SCRATCH+
Naar een product voor handschriftherkenning in erfgoedcollecties

door: Ivo Zandhuis (ivo@zandhuis.nl)
in opdracht van: Nationaal Archief (Henny van Schie)

versie: na_scratch4all_20091221

Inhoudsopgave

1. Inleiding

1.1. Achtergrond

In kader van het NWO-programma CATCH hebben het Nationaal Archief en de Rijksuniversiteit Groningen samengewerkt bij de ontwikkeling van technologie voor het geautomatiseerd herkennen van historische handschriften. Dit is het SCRATCH project (SCRipt Analysis Tools for the Cultural Heritage) genoemd en liep van 2005 tot 2009.

Het SCRATCH-project heeft zich gericht op een serie registers dat onderdeel uitmaakt van het archief van de Kabinet des Konings/Kabinet der Koningin (KdK) dat wordt bewaard bij het Nationaal Archief. Vanwege de typische ordening van dit archief is de handschriftherkenning hier voor de toegankelijkheid van groot belang. Bovendien werden deze registers tientallen jaren achter elkaar door dezelfde klerk geschreven, waardoor er een grote hoeveelheid trainingsmateriaal is om de computer te "leren lezen".

1.2. Aanleiding

Na de afsluiting van SCRATCH project is in het kader van CATCH+ subsidie aangevraagd om de uitkomsten van het project verder uit te werken tot een product dat op eenvoudige wijze ook door andere erfgoed-instellingen kan worden gebruikt. Om dit product te kunnen maken is een overzicht noodzakelijk van de onderdelen die zijn ontwikkeld. Dit overzicht was echter niet bij alle betrokkenen voldoende aanwezig.

1.3. Belang

Bij het toegankelijk maken van handgeschreven erfgoed hebben de deelnemers in SCRATCH een voorsprong opgebouwd op andere groepen in de wereld. We willen deze voorsprong voor Nederlands erfgoed behouden.

Voor een erfgoedinstelling is het van belang een aanspreekpunt te hebben dat een product (of een dienst; we gebruiken verder het woord "product") kan aanbieden waarmee een handgeschreven erfgoedcollectie kan worden ontsloten.

1.4. Centrale vraag

- Welke componenten die in SCRATCH zijn ontwikkeld, zouden kunnen worden ingezet bij het ontwikkelen van het beoogde product?
- Hoe kan de ontwikkeling van het beoogde product worden aangepakt?
- Welke (basale) eisen worden er aan het beoogde product gesteld?

1.5. Aanpak

Om overzicht te krijgen in de onderdelen en inzichten die gedurende SCRATCH zijn ontwikkeld, zijn gesprekken gevoerd met Lambert Schomaker (wetenschappelijk projectleider van SCRATCH), Tijn van der Zant (promovendus) en Fons Laan (wetenschappelijk programmeur). Daarnaast zijn enkele artikelen[1] bestudeerd, het digitale archief dat Fons heeft nagelaten en documentatie. Vooral door "rond te dwalen" in de laboratorium-omgeving[2] is veel inzicht verkregen in de huidige stand van zaken.

1 in het bijzonder: T. v.d. Zant e.a., Where are the Search Engines for Handwritten Documents?, in: INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 34 No. 2, June, 2009, 228–239
2 http://aps.ai.rug.nl/KdK/ (toegankelijk met behulp van wachtwoord)

Over de aanpak van de ontwikkeling van een product zijn gesprekken gevoerd met Julia Vytopil en Hennie Brugman van CATCH+ en Gert-Jan van Dijk van Target Holding.

Samen met Henny van Schie is een basale lijst van functionele en niet-functionele eisen geformuleerd.

1.6. Status van dit rapport

Het is binnen de gestelde termijn niet mogelijk gebleken om volledig inzicht te krijgen in de componenten die zijn ontwikkeld. Hierdoor bestaat de kans dat omschrijvingen te kort door de bocht zijn.

De complexiteit is ontstaan omdat componenten zijn ontwikkeld om een specifiek wetenschappelijk experiment uit te voeren en het minder van belang werd gevonden om een samenhangende omgeving te ontwikkelen: het project is niet als software-engineeringsproject ingericht.

Alvorens deze teksten worden gebruikt om de ontwikkeling van het product verder vorm te geven, verdient het daarom aanbeveling ze nogmaals te toetsen, te actualiseren en waar nodig te detailleren.

