <Field Name="Vendor">Jan Novák</Field>
   <Field Name="Invoice_Number">8675-3421</Field>
   <Field Name="Invoice_Date">2018/11/11</Field>
   <Field Name="Tax">2669,08</Field>
   <Field Name="Invoice_Total">15379,00</Field>
   <Field Name="Price">12709,92</Field>
   <Field Name="Adress">Jandova 2880 708 Ostrava</Field>
   <Field Name="VarSym">86753421</Field>
  </Page>
 </Document>
 <Document Type="Invoice" Id="20230315.000000.02">
  <Page Type="Main_Page">
   <Field Name="Vendor">IMRICH BĚLEHLAV</Field>
   <Field Name="Invoice_Number">201945170</Field>
   <Field Name="Invoice_Date">29.2.2018</Field>
   <Field Name="Tax">1470,00</Field>
   <Field Name="Invoice_Total">8470,00</Field>
   <Field Name="Price">7000,00</Field>
   <Field Name="Adress">NÁRODNÍ 204 110 00 PRAHA</Field>
   <Field Name="VarSym">
   </Field>
  </Page>
 </Document>
 <Document Type="Invoice" Id="20230315.000000.03">
  <Page Type="Main_Page">
   <Field Name="Vendor">Imrich Bělehlav</Field>
   <Field Name="Invoice_Number">201945170</Field>
   <Field Name="Invoice_Date">29.02.2018</Field>
