Boeken digitaliseren Stap 3 en 4

Handleiding van Helpmij.nl

Auteur: Kate95

januari 2017

De scans voorbewerken en omzetten naar tekst via OCR

Uitgangspunt

Je hebt van een boek alle pagina's gescand. Nu wil je daar graag een bewerkbare tekst van maken, om er later een epubje van te maken. Een epub kun je lezen op een e-reader, smartphone of pc, bijvoorbeeld met het geweldige programma **Calibre**.

Deze stappen zijn nodig om een boek te digitaliseren:

1. Het selecteren van de documenten
2. De pagina's scannen en hernummeren
3. De scans voorbewerken, netjes maken
4. Het omzetten naar bewerkbare tekst via OCR.
5. De tekst opschonen
6. De tekst omzetten naar een epub of ander formaat

In dit artikel behandel ik stap 3 en 4.

Stap 3: de tekst voorbereiden in Scantailor

De scans moeten worden voorbewerkt om het omzetten naar tekst zo vlotjes mogelijk te laten lopen. Rechte pagina's en geen vlekjes zorgen voor een beter resultaat. Je kunt deze stap overslaan, maar dan krijg je meer fouten in je tekst. Elk vlekje op een pagina wordt namelijk als leesteken geïnterpreteerd.

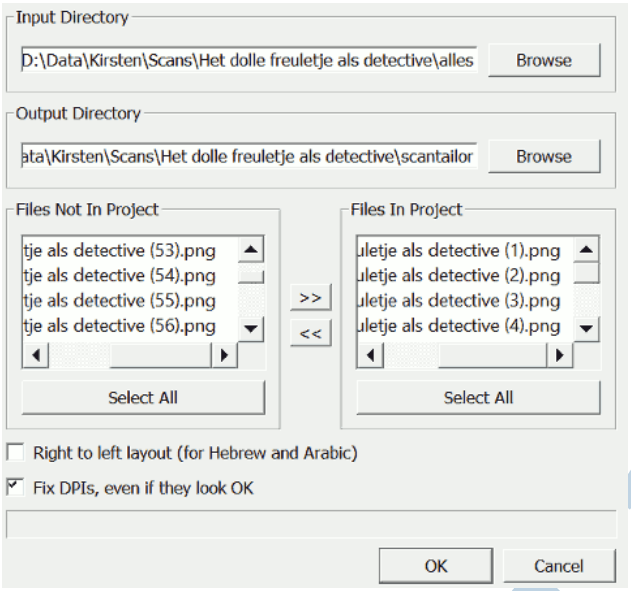
Om de scans te fatsoeneren, gebruik ik het gratis programma **Scantailor**. Dit programma is open source, gratis én beschikbaar voor Windows en Linux. Te downloaden op: http://scantailor.org/downloads Het programma is in het Engels.

Testen

Het is slim om het hele proces eerst te testen. Neem daarvoor eerst een paar pagina's, in plaats van meteen het hele boek. Dan kun je wat spelen met de instellingen en kijken of het goed werkt en alles eventueel aanpassen.

Nieuw project

Wanneer je Scantailor start, begin je met een **Nieuw Project**. Geef de map op van de scans en de map waar je de bewerkte scans heen wilt schrijven. De afbeeldingen die je wilt bewerken, zet je in de rechterkolom.
Klik ook aan "**Fix DPIs even if they look OK**". Dat is om afbeeldingen te verbeteren indien nodig. Is de resolutie te laag, vul dan in dat elke bladzijde 300 x 300 dpi moet worden.



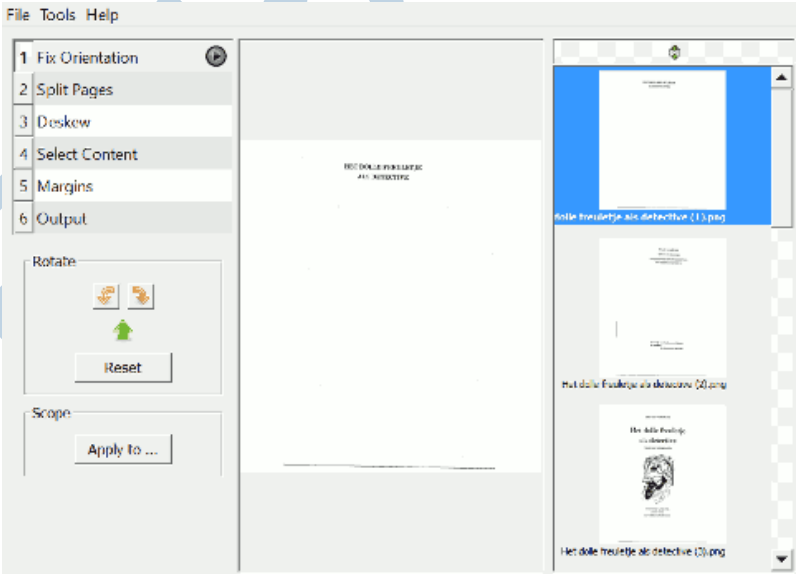
Opslaan

Tussendoor het project opslaan is natuurlijk altijd een heel goed idee.