2. Informatie-architectuur

In de laboratorium-omgeving wordt onderscheid gemaakt tussen verschillende informatie-objecten. Telkens maakt een kleiner informatie-object deel uit van het grotere geheel.

Deze hiërarchie wordt ook gebruikt bij het pad en de bestandsnaamgeving van *images* die in de *storage* zijn opgeslagen. De structuur hiervan is vastgelegd in een XML-bestand dat wordt gebruikt om de processen (zie hoofdstuk 3) te sturen en wordt daarom "stuurfile" genoemd.

2.1. Instelling

De instelling ("institution") is de organisatie die de te lezen images beschikbaar heeft gesteld. In het SCRATCH project was dit het Nationaal Archief.

2.2. Collectie

Een collectie ("collection") is de grootste eenheid waartoe de images behoren, binnen de instelling. Een archivaris zal liever spreken van "archief". In het SCRATCH project was dit het archief van het Kabinet des Konings.

2.3. Boek

Het boek ("book") is de eenheid waarin de images als een geheel binnen de collectie worden geïdentificeerd. In het geval van het Kabinet der Koningin is sprake van een register. De experimenten tijdens SCRATCH hebben zich gericht op het register uit 1903[3].

2.4. Pagina

De pagina ("page") is de eenheid binnen het boek, die als zodanig kan worden geïdentificeerd. Van elke pagina is één image. Bij het KdK 1903 boek werden daarvoor de scans van twee naast elkaar liggende pagina's (een spread) automatisch gescheiden in twee images.

2.5. Paragraaf

De pagina kan worden onderverdeeld in inhoudelijk logische eenheden. In de KdK-registers, is dat de inschrijving van een verbaal met datum en vindplaats.

2.6. Regel

De pagina's worden onderverdeeld in regels ("line").

2.7. Woord

In een regel worden Word-zones onderscheiden. Het staat niet bij voorbaat vast dat dit hetzelfde is als een woord, omdat er witruimtes binnen een woord kunnen zijn

3 Nationaal Archief, Den Haag, Kabinet der Koningin, (1814) 1898-1945 (1988), nummer toegang 2.02.14, inventarisnummer 7823

geschreven, die tot valse scheidingen leiden. Ook kan een woord aan het eind van de regel zijn afgebroken.

Word-zones kunnen waar nodig worden gecombineerd tot woorden.

2.8. Letter

Binnen het woord worden ook letters onderscheiden. Wellicht leidt deze detectie in de toekomst tot de herkenning van woorden die daarvoor niet afzonderlijk hoeven te worden getraind.

3. Procesarchitectuur

De verwerking van images tot een zoekdienst bestaat uit het uitvoeren van de volgende processen:

- capture
- segmentation
- annotation
- classification

Na het doorlopen van deze processen kan aan publiek en archiefmedewerkers de zoekdienst worden aangeboden.

Merk op dat -hoewel we het hier lineair presenteren - deze processen zich niet perse in deze volgorde afspelen. Zo wordt het resultaat van de annotatie gebruikt om de woorden correct te kunnen detecteren en wordt het vergelijken van de uitgeknipte woorden gebruikt om het annotatie-proces te ondersteunen.

3.1. Capture

Onder *capture* wordt verstaan het opnemen van de scans en de benodigde metadata om het systeem te kunnen laten werken.

Scanning

Het proces begint met het scannen van een (deel van een) collectie. Deze scans moeten voldoen aan eisen aan de kleur, de schuinte waarmee is gescand, de schaduw tussen de bladzijdes en dergelijke. Op dit moment zijn dat eisen waaraan niet zonder meer door een scanbedrijf kan worden voldaan.

Inlezen

De scans worden opgeslagen in het systeem. Daarnaast wordt de naamgeving van de images in de collectie en het boek beschreven in de stuurfile. Met behulp van een script worden alle benodigde afgeleide images en paden aangemaakt.

Metadata

Elke pagina moet een eigen unieke omschrijving hebben: een paginanummer, of het eerste woord of iets dergelijks.

3.2. Segmentation

Segmentation is het proces waarbij de image in kleinere images wordt opgebroken. De kleinere images bevat dan paragrafen, regels, word-zones, woorden of letters. De plaats van het gedetecteerde segment wordt vastgelegd met behulp van XY-coördinaten binnen de (grote) image. De XY-coördinaten worden vastgelegd in de bestandsnaam van het uitgeknipte segment.