   <Field Name="Tax">1470,00</Field>
   <Field Name="Invoice_Total">8470,00</Field>
   <Field Name="Price">7000,00</Field>
   <Field Name="Adress">Národní 204 110 00 Praha</Field>
   <Field Name="VarSym">20114760</Field>
  </Page>
 </Document>
 <Document Type="Invoice" Id="20230315.000000.04">
  <Page Type="Main_Page">
   <Field Name="Vendor">BOŘIVOJ HEJSEK</Field>
   <Field Name="Invoice_Number">2018-1014</Field>
   <Field Name="Invoice_Date">2020/12/20</Field>
   <Field Name="Tax">1386,00</Field>
   <Field Name="Invoice_Total">7986,00</Field>
   <Field Name="Price">6600,00</Field>
```

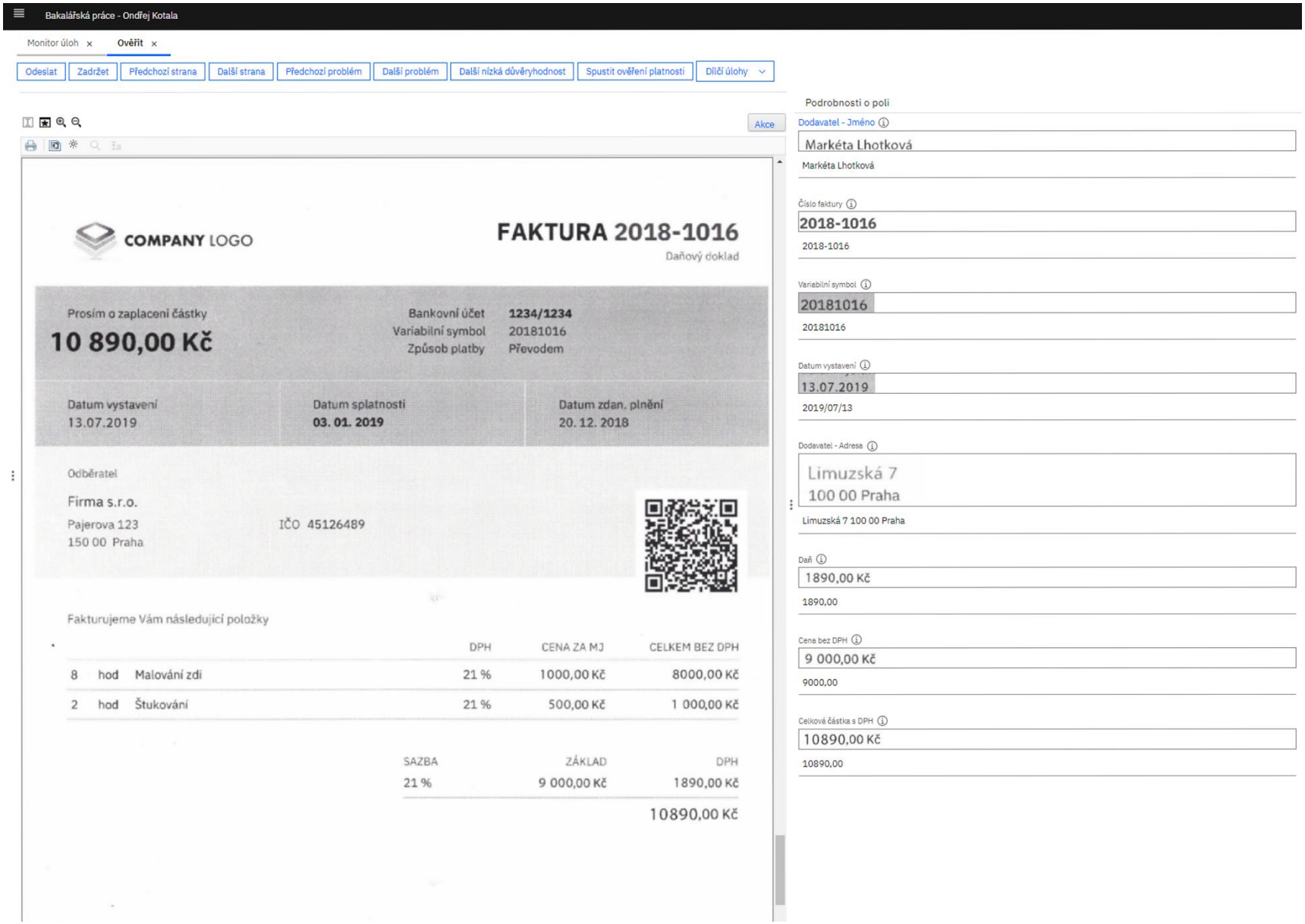
```
    <Field Name="Adress">HOPSINKOVA 28 100 00 PRAHA</Field>
    <Field Name="VarSym">20141014</Field>
   </Page>
  </Document>
  <Document Type="Invoice" Id="20230315.000000.05">
   <Page Type="Main_Page">
    <Field Name="Vendor">Bořivoj Hejsek</Field>
    <Field Name="Invoice_Number">2018-1014</Field>
    <Field Name="Invoice_Date">2018/12/20</Field>
    <Field Name="Tax">1386,00</Field>
    <Field Name="Invoice_Total">7986,00</Field>