Start

Het scherm ziet er als volgt uit:

Links de door te werken stappen, in het midden één enkele bladzijde en rechts alle bladzijden.



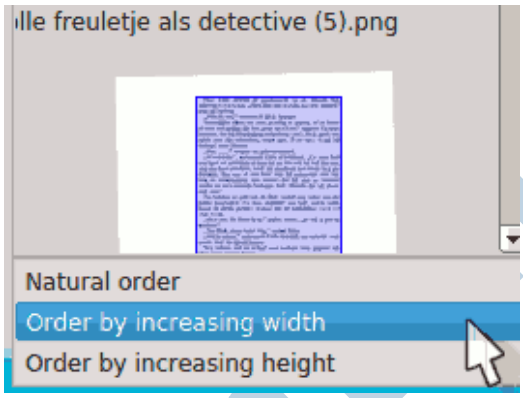
Automaat

Automatisch alle pagina's doorwerken, doe je door op de pijl te klikken in een stap. Heb je haast, klik dan meteen in stap 5 en daarna stap 6 op de pijl. Dan doorloopt het programma na elkaar alle stappen automatisch. Kijk of alles naar je zin is. Zo niet, dan kun je per stap de instellingen veranderen.

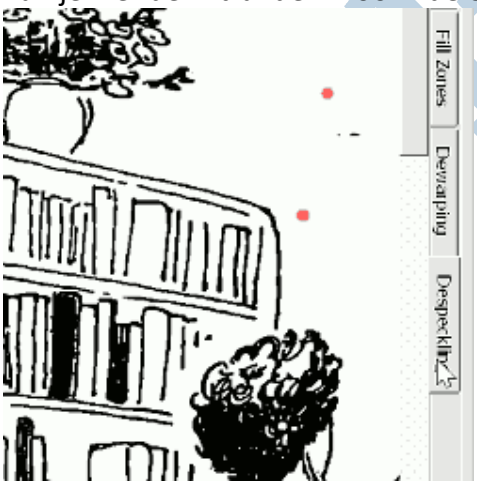
Handmatig aanpassen

Dit handmatig aanpassen doe je in de stappen van boven naar beneden:

1. **Fix orientation** = Pagina's draaien. Klik op de pijl en kijk of alles naar wens is. Met de knop **Scope** en **Apply to** kun je aangeven voor welke pagina's jouw instellingen gelden.
2. **Split pages** = Pagina's splitsen. Klik eerst weer op de pijl. De tekst moet in een gekleurd vlak staan. Heb je meerdere kolommen, of bijvoorbeeld twee pagina's naast elkaar, en is dat niet automatisch gedetecteerd, dan kun je dit hier aanpassen.
3. **Deskew** = Rechtzetten. Je kunt dit handmatig aanpassen door de bolletjes te slepen.
4. **Select content** = De inhoud selecteren. Ergens stonden bij mij wat streepjes die geen deel uitmaakten van de tekst.
   Door te slepen kun je hier de selectie aanpassen. Heel handig is de optie om de pagina's op grootte te zetten. Deze functie zit rechts onderaan. Hiermee vind je snel de pagina's die afwijken qua grootte.



5. **Margins** = Marges. Kijk hier even of alle inhoud in het witte vak staat.
   Zou je van de scans bijvoorbeeld een pdf maken, zonder de tekst via OCR te bewerken, dan kun je hier de witranden mooi indelen.



6. **Output** = Het uiteindelijke resultaat.
   Ouput resolution: de ideale output is tweemaal de input, dus **600 dpi**. Dat geeft grotere bestanden, en ook de beste OCR resultaten, heb ik gemerkt.
   Mode: de kleur staat hier op **Black and White**.
   Thinner/ Thicker: hiermee kun je de lijndikte van de tekst aanpassen. Dat kan de tekst helderder maken.
   Despecling: om vlekjes weg te poetsen. Ik zet deze op de kleine kwast.
   Rechts naast de grote pagina zitten nog meer tabs, waarop je kunt zien wat het resultaat is van de instellingen. In het tabje **Fill Zones** kun je nog vlekken of lijnen handmatig wegpoetsen.

