

## WinVIS DigiDossier tekstherkenning (OCR)

### Inleiding

In de WinVIS module DigiDossier kunnen we alle denkbare documenten en bestanden archiveren en aan een relatie koppelen. Ongemerkt archiveren we zo vele pdf documenten, die we prima kunnen bekijken, maar waar we vaak niet zomaar teksten uit kunnen overnemen. Dan komt de vernieuwde **OCR (optical character recognition)** functie voor tekstherkenning direct van pas.

We gebruiken hiervoor enkele externe programma's, die de omzetting van een pdf bestand naar een tekstbestand regelen. Kern van de tekstherkenning is het onderdeel **Tesseract-OCR**, dat bijvoorbeeld bij de Universitätsbibliothek Mannheim gebruikt wordt voor het digitaliseren van historische Duitse dagbladen.

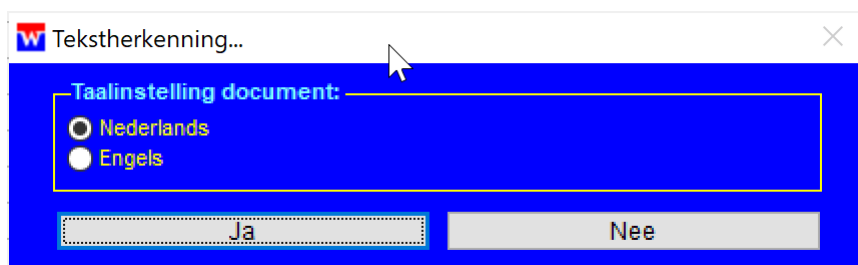
### 1. Hoe werkt het?

In het hoofdscherm van DigiDossierLog vindt u een knop met "ABC". Hiermee start u de tekstherkenningsmodule.



De tekstherkenning werkt voor alle pdf documenten. Wanneer u op de knop drukt krijgt u eerst een keuzevenster voor de taal, die u wilt gebruiken voor de tekstherkenning.

Op dit moment leveren we taalbestanden mee voor Engels en Nederlands. Mocht er behoefte zijn aan meer talen, dan horen we dat graag, zodat we die kunnen toevoegen.



Het pdf bestand wordt vervolgens ingelezen en de pagina's worden omgezet naar een zgn. tiff-bestand. Dit is een grafisch bestandsformaat, dat min of meer de standaard is voor tekstherkenningssoftware, waaronder dus ook Tesseract-OCR.

De feitelijke omzetting van pdf naar tiff gebeurt door het programma Image Magick in combinatie met GhostScript.

Het tiff-bestand wordt vervolgens ingelezen in Tesseract-OCR. WinVIS DigiDossierLog controleert bij het starten van het programma of de benodigde componenten aanwezig zijn. Ziet u drie groene bolletjes in onderstaand schermblok, dan zal de tekstherkenningsknop ook geactiveerd zijn.



De installatie van deze drie componenten gebeurt d.m.v. het WinVIS update- en installatiebestand. Het vinkje voor de DigiDossier onderdelen t.b.v. de tekstherkenning moet dan natuurlijk wel zijn blijven staan bij de installatie van dit bestand.

Alle herkende tekst wordt opgenomen in een tekstbestand. De bestandsnaam is afgeleid van de oorspronkelijke documentnaam van de pdf.

Heeft een pdf maar één pagina, dan werkt de tekstherkenning automatisch op die ene pagina. Zijn er echter meer pagina's, kunt u kiezen voor welke pagina's u de tekstherkenning wilt toepassen; maar voor alle pagina's kan dat natuurlijk ook!

Overzicht documentpagina's...

Pagina's: Schermvoorbeeld:

Pagina naam	Selecteer
Toetsing bouwbesluit_001.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_002.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_003.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_004.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_005.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_006.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_007.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_008.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_009.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_010.tiff	<input checked="" type="checkbox"/>
Toetsing bouwbesluit_011.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_012.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_013.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_014.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_015.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_016.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_017.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_018.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_019.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_020.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_021.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_022.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_023.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_024.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_025.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_026.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_027.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_028.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_029.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_030.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_031.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_032.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_033.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_034.tiff	<input type="checkbox"/>
Toetsing bouwbesluit_035.tiff	<input type="checkbox"/>

Selecteer alle bovenstaande pagina's

VOORSCHRIFTEN UIT HET DOELPUNT VAN VEILIGHEID:

**Artikel 1 Begripsbepalingen.**  
In deze bouwbesluitinformatie worden de begripsbepalingen gehanteerd als genoemd in artikel 1 Bouwbesluit.

