



Faculteit Bedrijf en Organisatie

Automatische transformatie van ingescande tabellen naar gestructureerde digitale data

Milad Nazari

Scriptie voorgedragen tot het bekomen van de graad van  
professionele bachelor in de toegepaste informatica

Promotor:  
Martijn Saelens  
Co-promotor:  
Bram Vandewalle

Instelling: Predictiv NV

Academiejaar: 2019-2020

Derde examenperiode



Faculteit Bedrijf en Organisatie

Automatische transformatie van ingescande tabellen naar gestructureerde digitale data

Milad Nazari

Scriptie voorgedragen tot het bekomen van de graad van  
professionele bachelor in de toegepaste informatica

Promotor:  
Martijn Saelens  
Co-promotor:  
Bram Vandewalle

Instelling: Predictiv NV

Academiejaar: 2019-2020

Derde examenperiode



## **Woord vooraf**

Ik zou graag meneer Vandewalle willen bedanken voor enerzijds dit bachelorproefonderwerp en anderzijds voor de inhoudelijke ondersteuning, feedback en hulp die hij aangeboden en gegeven heeft. Hiernaast wil ik eveneens meneer Saelens danken voor de opvolging van mijn bachelorproef. Tenslotte wil ik mijn ouders eveneens bedanken voor hun onschatbare steun.



# 1. Samenvatting

Dataverwerkingen worden veelal nog op niet-digitale wijze opgeslagen en gelezen, onder meer d.m.v. papier. Data wordt op papier veelal in tabulair vorm teruggevonden. Het doel van dit onderzoek is het creëren van een open source prototype-software die het mogelijk maakt om tabellen in ingescande documenten te digitaliseren. Dit zou niet enkel digitaliseringssprocessen versnellen maar zou het eveneens gebruikt kunnen worden voor verschillende digitalisatietaken, zoals de digitalisatie van medicatieschema's.

Er werd eerst een literatuurstudie uitgevoerd om de stand van zaken rond tabeltransformatie te verduidelijken. Vervolgens zijn systeemvereisten en non-goals voor de proof-of-concept gespecificeerd. Op basis hiervan werden de algoritmes voor tabeltransformatie en de technologieën geselecteerd. Hierna werd de proof-of-concept geïmplementeerd en in detail toegelicht, bovendien werd een nieuwe algoritme voorgesteld voor structuuranalyse. Uiteindelijk werd de software op een dertig tal afbeeldingen getest.

Hoewel bij de test alle tabellen juist gedetecteerd werden, werd de nauwkeurigheid van de tabeltransformatie verlaagt door de minder performante tabelstructuuranalyse. De voorgestelde algoritme echter verhoogde de nauwkeurigheid van de structuuranalyse, al kan hierdoor de software nog niet als perfect beschouwd worden.

Men kan bij dit onderzoek concluderen dat tabeltransformatie een complex domein is. Kant en klaar software-pakketten bestaan, maar zijn betrekend en niet open source. Verder kan besloten worden dat tabelstructuuranalyse uit meerdere, van elkaar afhankelijke, subprocessen bestaat. Tenslotte kan men concluderen dat tabeldetectie enerzijds zeer nauwkeurig is, terwijl anderzijds structuuranalyse minder optimale resultaten kan leveren. De voorgestelde algoritme die een niet onbelangrijke verbetering van de tabeltransformaties teweegtbrengt, toont aan dat optimalisatiemogelijkheden zeker nog mogelijk zijn.



# Inhoudsopgave

<b>1</b>	<b>Samenvatting</b>	<b>5</b>
<b>2</b>	<b>Inleiding</b>	<b>19</b>
2.1	Probleemstelling	19
2.2	Onderzoeksvraag	22
2.3	Onderzoeksdoelstelling	22
2.4	Opzet van deze bachelorproef	22
<b>3</b>	<b>Stand van zaken</b>	<b>25</b>
3.1	Tabulair data	25
3.1.1	Definitie	25
3.1.2	Anatomie	27
3.1.3	Functie	27
3.1.4	Creatie en representatie	28