Om segmentation te kunnen doen, wordt een image eerst met beeldbewerkingssoftware tot een image met specifieke eigenschappen (denk aan zwart-wit) omgezet.

Paragrafen

Marius Bulacu (post-doc) ontwikkelde een lay-out analyse, waarbij onder andere paragrafen worden gedetecteerd[4].

4 Marius Bulacu, Rutger van Koert, Lambert Schomaker, Tijn van der Zant (2007) "Layout analysis of handwritten historical documents for searching the archive of the Cabinet of the Dutch Queen", Proc. of 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007), IEEE Computer Society, pp. 357-361, vol. I, 23 - 26 September, Curitiba, Brazil.

Regels
Regels worden gedetecteerd door het aantal zwarte pixels te tellen op een horizontale lijn. Bij een bepaald aantal pixels wordt aangenomen dat deze horizontale lijn deel uitmaakt van een regel.

Woorden
Voor het detecteren van woorden zijn verschillende technieken uitgeprobeerd. Deze technieken worden gedemonstreerd in de zoekmachine naar Word-zones (paragraaf 3.5).

3.3. Annotation

Annotation is het proces waarin vrijwilligers en archivarissen een deel van het handschrift voorzien van transcriptie.

Annotator van Lambert Schomaker
In de web-based annotatietool van Lambert Schomaker kunnen vrijwilligers invoeren welke tekst ze lezen in de regels die in de gebruikersinterface worden gepresenteerd.

http://aps.ai.rug.nl/cgi-bin/monk?db=1000&ipage=54&cmd=broken_page&pagemode=broken

Op basis van deze input worden afzonderlijke woorden herkent. Deze worden ter controle aan de vrijwilliger gepresenteerd, waardoor het algoritme sneller wordt getraind.

Annotator van Fons Laan: KdK-editor
Deze editor is gemaakt in C++ en is aanwezig in het digitale archief van Fons op het Nationaal Archief. Hij heeft een executable (KdK-editor.exe) gemaakt, waardoor de tool op een Windows-werkstation kan worden gebruikt.

Voorwaarde voor het gebruik is dat de tool de beschikking heeft over de te annoteren images en de layout-analyse, waarin de line-segmentation is vastgelegd (zie par 3.2). De tool maakt via internet gebruik van de opgeknipte images die zijn opgeslagen in Groningen.

De tool maakt gebruik van een Open Source XML-database, waarin de lay-out en het transcriptie-bestand worden opgeslagen. Er wordt geen informatie teruggestuurd naar de laboratorium-omgeving in Groningen.

3.4. Classification

Classification is het proces waarin een image van een woord wordt "geclassificeerd", dat wil zeggen (hypothetisch) gekoppeld aan een woord waarvan de betekenis wel bekend is. Na classification worden de woorden opgeslagen in een database, waardoor een snelle zoekactie mogelijk is. Voor de classification is geëxperimenteerd met verschillende technieken.

Lambert Schomaker e.a. hebben geëxperimenteerd met een techniek, waar het rekenproces langdurig was en werd uitgevoerd om een supercomputer[5].

Tijn van der Zant heeft een algoritme ontwikkeld dat niet in de laboratorium-omgeving is geïmplementeerd, maar op zijn laptop. Hij verwacht begin 2010 over deze techniek een artikel te schrijven. De herkende woorden worden vanaf zijn laptop niet opgenomen om de laboratorium-omgeving. Dit algoritme is sneller dan het eerder gebruikte algoritme.

5 Schomaker, L.R.B., K. Franke, and M. Bulacu. 2007. Using codebooks of fragmented connected component contours in forensic and historic writer identification. Pattern Recognition Letters 28(6): 719–27.

3.5. Keyword search

Zoekmachine van Tijn

Deze zoekmachine zoekt en vindt regels in de tekst (line-strips). De gevonden regel kan worden teruggevoerd aan het systeem om vergelijkbare regels (dat wil zeggen regels die er visueel op lijken) te zoeken.

Deze zoekmachine demonstreert de mogelijkheid van het associëren van vormen op regel niveau.

http://rugtest5.service.rug.nl:9001/Monk

Zoekmachine van Marius

De zoekmachine van Marius geeft paragrafen terug: fragmenten in de KdK die zinvolle eenheden tekst vormen. De gevonden woorden worden in het geel gemerkt: hierbij een client-side oplossing gekozen. Indien een term wordt ingevoerd die niet in de index aanwezig is, wordt een suggestie gedaan die *wel* in de index staat.