    <Field Name="Price">6600,00</Field>
    <Field Name="Adress">Hopsinkova 28 100 00 Praha</Field>
    <Field Name="VarSym">20181014</Field>
   </Page>
  </Document>
  <Document Type="Invoice" Id="20230315.000000.06">
   <Page Type="Main_Page">
    <Field Name="Vendor">Martin Pajer</Field>
    <Field Name="Invoice_Number">183-20118</Field>
    <Field Name="Invoice_Date">2018/12/20</Field>
    <Field Name="Tax">1386,00</Field>
    <Field Name="Invoice_Total">7986,00</Field>
    <Field Name="Price">6600,00</Field>
    <Field Name="Adress">Nákladní 103 101 00 Praha</Field>
    <Field Name="VarSym">18320118</Field>
   </Page>
  </Document>
  <Document Type="Invoice" Id="20230315.000000.07">
   <Page Type="Main_Page">
    <Field Name="Vendor">JAN NOVÁK</Field>
    <Field Name="Invoice_Number">8675-3421</Field>
    <Field Name="Invoice_Date">2018/11/11</Field>
    <Field Name="Tax">2669,00</Field>
    <Field Name="Invoice_Total">15379,00</Field>
    <Field Name="Price">12710,00</Field>
    <Field Name="Adress">JANDOVA 2880 708 00 OSTRAVA</Field>