<b>3.2 Tabeltransformatie</b>	<b>28</b>
3.2.1 Tabeldetectie .....	29
3.2.2 Tabelstructuuranalyse .....	31
3.2.3 End-to-end-systemen .....	32
<b>4 Methodologie .....</b>	<b>33</b>
<b>4.1 Systeemvereisten</b>	<b>33</b>
4.1.1 Goals .....	33
4.1.2 Non-goals .....	34
<b>4.2 Selectie technologieën</b>	<b>34</b>
4.2.1 Tabeldetectie en Tabelstructuuranalyse .....	34
4.2.2 Programmeertaal .....	36
4.2.3 Interne tabelmodel .....	37
4.2.4 OCR .....	37
4.2.5 Back end server .....	38
4.2.6 Front end .....	38
<b>4.3 Performantiemeting</b>	<b>38</b>
<b>5 Proof of concept .....</b>	<b>39</b>
<b>5.1 Architectuur</b>	<b>40</b>
5.1.1 Documentinput .....	40
5.1.2 Preprocessing .....	41
5.1.3 Tabeldetectie .....	41
5.1.4 Tabelstructuuranalyse .....	42
5.1.5 Postprocessing .....	50
5.1.6 Resultaatweergave .....	50

<b>5.2 Praktisch gebruik</b>	<b>51</b>
5.2.1 Hardware- en software vereisten .....	51
5.2.2 Installatie en gebruik .....	51
<b>6 Resultaten .....</b>	<b>53</b>
<b>7 Conclusie .....</b>	<b>55</b>
<b>A Details van resultaten .....</b>	<b>57</b>
A.1 Document 1	58
A.2 Document 2	61
A.3 Document 3	64
A.4 Document 4	66
A.5 Document 5	68
A.6 Document 6	71
A.7 Document 7	74
A.8 Document 8	77
A.9 Document 9	79
A.10 Document 10	82
A.11 Document 11	84
A.12 Document 12	86
A.13 Document 13	89
A.14 Document 14	90
A.15 Document 15	92
A.16 Document 16	94

A.17 Document 17	95
A.18 Document 18	97
A.19 Document 19	99
A.20 Document 20	101
A.21 Document 21	103
A.22 Document 22	106
A.23 Document 23	107
A.24 Document 24	109
A.25 Document 25	111
A.26 Document 26	113
A.27 Document 27	115
A.28 Document 28	117
A.29 Document 29	120
A.30 Document 30	122
<b>B Onderzoeksvoorstel .....</b>	<b>127</b>
B.1 Introductie	127
B.2 State-of-the-art	128
B.3 Methodologie	128
B.4 Verwachte resultaten	129
B.5 Verwachte conclusies	129
<b>Bibliografie .....</b>	<b>131</b>

# Lijst van figuren

2.1	Voorbeeld van een tabelafbeelding. Bron: Microsoft (2020) .....	20
2.2	Voorbeeld medicatieschema. Bron: Tsjoen (2020) .....	21
3.1	Een tabel van evaluaties. Het geeft dezelfde informatie weer als tabelfiguur 3.2. Bron: Long, 2010 .....	26
3.2	Een tabel van evaluaties. Het geeft dezelfde informatie weer als tabelfiguur 3.1. Bron: Long, 2010 .....	26
3.3	De anatomie van de structurele rij-kolomvoorstelling van een tabel. Bron: Wang, 1996 .....	27
4.1	Anatomie van een Pandas Dataframe. Bron: GeeksforGeeks (2020) 37	37
5.1	Architectuur van de proof-of-concept .....	40
5.2	GUI van de software bij eerste gebruik. ....	41
5.3	Origineel document .....	42
5.4	Gedetecteerde tabellen. ....	42
5.5	Origineel geïsoleerd tabel. ....	45
5.6	Adaptive thresholding. ....	45
5.7	Bitwise-not-operatie. ....	46
5.8	Erosie d.m.v. een horizontale structuur. ....	46

5.9	Dilatatie d.m.v. de horizontale structuur.	46
5.10	Simpele dilatatie.	47
5.11	Simpele erosie.	47
5.12	Lijnen gedetecteerd door algoritme A.	47
5.13	Verwijdering van tekstelementen.	48
5.14	Dilatatie.	48
5.15	Erosie.	48
5.16	Lijnen gedetecteerd door algoritme B.	49
5.17	Lijnen gedetecteerd door algoritme A.	49
5.18	Lijnen gedetecteerd door algoritme B.	49
5.19	GUI van de software wanneer de tabeltransformatie uitgevoerd is.	50
A.1	Origineel document.	58
A.2	Gedetecteerde tabellen.	59
A.3	Origineel document.	61
A.4	Gedetecteerde tabellen.	62
A.5	Origineel document.	64
A.6	Gedetecteerde tabellen.	65
A.7	Origineel document.	66
A.8	Gedetecteerde tabellen.	67
A.9	Origineel document.	68
A.10	Gedetecteerde tabellen.	69
A.11	Origineel document.	71
A.12	Gedetecteerde tabellen.	72
A.13	Origineel document.	74
A.14	Gedetecteerde tabellen.	75
A.15	Origineel document.	77
A.16	Gedetecteerde tabellen.	78
A.17	Origineel document.	79
A.18	Gedetecteerde tabellen.	80
A.19	Origineel document.	82
A.20	Gedetecteerde tabellen.	83

