# Návrh softwarového projektu

### Názov projektu

Lana (Layout ANAlyzer of scanned documents)

### Vedúci projektu

Doc. RNDr. Tomáš Skopal, PhD.
Kontakt: skopal@ksi.mff.cuni.cz

### Cieľová platforma

Predpokladaná klientská platforma bude PC s inštalovaným pluginom Flash Player v internetovom prehliadači. Serverová časť bude pracovať v Linuxe

### Počet riešiteľov

4-členný tím:

- Gálfy Stanislav
- Michalko Jakub
- Vévoda Petr
- Woska Aleš

### Termín odovzdania

Do 9 mesiacov od vypísania projektu

## Popis

Projekt je zadávaný a realizovaný v spolupráci s firmou grappes s.r.o., ktorá sa zaoberá vývojom internetových aplikácií. Cieľom projektu je vyvinúť software, ktorý dokáže z oskenovaného dokumentu vyextrahovať text, určiť typ dokumentu (faktúra, životopis a pod.) a vo formáte XML dát vrátiť užívateľovi vyextrahované dáta.

## Najdôležitejšie vlastnosti

### Rozpoznávanie dokumentu

Predpokladanými vstupnými dátami budú oskenované dokumenty. Rozpoznávanie

dokumentov (obrázkov) bude prebiehať na serveri. Pre vyextrahovanie znakov z obrázku bude použitý Tesseract-ocr verzie 3 s knižnicou Leptonica (súčasť Tesseract-ocr), ktorá zoskupuje rozoznané znaky do blokov. Aplikácia sa najprv podľa layoutu dokumentu snaží nájsť čo najpodobnejší dokument (spôsobom popísaným v práci “Distance Measures for Layout-Based Document Image Retrieval” **[1]**) a podľa neho priradiť význam jednotlivým blokom textu.

Každému bloku textu budú priradené deskriptory (ako napríklad obsahuje tel. číslo, adresu a pod), ktoré budú použité pre upresnenie významu textu a ako pomocné informácie pri vyhľadávaní dokumentu s podobným layout-om (napríklad ak budú nájdené dva dokumenty s podobným layoutom). Nájdený podobný dokumentu bude použitý pre určenie typu dokumentu a pre priradenie významu jednotlivým blokom textu. Výstupný XML dokument bude validný voči vopred známej XSD definícii, ktorá je závislá na jeho type (faktúra má iné XSD než životopis).

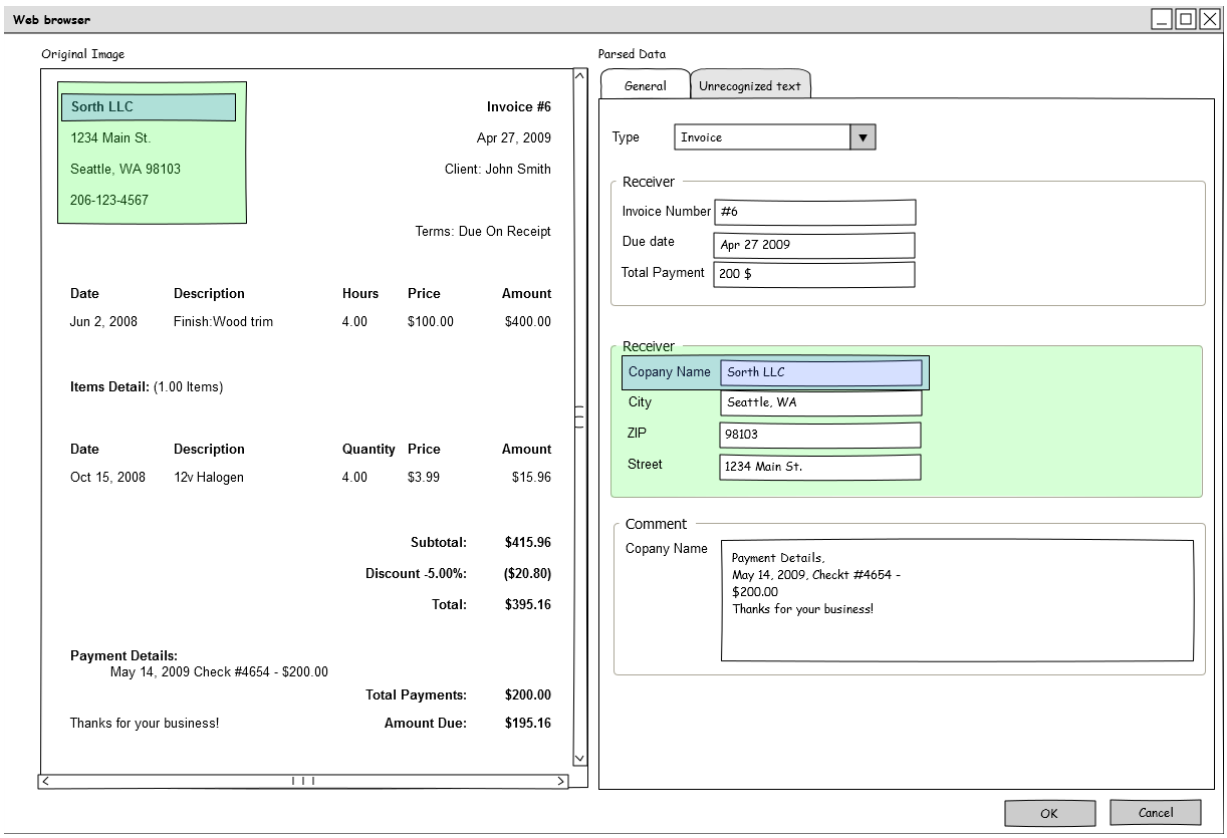
Hlavným spôsobom, ako sa aplikácia učí rozoznávať dokumenty je zo vstupu od užívateľa. Užívateľ v rozoznaných dokumentoch môže robiť korekciu slov (kde OCR zle rozoznalo slovo), alebo priradiť správny význam bloku textu (viď “Užívateľské rozhranie”). V prípade neidentifikovaných dokumentoch, užívateľ priradí každému bloku textu jeho význam. Takto nadefinovaný dokument (každý blok textu má určený svoj význam + korektúra slov) je uložený na serveri a jeho layout bude použitý ako vzor pri ďalšom vyhľadávaní.