    <Field Name="VarSym">86753421</Field>
   </Page>
  </Document>
  <Document Type="Invoice" Id="20230315.000000.08">
   <Page Type="Main_Page">
    <Field Name="Vendor">Markéta Lhotková</Field>
    <Field Name="Invoice_Number">2018-1016</Field>
    <Field Name="Invoice_Date">2019/07/13</Field>
    <Field Name="Tax">1890,00</Field>
    <Field Name="Invoice_Total">10890,00</Field>
    <Field Name="Price">9000,00</Field>
    <Field Name="Adress">Limuzská 7 100 00 Praha</Field>
    <Field Name="VarSym">20181016</Field>
   </Page>
  </Document>
 </Batch>
```

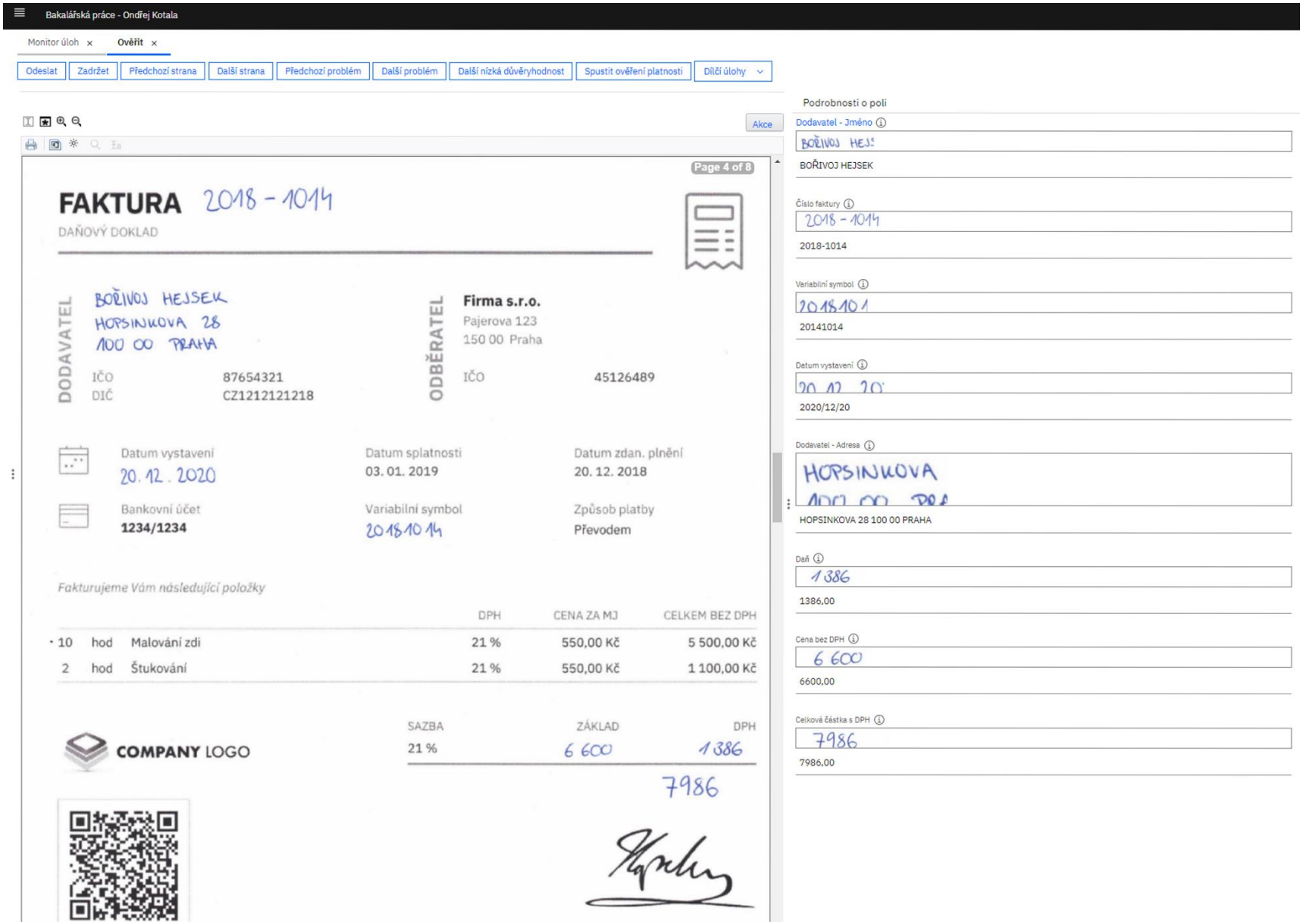
Příloha B Výsledky druhého testu vytěžování ve verifikačním panelu

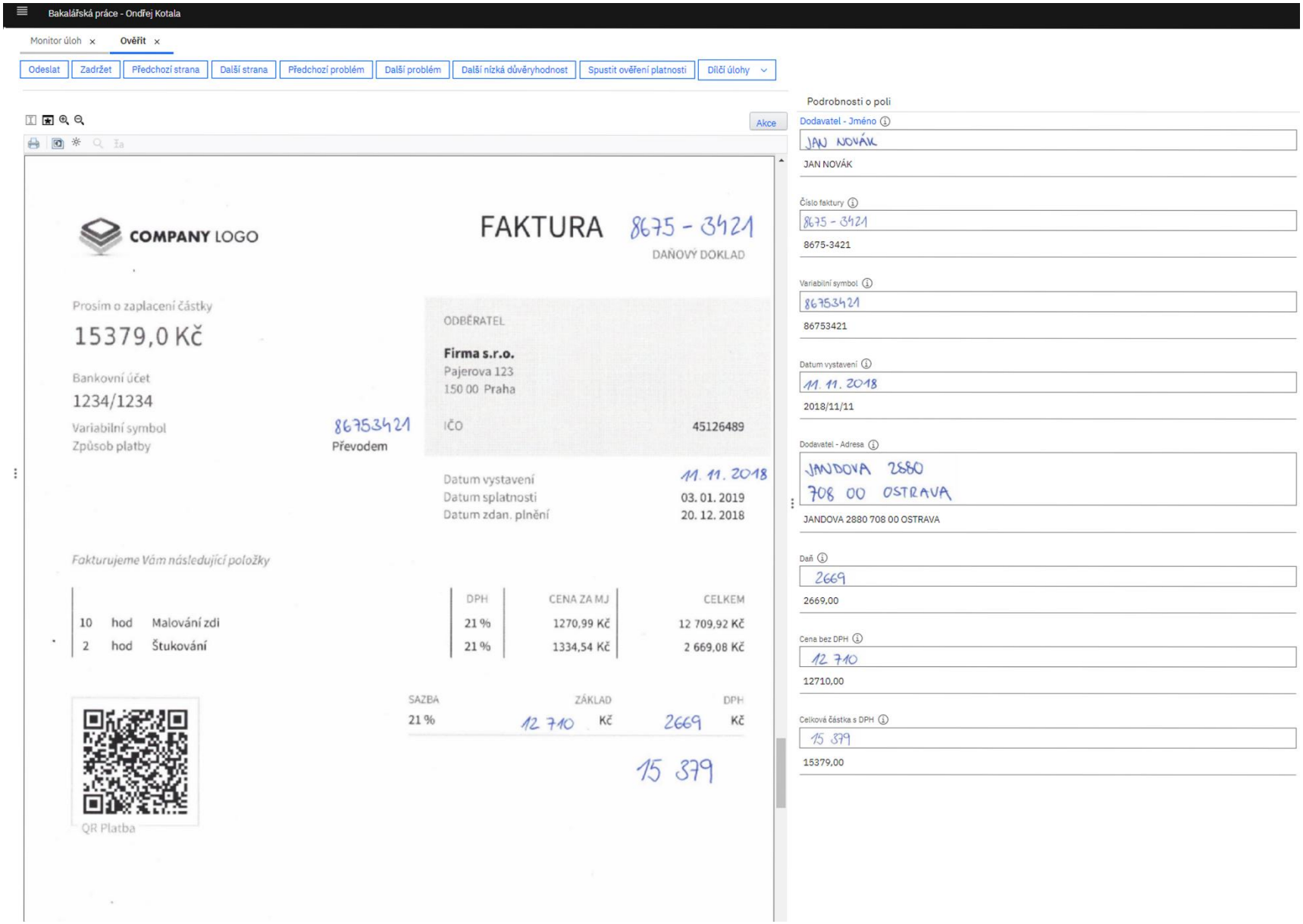



Bakalářská práce - Ondřej Kotala
Monitor úloh
Ověřit
Odeslat
Zadržet
Předchozí strana
Další strana
Předchozí problém
Další problém
Další nízká důvěryhodnost
Spustit ověření platnosti
Dílčí úlohy
Akce
Page 5 of 8
FAKTURA 2018-1014
DAŇOVÝ DOKLAD
DODAVATEL
Bořivoj Hejsek
Hopsinkova 28
100 00 Praha
IČO 87654321
DIČ CZ1212121218
ODBĚRATEL
Firma s.r.o.
