EvaluatieRapport

Transfertool

Naam: Shibin Pan

Klas: Seito21A

Stagebedrijf: Prodist ERP Software

Stagebegeleider: William Cornelissen

Datum: 16/02/2024

Versie: 0.1

# Inleiding

De examenopdracht wordt het bouwen van een tool die van PDF naar ProdistERP XML-bestand omzet. De bedoeling is dat deze tool een/meerdere PDF-bestand kan analyseren en de inhoud zo goed mogelijk zien te vertalen naar een XML-bestand, zodat ERP-applicatie uiteindelijke kan inlezen en importeren.

De tool hoeft niet voor alle willekeurige PDF-bestanden te gelden. De examenopdracht is beperkt tot de PDF-bestanden van een hotelketen. Dus de opmaak van de PDF-bestanden ziet er hetzelfde uit.

Dit project heeft tot doel om de samenwerking van ProdistERP Software en zijn klanten te verbeteren en te versterken.

In dit evaluatierapport wordt ingegaan op hoe het proces tijdens de examenopdracht is verlopen, het product en over mezelf.

Inhoudsopgave

# Het proces

De gemaakt afspraak voor eventueel documentaties zijn Functioneel ontwerp, Research analyse en Technisch ontwerp.

In het Functioneel ontwerp staat de huidige- en gewenst situatie, eventueel het stroomdiagram om te laten zien wat de stappen zijn om het project succesvol af te ronden. Omdat ik eerst van plan was om een Prototype te maken, wordt daar dus twee plannen(stappenplan) gemaakt.

In de Research analyse wordt een onderzoek gedaan om beter inzicht te krijgen voor de beste aanpak. Per onderdeel wordt de huidige oplossingen, bestaande technologieën, schaalbaarheid etc, uitgelegd.

In het Technisch ontwerp staat welke programmeertaal er wordt gebruikt, en welke tools er eventueel gebruikt kunnen worden.

Een Github omgeving gemaakt, ik heb daarin al mijn documentaties en mijn sourcecode, om te voorkomen dat alles verloren gaat. Er wordt zoveel mogelijk gepusht na elke kleine ding die gedaan is. Ook kan ik dan naast mijn laptop ook nog op zakelijke computer laten bekijken.

Trello wordt aangemaakt die elke dag bijgehouden wordt, zodat ik overzicht heb wat ik aan het doen ben, en wat eraan komt.

Verder wordt het Logboek elke dag eind van de dag ingevuld, per onderdeel wordt er uitgelegd om beter in details te geven.

Omdat het stagebedrijf niet met methode SCRUM werkt waar wij op school gewend zijn, dus er wordt geen Daily of Sprint review gedaan, maar ik probeer wel elke dag een update te geven aan William voor eventueel het ontwikkelproces, om even wat te laten zien wat ik de afgelopen dagen heb gedaan, en wat ik de aankomende dagen ga doen.

Op eerste week heb ik een opwarmopdracht gekregen om een tool te maken, dat je jouw bestanden kan uploaden naar een specifieke map. Die heeft me echt op het idee gebracht om eerst een prototype te maken. In plaats van uploaden, kan ik de bestanden gelijk uitlezen, en dan plaatsen naar die map.

Omdat de tweede week eigenlijk de voorjaarsvakantie is, heeft mijn Examenopdrachtomschrijving voor de hele week in de mail van meneer de Boer gelegen. In de derde week heb ik die nog wat aanpassingen gedaan en opnieuw gestuurd.

Het prototype kon ik eigenlijk eind van derde week al laten testen (omdat ik de tweede week met documentaties bezig ben), de reden dat het zo snel gaat is, dat ik van tevoren een onderzoek heb gedaan, hoe ik het beste kan aanpakken, en wat voor tools ik gebruik kan maken. Dit onderzoek heeft me heel erg goed geholpen met eventueel bestaande bibliotheken voor het verwerken van PDF bestanden, verder heb ik ook nog even opgezocht hoe ik bijvoorbeeld “FileSystemWatcher” kan gebruiken.