**Artikel Gebruiksveiligheid**

**220 Verlichting**  
In de besloten verkeersruimtes alsmede in de besloten ruimten waardoor de vluchtmogelijkheid voert te tenminste een verlichtingsinstallatie aanwezig welke is aangesloten op de voorziening voor elektriciteit en die over de ten minste veraste breedte van de in die sector of ruimte gelegen besloten verkeersruimten, gemeten op de vloer van die verkeersruimten een verlichtingssterkte van ten minste 10 lux kan geven.  
De verlichtingsinstallatie moet zijn aangesloten op een voorziening van noodstroom die binnen 15 seconden na het uitvallen van de voorziening voor elektriciteit gedurende ten minste 90 minuten een verlichtingssterkte kan geven van ten minste 1 lux.

**Artikel Brandveiligheid**

**221 Beperking van ontbrekking van brand**

- a Een constructie-onderdeel, met uitzondering van een dak, moet ten minste behoren tot klasse 4 van de in die norm bedoelde bijdrage tot brandvoortplanting, met dien verstande dat de naar een vluchtweg toegelate zijde van een constructie-onderdeel ten minste moet behoren tot klasse 2.
- b In afwijking van het bovenstaande moeten een vloer en een tredvlak, bepaald overeenkomstig NEN 1776, ten minste behoren tot klasse T3 van de in die norm bedoelde bijdrage tot brandvoortplanting, met dien verstande dat een vloer en een tredvlak waarover een vluchtweg voert, moeten behoren tot klasse T1.
- c Het bovenstaande is niet van toepassing op ten hoogste 5% van de totale oppervlakte van de in die leden bedoelde constructie-onderdelen van elke afzonderlijke ruimte van het kantoorgebouw.

**222 Beperking van uitbreiding van brand**

- a Een brandcompartiment van een of meer met elkaar in verbinding staande afzonderlijke besloten ruimten waardoor geen vluchtweg voert, met een totale gebruiksoppervlakte van niet meer dan 1000 vierkante meter.
- b De in de NEN 8064 bedoelde weerstand tegen branddoorslag en brandoverslag tussen een brandcompartiment en een ander besloten ruimte ten minste 60 minuten zijn.
- c In afwijking van bovenstaande, mag tussen een brandcompartiment en een ander besloten ruimte zijn voorzien met een weerstand tegen branddoorslag en brandoverslag van ten minste 30 minuten, wanneer het gebouw waarin zich geen verblijfsgebied bevindt hoger dan 5 meter boven het aansluitende terrein, gemeten ter plaatse van de toegang van het gebouw.
- d In afwijking van bovenstaande, geldt met betrekking tot de weerstand tegen branddoorslag en brandoverslag tussen een brandcompartiment en een achtsel, lozer of lousel, welke voert langs ten hoogste een brandcompartiment als bedoeld in lid a, dan wel is gelegen in een uitlopend is bestemd voor een of meer boven elkaar gelegen slaalruimtes of badruimten, geen eis.
- e In een inwendige scheidingconstructie tussen een brandcompartiment en een andere besloten ruimte, niet zijnde een begruimte, mag zich geen beweegbaar constructieonderdeel bevinden, anders dan een zelfsluitende deur.

Start tekstherkenning      Annuleren

Voor iedere geselecteerde pagina wordt een tekstbestand aangemaakt. Deze worden opgeslagen in uw folder "Mijn Documenten". In bovenstaand voorbeeld is "Toetsing bouwbesluit\_011.tiff" geselecteerd, hetgeen een tekstbestand zal opleveren met de naam "Toetsing bouwbesluit\_011.txt".

## 2. Voorbehoud

Het kunnen gebruiken van de WinVIS DigiDossier tekstherkenning (OCR) functie is afhankelijk van software van derden, zoals u hierboven hebt kunnen lezen.

In principe hoeft u deze programma's niet te updaten. Het werkt immers goed zoals het nu is, met de nu bekende nieuwste versies van deze software. Het is echter mogelijk dat een wijziging in Windows of in andere software de werking van de tekstherkenning zou kunnen verstoren. In dat geval verzoeken wij u dit zo gedetailleerd mogelijk aan ons door te geven, zodat we dit kunnen onderzoeken en met een oplossing kunnen komen.

Mochten er relevante updates zijn van de software van derden, dan zullen we dit t.z.t. opnemen in een update. Ga hier dus niet zomaar zelf aan sleutelen!

## 3. Kosten

Aan het gebruik van de WinVIS DigiDossier tekstherkenning (OCR) functie zijn geen extra kosten verbonden. Deze functie is beschikbaar voor alle WinVIS gebruikers met een geldige licentie.

## 4. Document revisiehistorie

2022-05	eerste versie helpdocument
---------	----------------------------