Pajerova 123
150 00 Praha
IČO 45126489
Datum vystavení 20. 12. 2018
Datum splatnosti 03. 01. 2019
Datum zdan. plnění 20. 12. 2018
Bankovní účet 1234/1234
Variabilní symbol 20181014
Způsob platby Převodem
Fakturujeme Vám následující položky
DPH CENA ZA MJ CELKEM BEZ DPH
10 hod Malování zdi 21 % 550,00 Kč 5 500,00 Kč
2 hod Štukování 21 % 550,00 Kč 1 100,00 Kč
COMPANY LOGO
SAZBA ZÁKLAD DPH
21 % 6 600,00 Kč 1 386,00 Kč
7 986,00 Kč
Podrobnosti o poli
Dodavatel - Jméno
Bořivoj Hejsek
Bořivoj Hejsek
Číslo faktury
2018-1014
2018-1014
Variabilní symbol
20181014
20181014
Datum vystavení
20. 12. 2018
2018/12/20
Dodavatel - Adresa
Hopsinkova 28
100 00 Praha
Hopsinkova 28 100 00 Praha
Daň
1 386,00 Kč
1386,00
Cena bez DPH
6 600,00 Kč
6600,00
Celková částka s DPH
7 986,00 Kč
7986,00

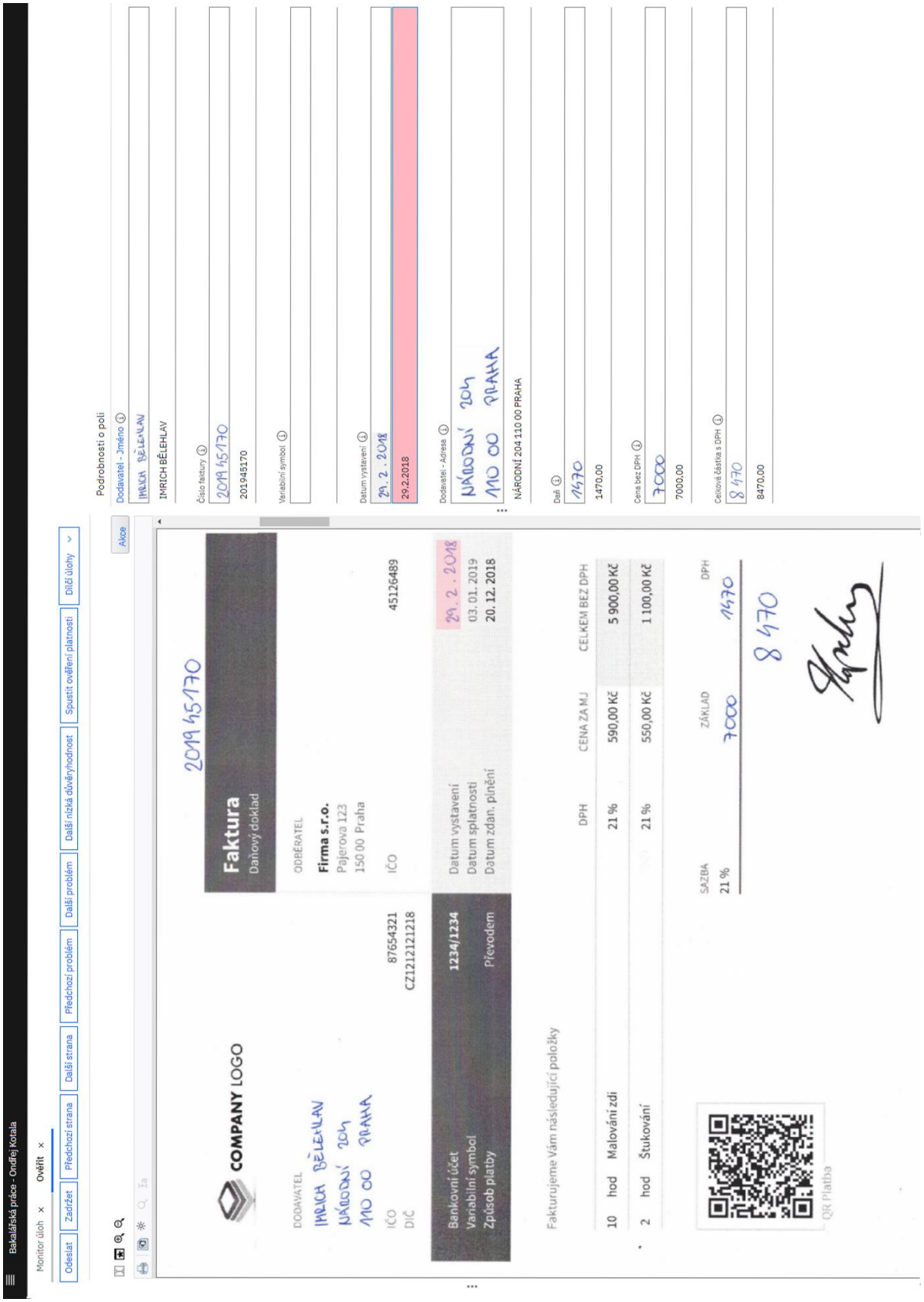

Bakalářská práce - Ondřej Kotala
Monitor úloh
Ověřit
Odeslat
Zadržet
Předchozí strana
Další strana
Předchozí problém
Další problém
Další nízká důvěryhodnost
Spustit ověření platnosti
Dílčí úlohy
Akce
Page 6 of 8
COMPANY LOGO
Faktura 183-20118
DAŇOVÝ DOKLAD
DODAVATEL
Martin Pajer
Nákladní 103
101 00 Praha
IČO 87654321
DIČ CZ1212121218
ODBĚRATEL
Firma s.r.o.