Output

Zo, alle instellingen naar wens? Klik dan op het pijltje naast de 6 en de pc gaat staan rekenen. Het resultaat staat in de map die je in het begin hebt opgegeven. Sla je project nog een keer op.

Stap 4: De plaatjes omzetten naar bewerkbare tekst via OCR

OCR staat voor *optical character recognition*. Het is het proces om tekst van een plaatje om te zetten naar bewerkbare tekst.

GImageReader installeren

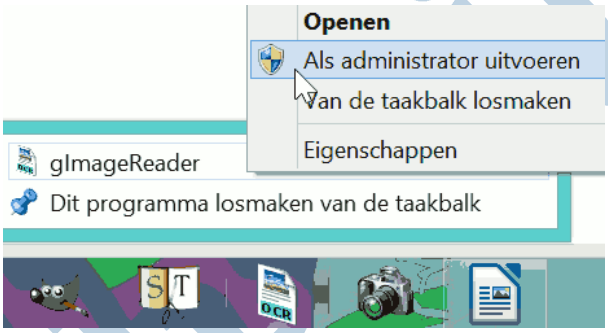
Voor OCR is er het programma *gImageReader*. Dat programma gebruikt onderliggend *Tesseract*. Het is beschikbaar voor Windows en Linux. Het is gratis én open source.
Let wel: onder Windows zijn goede gratis programma's voor OCR dungezaaid. GImageReader heeft de mogelijkheid om meerdere afbeeldingen in één keer om te zetten naar tekst. Onder Linux is de keuze heel wat ruimer, OCRfeeder of YAGF zijn ook prima keuzes.

Je moet even moeite doen voor de installatie, maar dan heb je een waardevol programma! Daarna is alles een fluitje van een cent.

Download de Windowsversie van: https://github.com/manisandro/gImageReader/releases

Aanvullende taalpakketten installeren

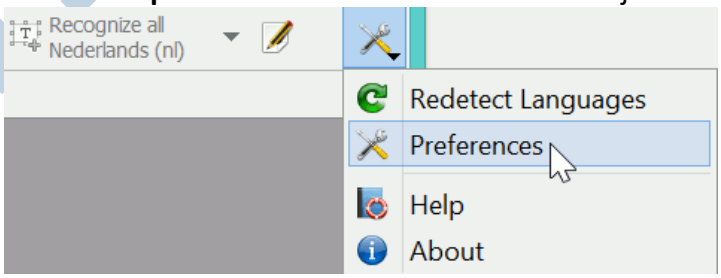
GimageReader heeft nog wat de Nederlandse taalpakketten nodig. Om die te installeren, moet je eenmaal het programma opstarten (openen door 2 maal rechts te klikken) met beheerdersrechten.



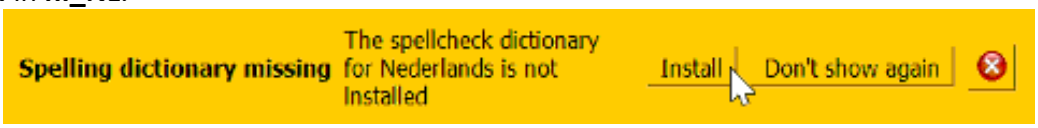
Laad eerst één plaatje in het programma. Daarna kun je de taalopties instellen.

Via de functie **Menubalk > Recognize all English > Manage Languages** download je Nederlands voor de OCR functie.
Klik aan **Nederlands** en dan **Apply** en **Close**.
Dan nog het woordenboek voor de correcte spellingscontrole.

Dit doe je via **Menubalk > Knop rechts > Preferences**. Vink alle vakjes aan.



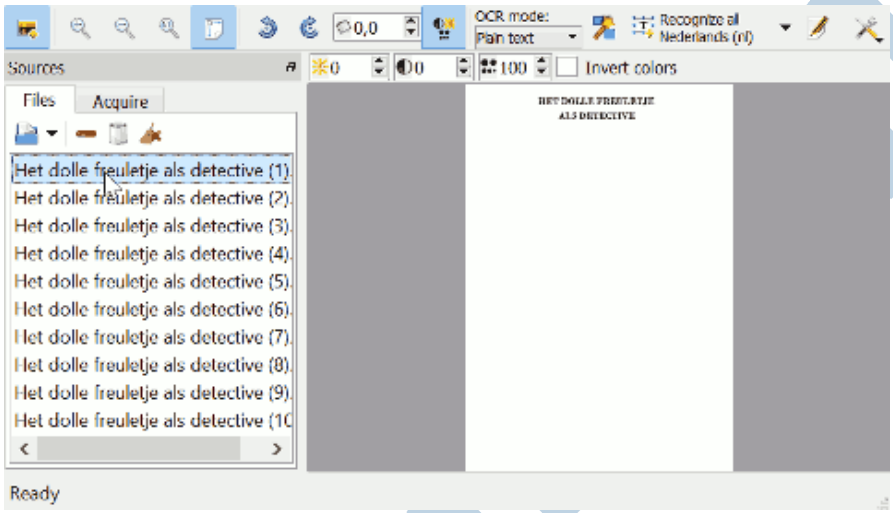
In het vak eronder zoek je vervolgens naar de regel waarin het Nederlands staat. Daar moet je de code voor het Nederlands aanpassen, want die is niet correct. Verander daar **nl** in **nl_NL**.