Deze zoekmachine demonstreert de mogelijkheden van het herkennen van de lay-out van de pagina.

http://aps.ai.rug.nl/KdK/MariusMonk/index.php

Zoekmachine voor Word-zones

Ook deze zoekmachine is paragraaf gebaseerd. Er is voor gekozen om een parallellogram te gebruiken om het (schuingeschreven) gevonden woord te markeren.

Deze zoekmachine demonstreert op wat voor manieren word-zones kunnen worden gedetecteerd en dat er op kan worden gezocht.

http://rugtest5.service.rug.nl/scratch/kdkwz/search/

Associëren van vormen

Er is een zoekmachine gemaakt, waarbij gezocht wordt naar uitgeknipte woorden, waarvan is vastgesteld dat ze met enige zekerheid hetzelfde woord zijn.

Deze zoekmachine demonstreert de mogelijkheid van het associëren van vormen op woord niveau.

http://aps.ai.rug.nl/cgi-bin/wrdlist?cmd=TrainedWords&db=*&trainedwordmethod=BestSelection

4. Analyse

Lambert Schomaker heeft een laboratorium-omgeving ingericht waarin is en wordt geëxperimenteerd met de verschillende stappen die nodig zijn om van scans van een handgeschreven tekst te komen tot een zoekmachine waarin op de tekst kan worden gezocht. Aan deze omgeving is bijgedragen in het kader van het SCRATCH- en het MORPH project.

Bij het SCRATCH project is een pijplijn ontwikkeld, waar "met zeven mijlslaarzen" door de procedures wordt gelopen. Dit om een *proof of concept* te realiseren. Hierdoor zijn niet op alle plaatsen in deze pijplijn robuuste oplossing gekozen; robuust in de zin dat zij niet zonder aanpassingen werkzaam zullen zijn voor alle denkbare scans en handschriften en niet schaalbaar is naar grootschalige toepassing door meerdere erfgoedinstellingen.

Lambert Schomaker is de enige persoon die voor continuïteit zorgt. Hij heeft overzicht over het systeem, haar componenten en de werking ervan. Hij heeft bovendien ook de belangstelling en het (wetenschappelijke) belang om deze omgeving voort te zetten en te verbeteren; anderen hebben andere interesses, andere ambities, al dan niet bij andere (wetenschappelijke) organisaties.

5. Aanpak van de ontwikkeling

De conclusies, omschreven in hoofdstuk 4, maken het noodzakelijk om ook anderen bij de ontwikkeling van het beoogde product te betrekken. Hierbij kan

1. een productie-omgeving in het leven worden geroepen bij een organisatie die de ontwikkeling van het product inricht als software-engineeringsproject;
2. vanuit de laboratorium-omgeving geleidelijk functionaliteit worden overgedragen aan een productie-omgeving;
3. de productie-omgeving als product aan erfgoedinstellingen worden aangeboden.

Voor deze productie-omgeving zijn drie zaken noodzakelijk:

1. Toegang tot reken- en opslagcapaciteit, die nu wordt ingevuld door het rekencentrum van de RUG;
2. Deskundigheid in het opzetten van de beoogde productie-omgeving met haar Kunstmatige Intelligentie componenten en gebruiksvriendelijke gebruikersinterfaces;
3. Een korte, betrouwbare communicatielijn tussen de eigenaar van de laboratorium-omgeving en de eigenaar van de productie-omgeving.

Deze drie zaken worden mogelijk door de oprichting van het TARGET onderzoeksprogramma.

5.1. Target

Het TARGET is een onderzoeksprogramma van ruim 30 miljoen euro waarin 10 partijen participeren. Onder deze partijen zijn grote en kleine bedrijven en onderzoeksinstellingen van de RUG[6].

Eén van deze partijen is Target Holding. Deze werkmaatschappij is speciaal opgericht om onderzoeksresultaten die in het target project worden verwacht te valoriseren[7] en intellectueel te beheren. De Holding staat onder leiding van Gert-Jan van Dijk[8].

Ook het onderzoeksinstituut ALICE waar Lambert Schomaker leiding aan geeft participeert. De laboratorium-omgeving van Schomaker is een demonstratieproject voor Target, waardoor deze bestaande infrastructuur in Groningen in stand zal blijven gedurende het target-project. Target Holding zou als aanspreekpunt kunnen dienen om diensten voor erfgoed instellingen te ontwikkelen op basis van de SCRATCH resultaten.