<b>LIJST VAN FIGUREN</b>	<b>13</b>
A.21 Origineel document.	84
A.22 Gedetecteerde tabellen.	85
A.23 Origineel document.	86
A.24 Gedetecteerde tabellen.	87
A.25 Origineel document.	89
A.26 Gedetecteerde tabellen.	89
A.27 Origineel document.	90
A.28 Gedetecteerde tabellen.	91
A.29 Origineel document.	92
A.30 Gedetecteerde tabellen.	92
A.31 Origineel document.	94
A.32 Gedetecteerde tabellen.	94
A.33 Origineel document.	95
A.34 Gedetecteerde tabellen.	95
A.35 Origineel document.	97
A.36 Gedetecteerde tabellen.	98
A.37 Origineel document.	99
A.38 Gedetecteerde tabellen.	100
A.39 Origineel document.	101
A.40 Gedetecteerde tabellen.	101
A.41 Origineel document.	103
A.42 Gedetecteerde tabellen.	104
A.43 Origineel document.	106
A.44 Gedetecteerde tabellen.	106
A.45 Origineel document.	107
A.46 Gedetecteerde tabellen.	108
A.47 Origineel document.	109
A.48 Gedetecteerde tabellen.	109
A.49 Origineel document.	111
A.50 Gedetecteerde tabellen.	112
A.51 Origineel document.	113
A.52 Gedetecteerde tabellen.	114

A.53	Origineel document.	115
A.54	Gedetecteerde tabellen.	116
A.55	Origineel document.	117
A.56	Gedetecteerde tabellen.	118
A.57	Origineel document.	120
A.58	Gedetecteerde tabellen.	121
A.59	Origineel document.	122
A.60	Gedetecteerde tabellen.	123

## Lijst van tabellen

4.1	Voor- en nadelen van tabeltransformatiealgoritmes .....	35
4.2	Tabeldetectieperformanties van de verschillende algoritmes .....	36



# Woordenlijst

## CNN

Een Convolutionele Neuraal Netwerk (CNN) is een klasse van diepe neurale netwerken veelal gebruikt voor computer visie.. 30, 31, 32

## F1-score

De F1-score is een numerieke waarde die men krijgt door de product van de Recall en de Precisie te delen door de som van de Recall en de Precisie, en deze waarde nog met 2 te vermenigvuldigen.. 36

## GNN

Een Graaf Neuraal Netwerk (GNN) is een neurale netwerkarchitectuur die relaties en interacties tussen knopen kan modelleren en deze voorstellen als numerieke waarden.. 31, 32

## Hidden Markov Model

Hidden Markov Model (HMM) is een model uit de statistiek waarin het te modelleren systeem een markov-proces is met onbekende parameters. Het doel is de verborgen parameters te bepalen op basis van de waarneembare parameters. De op deze manier verkregen parameters kunnen vervolgens worden gebruikt voor toepassingen als patroonherkenning.. 30

## IoU

De Intersection Over Union (IoU) is een concept die veelal gebruikt wordt bij objectdetectie. De IoU kan men berekenen door de intersectieoppervlakte te delen over de unieoppervlakte van twee regio's: de ground truth regio en de voorspelde regio. Met de IoU wordt steeds een threshold gekoppeld. Indien men voor een threshold van 0,5 kiest en de IoU voor een voorspelling gelijk is of groter is dan de

threshold van 0,5, dan heeft men een echt positief (true positive), anders heeft men te maken met een foutpositief. Een foutnegatief krijgt men wanneer een ground truth aanwezig is maar geen voorspelling gemaakt kon worden.. 36

**JSON**

JavaScript Object Notation (JSON) is een gestandaardiseerd dataformaat.. 34

**OCR**

Optical Character Recognition (OCR), optische tekenherkenning, is de transformatie van afbeeldingtekst in bewerkbare, digitale tekst.. 8, 19, 20, 22, 37, 40, 44, 47, 50, 53, 55

**Precisie**

De precisie is een numerieke waarde die bepaald wordt het aantal echt positieven te delen door de som van het aantal echt positieven en foutnegatieven.. 17, 36