In vierde week ben ik voornamelijk bugs aan het fixen van het prototype. Ik heb begin van de week meer PDF’s gevraagd om te testen. Einde van de dag 141 bestanden gekregen, en blijkt 11 bestanden niet doorheen kunnen, 9 van de 11 hebben te maken dat sommige cellen in de artikelentabel newline(\n) bevat, en dat neemt hij letterlijk ook mee, dat ik van tevoren geen rekening mee heb gehouden. 2 van de 11 hebben te maken dat ik een check heb gedaan, zodra hij detecteert dat de hotelnaam uit ‘NH’(hotelketen) bevat, dan gaat hij pas alles uitlezen en omzetten naar XML. Maar die 2 bestanden hebben juist geen ‘NH’ in hun hotelnaam, maar wel behoren tot die hotelketen. Dus heb verder nog even onderzoek gedaan of de tekst uit logo kan halen, want het logo blijft wel gewoon hetzelfde en op dezelfde positie. De tekstbugs heeft me 3-4 dagen gekost om ze te fixen(moet nog onderzoek doen over het uithalen van tekst in image).

Eind van de week heb ik pas de gepublished versie ook kunnen testen. Daarom is het contact met William stuk minder. Eerst werkt alles ook niet naar wens. De handmatige upload van bestanden werkt wel. Maar zodra ik een bestand in een map sleept, crasht het en hij geeft ook geen foutmelding(niet via sourcecode). Uiteindelijk heb ik online gezocht dat ik van Debug naar Release moet schakelen en hij geeft een error waarvan het niet mijn code is(ArgumentNull…). Uiteindelijk heb ik gevonden dat het te maken heeft met mijn configuratiebestand dat niet goed ingesteld is, dus heb die maar weggehaald. En het werkt. Gepublished versie ook getest, en ook op mijn zakelijke computer(Hij heeft wel toestemming nodig om de bestanden te schrijven en te verplaatsen).

Alleen blijft de logoerror ernaast, omdat die 2 bestanden geen ‘NH’ bevat in hun hotelnaam, dus het programma slaat hun over, terwijl ze wel behoren tot die hotelketen. Dus heb daar ook nog even onderzoek gedaan of het mogelijk is om tekst uit een image te halen.

Verder heb ik nog wat extra gekregen:

1. dat er een configuratie bestand. Hij leest het logo uit(als dat mogelijk is), zodra het een “NH”-hotelketen is, pakt hij die structuur. Andere hotel? Andere structuur. 2. Windows service, is iets voor later. Bij het opstarten van je computer kan je meteen het programma gebruik van maken, dus hoef je niet het programma eerst nog te openen.

Vijfde week is eigenlijk het prototype klaar. Het is ook getest door William. Alleen was wel de bedoeling dat deze week nog de XML om te zetten naar een ProdistERP-XML, maar was niet eraan toegekomen. Verder hebben we de code doorgelopen met Marco(technische begeleider). Daar hebben we een begin gemaakt voor het definitiebestand. Daar ben ik dus de hele week bezig geweest. Het uithalen van vaste gegevens vanuit het configuratie gaan redelijk goed, omdat ze met coördinaten worden uitgelezen. Verder heb je een bestelling met 1 artikel en soms weer 20 artikelen, dus daar wordt niet zomaar de coördinaten toegepast, maar door tekst. Hetzelfde geldt voor de levering , totaalbedrag en opmerking etc… Als je meer dan 1 artikel hebt in de factuur, dan wordt de levering, totaalbedrag, opmerking naar tweede pagina geschoven.

Alleen ik loop nog wel wat probleempjes/uitdagingen tegen:

1. Voor Pdf’s die uit meerdere pagina’s bestaan, gaan de loop ook meerdere keren, waardoor de eerste xml tag wel is gevuld met data, en bij de volgende loop staat de xml tag weer leeg.

2. Om per artikel in een bestelling te splitsen is nog een dingetje, want de teksten worden uitgehaald als platte teksten, niet als een tabel. Per bedrijf heeft ook z’n eigen opzet van de factuur.

Zesde week is vooral het definitiebestand afmaken, begin van de week de dubbele tags gefixt met een nieuwe list, zodra de tag nog niet bestaat, voeg het toe, anders sla hem over.