## Užívateľské rozhranie

Užívateľ sa do aplikácie bude prihlasovať pomocou internetovej aplikácie vyvinutej v Adobe Flex frameworku. Kvôli bezpečnosti, Adobe Flex nemá prístup k pripojeným zariadeniam, preto bude vyvinutá pomocná aplikácia bežia na lokálnom PC, ktorá umožní preposielať oskenované dokumenty zo skenera priamo do Adobe Flex aplikácie. Táto pomocná aplikácia bude primárne vyvinutá v jazyku C#, pomocou knižnice TWAIN. Ako doplnková funkcionalita, pre podporu viacerých OS, je zvažovaný vývoj tejto pomocnej aplikácie v jazyku Java.

Užívateľ bude mať k dispozícii rozhranie (viď Obrázok 1), v ktorom je zobrazený originálny obrázok (vľavo) a k nemu vyextrahovaný text (vpravo). Toto rozhranie umožňuje užívateľovi dodefinovať typ dokumentu, priradiť význam jednotlivému textu/bloku (napr. v ľavej časti označí blok/bloky textu a pomocou drag&drop im priradí význam), alebo urobiť korektúru v prípade chybného priradenia významu textu/bloku či rozoznania znakov (ak sa zle rozoznaná adresa, užívateľ môže v komponente v pravej časti, opraviť text). Významom je myslené meno, adresa alebo položka na faktúre. Aplikácia sa týmto spôsobom sama učí rozpoznávať typy dokumentov a priraďovať význam k textu/bloku.

Predpokladanou platformou bude PC s pluginom Flash Player v internetovom prehliadači, v prípade PC s pripojeným skenerom by mal byť OS podporovať platformu .net . Ako doplnková funkcionalita celého riešenia je zvažovaný vývoj aplikácie spolupracujúcej so serverom, pre smartphony.




*Obrázok 1: Mockup užívateľského rozhrania*

## Užívateľský účet

Každý užívateľ bude mať k dispozícii svoj účet a úložný priestor na vzdialenom serveri. Tento priestor bude obsahovať záznamy, kde každý záznam je tvorený dvojicou originálny obrázok a k nemu vyextrahované dáta. Užívateľ bude môcť v týchto záznamoch vyhľadávať, alebo robiť jednoduché reporty.

## Prepojenie s externými aplikáciami

Aplikácia bude poskytovať API pre externé aplikácie (cez WSDL), kde vyextrahované dáta budú k dispozícii v štandardizovanom XML formáte, závislom na type dokumentu. (napr. dáta faktúry budú štrukturované podľa predpisu ISDOC).

Súčasťou riešenia je prepojenie s už existujúcim CRM systémom Atollon Lagoon (www.atollon.com). Prepojenie umožní užívateľovi v CRM systéme vytvoriť objekt faktúry, alebo uchádzača priamo zo skenera, či obrázkového súbora uloženom v CRM systéme.

## Architektúra a technológie

Klient bude vyvinutý vo frameworku Adobe Flex (SDK 3.3+) ako RIA aplikácia. So

serverom bude komunikovať prostredníctvom WSDL (SOAP protokolu). Jedinou výnimkou, kôli rýchlosti, bude prenášanie súborov, ktoré nebude riešené pomocou WSDL.

Server bude vyvinutý v jazyku Java, ako Java servlet, ktorý bude používať PostgreSQL databázu a bude pracovať s Tessaract-ocr a jeho výstupmi.

# Vývoj

### Postup práce a rozdelenie

Celý vývoj aplikácie je určený pre 4-členný tím. Následujúce body popisujú približný popis funkcionality a zodpovedných riešiteľov:

- Vývoj serverovej časti pre správu užívateľských účtov (vrátane komunikácie s databázov), pripojenia, report nad dokumentami (hlavný riešiteľ Aleš Woska)
- Vývoj serverovej aplikácie pre rozoznanie dokumentu podľa šablóny, vyextrahovanie dôležitých dát a export do XML dokumentu (hlavný riešiteľ Petr Vévoda a Jakub Michalko)
- Vývoj klientskej aplikácie a prepojenie s CRM systémom Atollon Lagoon (hlavný riešiteľ Stanislav Gálfy)

# Obhajoba

Obhajoba bude zahŕňať demonštráciu bežiacej inštalácie a demonštráciu importu oskenovaného dokumentu priamo do prepojenej externej aplikácie.

# Odkazy

**[1] Joost van Beusekom, Daniel Keysers, Faisal Shafait, Thomas M. Breuel:** *Distance Measures for Layout-Based Document Image Retrieval,* **http://www.keysers.net/daniel/files/Beusekom--Layout-Distance--DIAL2006.pdf**