Pajerova 123
150 00 Praha
IČO 45126489
Bankovní účet 1234/1234
Variabilní symbol 18320118
Způsob platby Převodem
Datum vystavení 20. 12. 2018
Datum splatnosti 03. 01. 2019
Datum zdan. plnění 20. 12. 2018
Fakturujeme Vám následující položky
DPH CENA ZA MJ CELKEM BEZ DPH
10 hod Malování zdi 21 % 550,00 Kč 5 500,00 Kč
2 hod Štukování 21 % 550,00 Kč 1 100,00 Kč
SAZBA ZÁKLAD DPH
21 % 6 600,00 Kč 1 386,00 Kč
7 986,00 Kč
QR Platba
Podrobnosti o poli
Dodavatel - Jméno
Martin Pajer
Martin Pajer
Číslo faktury
183-20118
183-20118
Variabilní symbol
18320118
18320118
Datum vystavení
20. 12. 2018
2018/12/20
Dodavatel - Adresa
Nákladní 103
101 00 Praha
Nákladní 103 101 00 Praha
Daň
1 386,00 Kč
1386,00
Cena bez DPH
6 600,00 Kč
6600,00
Celková částka s DPH
7 986,00 Kč
7986,00

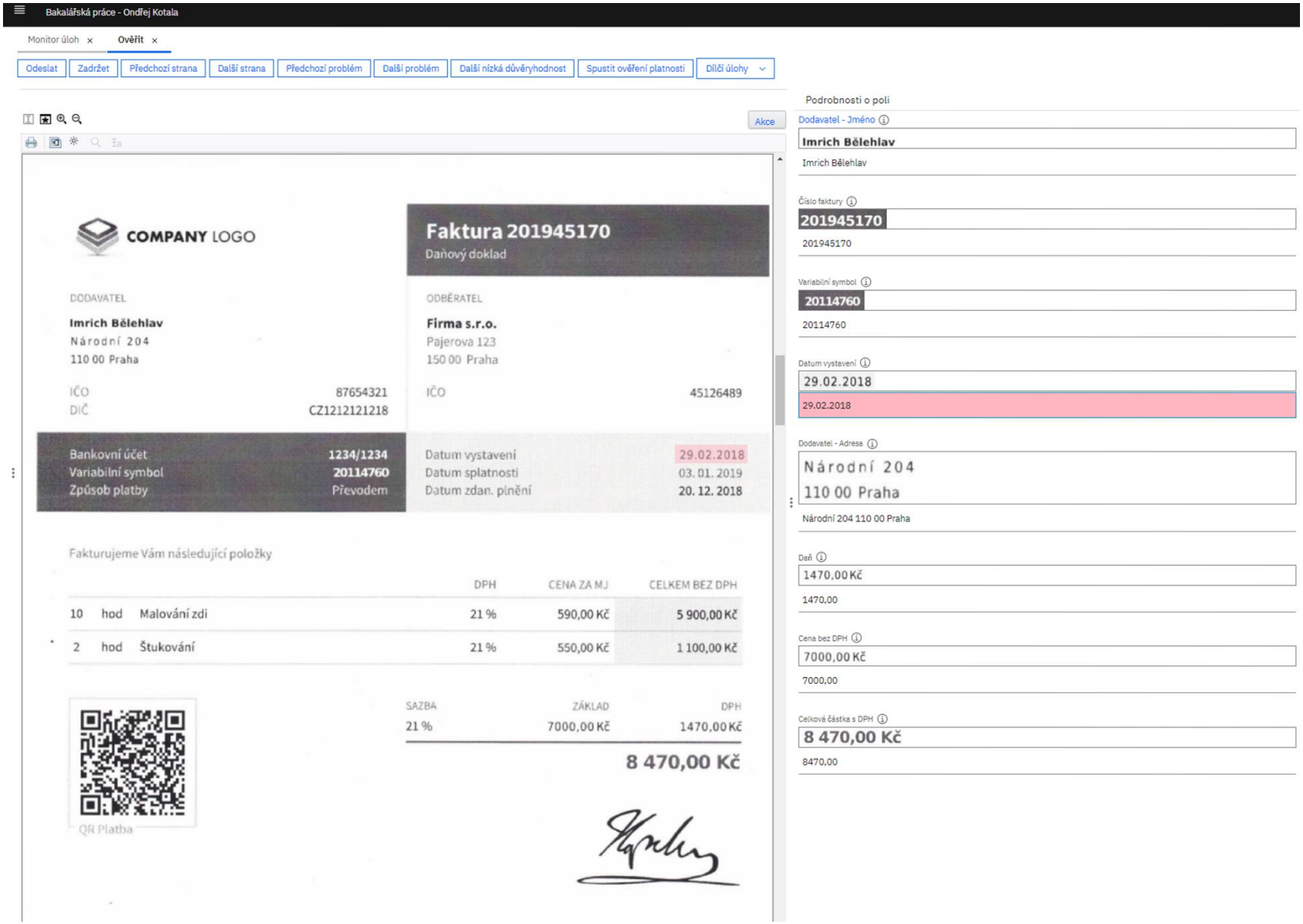

Bakalářská práce - Ondřej Kotala
Monitor úloh
Ověřit
Odeslat
Zadržet
Předchozí strana
Další strana
Předchozí problém
Další problém
Další nízká důvěryhodnost
Spustit ověření platnosti
Dílčí úlohy
Akce
COMPANY LOGO
FAKTURA 2018-1016
Daňový doklad
Prosím o zaplacení částky
10 890,00 Kč
Bankovní účet 1234/1234