Wanneer je op **OK** klikt, geeft het programma aan dat het missende woordenboek geïnstalleerd kan worden. Dat gaat goed als je op **Install** klikt. Klik nogmaals op **OK**. Hè hè de installatie is klaar.

OCR met 10 testplaatjes

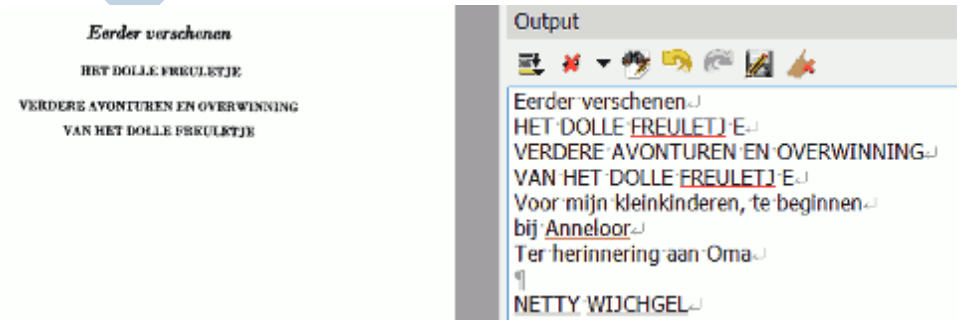
In Scantailor had ik 10 afbeeldingen bewerkt, om te testen. Die ga ik nu openen in gImageReader. Gewoon de bestanden selecteren in de verkenner en in gImageReader slepen werkt.

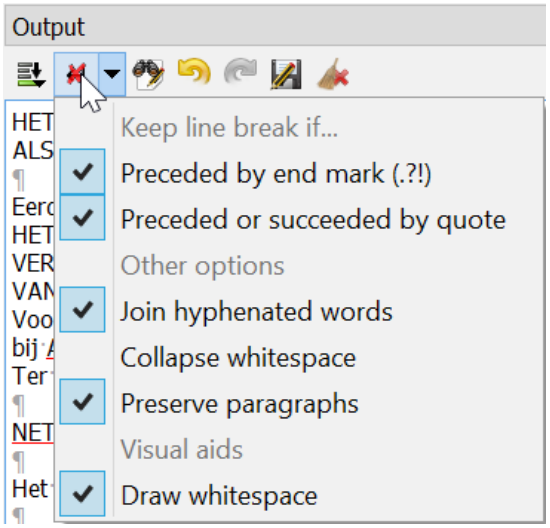


Nu stel ik de OCR taal in op **Nederlands** via het kleine pijltje rechts van de knop **Recognize all English > Nederlands > Dutch; Flemish (Netherlands)**.
Ik klik op **pagina 1** en op de knop **Recognize all Nederlands**. Dan opent er rechts een schermpje met daarin de tekst!
Als je alle pagina's tegelijk selecteert, en op de knop klikt, klik dan op **Multiple Pages** om meerdere pagina's te selecteren voor OCR.

Resultaat!

Mijn computer staat eventjes te rekenen en rechts verschijnt keurig de tekst, mét de Nederlandse spellingscontrole die aangeeft waar de fouten zitten. De tekst is direct te bewerken.





Verder is het mogelijk om overbodige enters te verwijderen en de afbreekstreepjes aan het einde van de zin. Alleen heb je hiermee weinig invloed op het eindresultaat.

Na enige tests volg ik daarom ook de gedachte: Vertrouwen is goed, controle is beter!
De tekst nabewerken doe ik daarom in later in de tekstverwerker LibreOffice.

Tekstverwerker

De tekst is nu op te slaan als txt bestand. Daarmee verdwijnen helaas leestekens.

Ik kopieer alles met *Ctrl+A* en *Ctrl+C* en plak het meteen in de tekstverwerker: zo blijven de leestekens zo precies mogelijk bewaard.
In het volgende artikel komen stap 5 en 6 aan de beurt: de tekst opschonen en exporteren naar een epub of ander formaat.

Succes tot zover!