En de artikeltags worden op dit moment nog met spatie verdeeld(Er zijn spaties tussen de tekst), het is niet helemaal gebruiksvriendelijk, maar ik weet ook niet hoe ik het anders kan doen. Maar hij werkt wel naar behoren.

Alle configuraties worden van tevoren gedefinieerd. Wat de tool doet, hij gaat eerst door alle json bestanden in huidige map heen, haalt de tags en identifiertext uit(checkt de identieke tekst/teksten die per pdf bevat). Zodra hij hem heeft gevonden, dan doet hij het volgende, anders heeft hij dus niet gevonden, en wordt het bestand naar “Afgewezen” verplaatst. Als hij de identifiertext heeft gevonden, gaat mijn tool beginnen met werken, alle tags wordt door mijn tool uitgehaald en vervolgens per onderdeel wordt de tekst vanuit pdf uitgehaald en in XML gezet. Zodra alles gelukt is, wordt dit bestand naar “Verwerkt” verplaatst.

PDF-bestanden die wel door mijn tool is gegaan(dus dat ze wel omgezet zijn naar XML), worden in submap “Verwerkt”, alle andere bestanden(die geen configbestand heeft met zijn identieke tekst/ze zijn überhaupt geen pdf bestanden) worden naar “Afgewezen” verplaatst. Maar ik kom achter op het moment als de pdf al in submap bestaat, en je kopieert hetzelfde bestand, dan crasht mijn tool en sluit die af. Maar dat heb ik gefixt door een check te doen, als ze niet in submap bestaat, dan worden ze verplaatst, anders verwijderen.

Aan het einde van de week heb ik een andere pdf-soort van ander bedrijf gekregen, dus heb ik van die een configuratiebestand gemaakt, grotendeels heeft gewerkt, alle vaste gegevens worden goed uitgelezen en goed in xml omgezet. Maar de artikelen niet helemaal, omdat hij niet dezelfde aantal tags heeft, dus de spatie-methode werkt niet meer. Dat heb ik probeer te fixen met andere pdf-library Spire. Voor NH wordt de artikelen uitgehaald met “|” ertussen, dus dat is niet meer moeilijk om ze te verdelen in tags. Maar bij andere pdf wordt er niks uitgehaald, omdat hij geen table-structuur heeft. Dus dit zou ook niet werken. Maar heb ik verder de hele tekst probeer uit te halen om te kijken of ze wel overeenkomt met structuur. Maar dat lukt ook niet helemaal. Dus misschien moet ik toch ocr gebruiken om de artikelen uit te halen. Eerder heb ik ook OCR gebruikt om de teksten uit te halen om te kijken hoe ze eruit zou komen te zien, maar dat was ook geen succes. Ten eerste duurt het veel langer om de tekst uit te halen. Ten tweede worden sommige teksten niet eens goed herkend waardoor er uiteindelijk rare teksten uitkomen->misschien andere OCR-software?

In de 7de week is de bedoeling dat de artikelen wel op goede manier worden uitgehaald. Begin van de week heb ik in totaal 4 libraries geprobeerd om de artikelen wel in een tabel uit te halen. Maar dat lukt gewoon niet, omdat pdf geen structuur heeft, het bevat alleen maar teksten etc. Maar ik heb van Marco een andere framework Syncfusion doorgekregen, het werkt beter dan verwacht maar niet helemaal perfect. Voor NH pdf -soort werkt het perfect, alle cellen worden in een tabelvorm gezet. Ik kan makkelijk de xmltags(vanuit configuratiebestand) koppelen met de values. Maar voor dat andere bedrijf, de meeste values worden als een tabelcel uitgehaald, maar er is een cel die wordt gesplit, terwijl dat in de pdf niet zo is. En ik snap ook niet waarom.

Syncfusion zit het dichtstbij naar wat ik wil. Maar omdat hij een cel niet goed definieert. Werkt mijn tool niet meer. Daarom Heb ik hulp omgevraagd. Dan is de bedoeling dat hij de bovenkant van de tabel herkent, en vervolgens de coördinaties ophaalt, met die coördinaties gaat hij steeds kijken of er per cel iets bestaat.