Variabilní symbol 20181016
Způsob platby Převodem
Datum vystavení
13.07.2019
Datum splatnosti
03. 01. 2019
Datum zdan. plnění
20. 12. 2018
Odběratel
Firma s.r.o.
Pajerova 123
150 00 Praha
IČO 45126489
Fakturujeme Vám následující položky
DPH
CENA ZA MJ
CELKEM BEZ DPH
8 hod Malování zdi 21 % 1000,00 Kč 8000,00 Kč
2 hod Štukování 21 % 500,00 Kč 1 000,00 Kč
SAZBA
ZÁKLAD
DPH
21 % 9 000,00 Kč 1890,00 Kč
10890,00 Kč
Podrobnosti o poli
Dodavatel - Jméno
Markéta Lhotková
Markéta Lhotková
Číslo faktury
2018-1016
2018-1016
Variabilní symbol
20181016
20181016
Datum vystavení
13.07.2019
2019/07/13
Dodavatel - Adresa
Limuzská 7
100 00 Praha
Limuzská 7 100 00 Praha
Daň
1890,00 Kč
1890,00
Cena bez DPH
9 000,00 Kč
9000,00
Celková částka s DPH
10890,00 Kč
10890,00


Bakalářská práce - Ondřej Kotala
Monitor úloh
Ověřit
Odeslat
Zadržet
Předchozí strana
Další strana
Předchozí problém
Další problém
Další nízká důvěryhodnost
Spustit ověření platnosti
Dílčí úlohy
Akce
Page 4 of 8
FAKTURA 2018 - 1014
DAŇOVÝ DOKLAD
DODAVATEL
BOŘIVOJ HEJSEK
HOPSINKOVA 28
100 00 PRAHA
IČO 87654321
DIČ CZ1212121218
ODBĚRATEL
Firma s.r.o.