5.2. CATCH+

Gezocht moet worden naar een goede besteding van de CATCH middelen in combinatie van TARGET. CATCH+ stelt daarbij eisen aan de beschikbaarheid van de software volgens Open Source principes.

6 http://www.rug.nl/target/index
7 http://www.rug.nl/target/valorisatie/; http://www.rug.nl/target/nieuws/actueel/Valorisatie%20Target%20van%20start
8 http://www.target-holding.nl/

6. Eisen aan het product

De productie-omgeving waarin het doorzoekbaar maken van gescande handschriften wordt gerealiseerd heeft twee gebruikersinterfaces nodig. Daarnaast is een goede ondersteuning noodzakelijk van deskundigen. Het product moet extern te beheren zijn, duurzaam en technisch onafhankelijk zijn van de hostingpartij.

6.1. Gebruikersinterface voor beheerders

Elke instelling die deelneemt door een gegevensbestand in de vorm van scans aan te bieden, is zelf verantwoordelijk voor het onderhoud van deze bestanden. Daarom is er een gebruikersinterface noodzakelijk die gebruikt wordt door gebruikers in de rol van beheerder. Dit zullen veelal medewerkers zijn van de erfgoedinstelling. Deze gebruikersinterface moet de volgende activiteiten mogelijk maken:

Primaire bronbewerking:

- het controleren van de aangeboden scans met behulp van kwalitatief objectiveerbare normen, in maat en getal
- het uploaden, verplaatsen en verwijderen van scans in de applicatieomgeving
- het toevoegen en wijzigen van metadata (bijv. bestandsnamen)
- het sorteren en selecteren van scans
- het optimaliseren van de scans voor patroonherkenning (deskewing, contrast- en kleurbewerking)
- het formuleren van de specifieke kenmerken van de te bewerken bron (beschrijvend en technisch)
- het definiëren van layout-karakteristieken (headers, kolommen, paragrafen, footers)
- het corrigeren van automatisch gegenereerde layout in een layout-editor
- het starten en afbreken van het patroonherkenningsproces
- het testen van de resultaten
- het in batch kunnen verwerken van bepaalde handelingen, hiervoor genoemd

Secundaire bronbewerking:

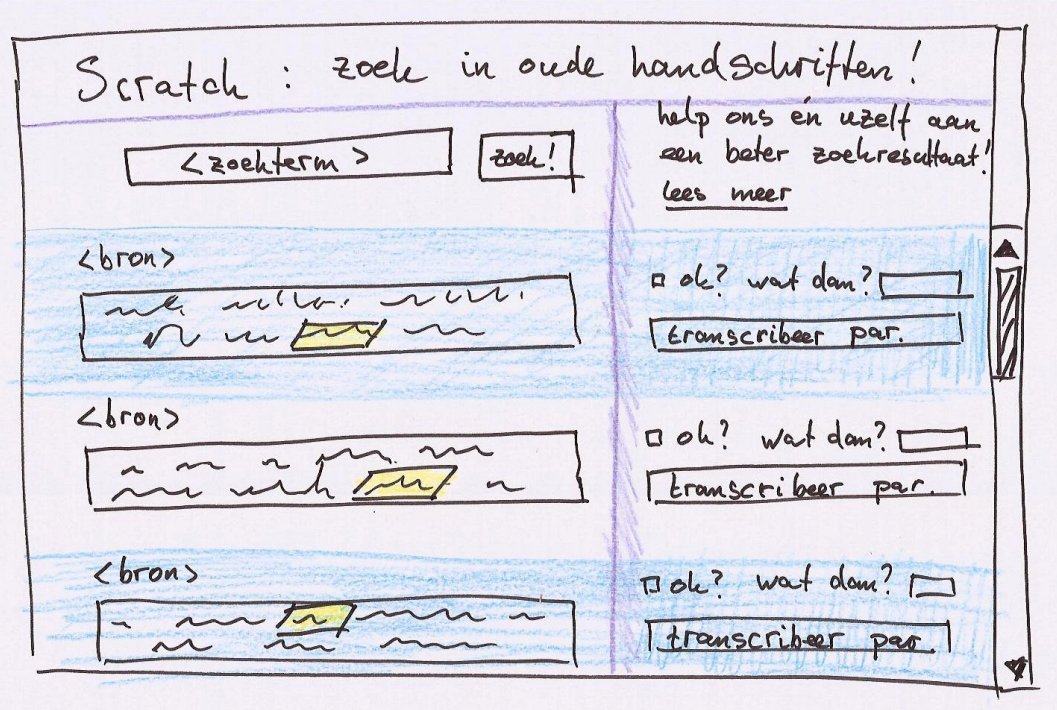
- het transcriberen van geschreven tekst
- het corrigeren en goedkeuren van getranscribeerde tekst
- het corrigeren en accepteren van automatisch herkende tekst