Maar ik kwam probleem tegen, als ik de tags(om te koppelen) vanuit de hele tekst(currenttext) ophaalt, en dan foreach tag te checken of hij in pdf bestaat, dan houdt hij geen rekening met newline in de cel, daarom komt er geen resultaat bij coordinaties. Of anders komt er meerdere resultaten uit omdat een woord kan meerdere keren in een pdf staan.

Maar als ik de tags vanuit configuratie ophaal. (We hebben gedaan dat de configuratie-tags gekoppeld is met de xmltags, daarom mogen in de configuratie-tags geen spaties zitten). Dan ziet de xmltags sws anders uit dan de bovenkant van de tabel. Dus dat is ook niet mogelijk.

Het contact met William is heel goed, elke dag kan ik wel een update geven om even wat te laat zien. Terwijl het contact met de product owner tijdens de proeve proef toen een aandachtspunt is.

In het prototype krijgt de gebruiker een scherm waarmee hij een pad kan selecteren, en de bestanden kan uploaden. De pad wordt dan opgeslagen voor later gebruik. En vervolgens wordt die ook in de gaten gehouden, zodra daar een nieuwe bestanden inkomen, wordt mijn tool geactiveerd, hij kijkt of voor elk bestand of hij überhaupt een PDF bestand is, dan wordt op een specifieke coördinaat de tekst uitgehaald, zodra hij dat niet herkend, dan geeft hij ook een foutmelding. Als alles doorheen is gegaan, worden alle gegevens verzameld, onder andere de afleverdatum, afleveradres, factuuradres en de bestelde artikelen, vervolgens naar een opgebouwd XML bestand omgezet. Het is wel de bedoeling dat dit XML bestand uiteindelijk ingelezen kan worden in ProdistERP.

Omdat William niet op donderdag en vrijdag op kantoor is, kan ik moeilijk hem wat te laten zien op die twee dagen. Maar ik geef wel meteen de maandag daarna een update wat ik de afgelopen week heb gedaan, en tegelijktijdig even wat verbeteringen/aanpassingen vragen.

Ik vind deze stage veel leerzamer dan mijn vorige stage. In mijn vorige stageperiode moest ik met het stagebedrijf meewerken aan een project. Alles hebben ze eigenlijk al klaargezet, en ze waren ook al een tijdje mee bezig. Dus ik heb niet de tijd om alles te laten verzinken aan het einde van de dag. Soms weet ik ook niet wat ik aan het doen was.

Maar met deze stage krijg ik een opdracht waarmee ik zelf aan de slag kan. Ik kan mijn eigen tempel pakken, waardoor ik meer verantwoordelijker voel over mijn eigen project. Er wordt geen Daily of Sprint Review gedaan, omdat dit stagebedrijf niet met SCRUM-methode werkt. Maar aan het begin of einde van de dag neem ik altijd even de tijd om logboek bij te werken en mijn trelloboard. Elke week wordt in dit document uitgelegd wat ik de afgelopen week heb gedaan, zodat ik later nog terug kan kijken. Eventueel voor het definitief Evaluatierapport, en ook later voor de presentatie.

Omdat de opdracht duidelijk is beschreven via de mail, is de examenopdrachtomschrijving aan het einde van dezelfde dag al klaar. De tweede week ben ik dan gefocust op de afgesproken documentaties. In derde week begonnen met het prototype. Het is eerste wel even struggelen hoe alles in elkaar zit. Maar het ontwikkelen is veel sneller verlopen dan ik heb verwacht omdat ik die opwarmopdracht kan gebruiken. Aan het einde van derde week kon ik mijn prototype al laten zien, met het uploaden van bestanden om vervolgens te laten omzetten naar XML, met het beluisteren naar directory. Dat heeft William ook niet verwacht.

Vanaf 5de week neem ik wat rustiger aan, omdat alles volgens planning loopt, en ik merk steeds lastiger om de hele dag door te halen. En verder het weekend even rust te pakken heeft ook enorm geholpen.

Vanaf 6de week ben ik mijn prototype aan het omvormen naar een meer kant en klaar product. Er wordt vanuit een configuratiebestand de xml tags aangemaakt. Dus mijn tool is ook later gebruikelijk voor eventueel andere bedrijven.