Pajerova 123
150 00 Praha
IČO 45126489
Datum vystavení
20. 12. 2020
Datum splatnosti
03. 01. 2019
Datum zdan. plnění
20. 12. 2018
Bankovní účet
1234/1234
Variabilní symbol
20181014
Způsob platby
Převodem
Fakturujeme Vám následující položky
DPH CENA ZA MJ CELKEM BEZ DPH
10 hod Malování zdi 21 % 550,00 Kč 5 500,00 Kč
2 hod Štukování 21 % 550,00 Kč 1 100,00 Kč
COMPANY LOGO
SAZBA ZÁKLAD DPH
21 % 6 600 1 386
7986
Podrobnosti o poli
Dodavatel - Jméno
BOŘIVOJ HEJS
BOŘIVOJ HEJSEK
Číslo faktury
2018 - 1014
2018-1014
Variabilní symbol
20181014
20141014
Datum vystavení
20. 12. 20
2020/12/20
Dodavatel - Adresa
HOPSINKOVA
HOPSINKOVA 28 100 00 PRAHA
Daň
1 386
1386,00
Cena bez DPH
6 600
6600,00
Celková částka s DPH
7986
7986,00


Bakalářská práce - Ondřej Kotala
Monitor úloh
Ověřit
Odeslat
Zadržet
Předchozí strana
Další strana
Předchozí problém
Další problém
Další nízká důvěryhodnost
Spustit ověření platnosti
Dílčí úlohy
Akce
COMPANY LOGO
FAKTURA 8675 - 3421
DAŇOVÝ DOKLAD
Prosím o zaplacení částky
15379,0 Kč
Bankovní účet
1234/1234
Variabilní symbol 86753421
Způsob platby Převodem
ODBĚRATEL
Firma s.r.o.
Pajerova 123
150 00 Praha
IČO 45126489
Datum vystavení 11. 11. 2018
Datum splatnosti 03. 01. 2019
Datum zdan. plnění 20. 12. 2018
Fakturujeme Vám následující položky
10 hod Malování zdi DPH 21 % CENA ZA MJ 1270,99 Kč CELKEM 12 709,92 Kč
2 hod Štukování DPH 21 % CENA ZA MJ 1334,54 Kč CELKEM 2 669,08 Kč
SAZBA ZÁKLAD DPH
21 % 12 710 Kč 2669 Kč
15 379
QR Platba
Podrobnosti o poli
Dodavatel - Jméno
JAN NOVÁK
JAN NOVÁK
Číslo faktury
8675 - 3421
8675-3421
Variabilní symbol
86753421
86753421
Datum vystavení
11. 11. 2018
2018/11/11
Dodavatel - Adresa
JANDOVA 2880
708 00 OSTRAVA
JANDOVA 2880 708 00 OSTRAVA
Daň
2669
2669,00
Cena bez DPH
12 710
12710,00
Celková částka s DPH
15 379
15379,00

Příloha C Výsledky třetího testu vytěžování ve verifikačním panelu




Bakalářská práce - Ondřej Kotala
Monitor úloh
Ověřit
Odeslat
Zadržet
Předchozí strana
Další strana
Předchozí problém
Další problém
Další nízká důvěryhodnost
Spustit ověření platnosti
Dílčí úlohy
Akce
COMPANY LOGO
Faktura 201945170
Daňový doklad
DODAVATEL
Imrich Bělehlav
Národní 204
110 00 Praha
IČO 87654321
DIČ CZ1212121218
ODBĚRATEL
Firma s.r.o.
Pajerova 123
150 00 Praha
IČO 45126489
Bankovní účet 1234/1234
Variabilní symbol 20114760
Způsob platby Převodem
Datum vystavení 29.02.2018
Datum splatnosti 03. 01. 2019
Datum zdan. plnění 20. 12. 2018
Fakturujeme Vám následující položky
DPH CENA ZA MJ CELKEM BEZ DPH
10 hod Malování zdi 21 % 590,00 Kč 5 900,00 Kč
2 hod Štukování 21 % 550,00 Kč 1 100,00 Kč
SAZBA ZÁKLAD DPH
21 % 7000,00 Kč 1470,00 Kč
8 470,00 Kč
QR Platba
Podrobnosti o poli
Dodavatel - Jméno
Imrich Bělehlav
Imrich Bělehlav
Číslo faktury
201945170
201945170
Variabilní symbol
20114760
20114760
Datum vystavení
29.02.2018
29.02.2018
Dodavatel - Adresa
Národní 204
110 00 Praha
Národní 204 110 00 Praha
Daň
1470,00 Kč
1470,00
Cena bez DPH
7000,00 Kč
7000,00
Celková částka s DPH
8 470,00 Kč
8470,00