6.2. Gebruikersinterface voor publiek

Ten tweede is een gebruikersinterface nodig voor gebruikers in de rol van publiek. Dit zullen veelal gebruikers zijn van het gedigitaliseerde materiaal en vrijwilligers die -al dan niet voor een beperkte hoeveelheid- annotaties kunnen toevoegen. Deze gebruikersinterface moet het mogelijk maken dat:

- logische eenheden (bv. paragrafen) worden gepresenteerd die lijken te voldoen aan een ingevoerde zoekterm, met daarin de gezochte term gemarkeerd.

De gebruikersinterface kan worden uitgebreid met de mogelijkheid om de transcriptie van een deel van de tekst (of een individueel woord) toe te voegen. Er kan worden uitgelegd dat hiermee de kwaliteit van het zoekresultaat van een specifieke woord kan worden verbeterd. Een zoeker kan daarom ook uitgelegd worden dat hij een bijdrage levert uit eigen belang: na het toevoegen van extra transcripties van zijn gevonden woord, zal - na een rekenslag - de kwaliteit van zijn eigen zoekresultaat verbeteren.



afbeelding: schets van de presentatie van het zoekresultaat om de gedachte te vormen.

Daarvoor is het in deze gebruikersinterface ook mogelijk dat:

- bij een correcte classificatie de gebruiker kan aangeven dat inderdaad het gezochte woord is gevonden.
- bij een foutieve classificatie de gebruiker kan aangeven welk woord (of word-zone) dan wel is gepresenteerd.
- van een logische eenheid een volledige transcriptie kan worden ingevoerd.

6.3. Ondersteuning door deskundigen

Het proces van capture, segmentation en classification zal -zeker gedurende de verfijning van het product- niet altijd foutloos verlopen. Het is daarom van groot belang een deskundige helpdesk te hebben, die het proces "onder de moterkap" volgt en in overleg met de erfgoedbeheerder variabelen kan instellen om het proces te ondersteunen.

Deze deskundigen krijgen daardoor bovendien inzichten in de problemen, die ze vervolgens op een generieke manier in de software kunnen verwerken.

6.4. Non-functionele eisen

Technische beheer

Het hosten van de applicatie moet door een externe partij kunnen worden verzorgd. Hiertoe wordt ook gerekend het hosten van de datasets, die overigens een andere kan zijn dan de applicatie-host. Het technisch beheer hoeft daarmee niet in de erfgoed organisatie te worden belegd: deze zijn immers gespecialiseerd in erfgoed en niet in technisch beheer.

Standaardisatie

Het is noodzakelijk dat de informatie in een gestandaardiseerd formaat (gedacht wordt aan METS en ALTO) kan worden uitgelezen. Dit om de duurzaamheid van het resultaat te kunnen waarborgen.

Ook het gebruik van persistent identifiers moet worden gestandaardiseerd. Het is dan mogelijk ook op de lange termijn en vanuit allerlei informatiesystemen naar de images te verwijzen.

Onafhankelijkheid

Er moet een architectuur worden ontworpen, waarbij de tools onafhankelijk zijn van de faciliteit die rekencapaciteit en opslagcapaciteit verzorgd. Daarmee wordt het mogelijk in de toekomst voor andere platformen en omgevingen te kiezen wanneer dit om wat voor reden dan ook aan de orde is.

6.5. Toekomst

Een functie, die in de toekomst, die in de toekomst wenselijk wordt geacht, is de mogelijkheid om op een *image* te kunnen klikken, waardoor een link wordt gevolgd naar een ander gerelateerd document. Hierbij moeten de links automatisch worden gegenereerd op basis van de verwijzingen in de tekst .

Deze functionaliteit is een wens die slechts mogelijk is, indien de middelen beschikbaar zijn voor het digitaliseren van de gerelateerde documenten. Deze zijn in omvang vele malen groter, maar er kan worden volstaan met een wat mindere kwaliteit scans.