**Recall**

De recall is een numerieke waarde die bepaald wordt door het aantal echt positieven te delen door de som van het aantal echt positieven en foutpositieven.. 17, 36

**REST**

Representational State Transfer (REST) is een software-architectuur voor Internetdiensten.. 34, 38

**ROI**

Een Region Of Interest (ROI) is een subset van een dataset, geïdentificeerd voor een specifiek doel.. 30

**SVM**

Een Support Vector Machine (SVM) is een binaire classificeerder; ze wijst aan de hand van een aantal kenmerken objecten toe aan een van twee klassen. Daarvoor moet ze eerst een numeriek model van deze objecten maken als punten in een vectorruimte.. 30

**tupel**

In de wiskunde en de informatica is een tupel (ook tuple) een eindige rij van objecten. In een tupel is de volgorde van belang; als de objecten in een andere volgorde staan is het een ander tupel. Ook hoeven de objecten niet van hetzelfde datatype te zijn.. 25

## 2. Inleiding

In deze sectie wordt de context en achtergrond rond deze bachelorproef meegedeeld. Alsook wordt de probleemstelling, de onderzoeksvragen en onderzoeksdoelstellingen uitgelegd. Daarbovenop wordt de opzet van de bachelorproef verduidelijkt.

### 2.1 Probleemstelling

Alhoewel meer en meer processen wereldwijd volledig digitaal plaatsvinden, worden toch nog een grote deel van procedures en data opslag uitgevoerd op niet-digitale manieren. Zo krijgen de meeste mensen hun facturen nog steeds per brief. Volgens de Federale Overheidsdienst Economie (2019) blijft het verzenden of ontvangen van facturen op papier een zeer gangbare praktijk. Zo verstuurde 90 % van de bedrijven er en 97 % ontving er in 2017. Daarbovenop worden kassatickets nog steeds afgedrukt op papier, en notities nemen op papier blijft de populaire keuze hoewel er tal van notitie-apps bestaan. Deze voorbeelden tonen aan dat essentiële data nog massaal op een niet-digitale, en dus niet-automatisch verwerkbare media bewaard wordt, namelijk op papier.

Tot enkele jaren geleden was dit probleem niet zo beduidend maar nu meer digitale platformen voor dataverwerking gebruikt worden, is het omzetten van data op papier naar digitale data, m.a.w. het digitalisatieproces steeds belangrijker geworden.

Hierdoor werden tal van digitalisatiesoftwareproducten ontwikkeld, zoals Abby FineReader en Adobe Acrobat Pro DC. Hoewel deze software producten veel features hebben, zoals OCR, tabelherkenning, formulierherkenning, etc, zijn ze betalend en closed source. Wat als gevolg heeft dat ze voor bedrijven een merkbare kost met zich meebrengen, naast een privacy- en veiligheidsrisico aangezien het om closed source software gaat.

Sommige bedrijven enkele van hun digitalisatie oplossingen open source gemaakt, zoals Google met diens bekende OCR-software, Tesseract OCR, die door iedereen gebruikt kan worden om tekst in foto's om te zetten in tekstdaten. Hoewel OCR op zich zeer belangrijk is voor digitalisatie, is het niet voldoende voor volledige digitalisatie. Zo kan men de relatie tussen verschillende documententiteiten, die normaal gezien grafisch wordt verduidelijkt, enkel met OCR digitaal niet overbrengen. In documenten worden relaties tussen woorden meestal a.d.h.v. een tabel verduidelijkt. Door gebruik te maken van OCR, verkrijgt men wel de tekst binnen een tabel, maar men verliest essentiële informatie rond de woorden, namelijk tot welke rij en kolom ze behoorden. Het valt tenslotte niet onder de verantwoordelijkheid van OCR-engines om naast tekstherkenning, ook nog tabeltransformatie uit te voeren.

	A	B	C	D
1	Product	Kw 1	Kw 2	Eindtotaal
2	Chocolade	€ 744,60	€ 162,56	€ 907,16
3	Gummibarchen	€ 5.079,60	€ 1.249,20	€ 6.328,80
4	Scottish Longbreads	€ 1.267,50	€ 1.062,50	€ 2.330,00
5	Sir Rodney's Scones	€ 1.418,00	€ 756,00	€ 2.174,00
6	Tarte au sucre	€ 4.728,00	€ 4.547,92	€ 9.275,92
7	Chocoladekoekjes	€ 943,89	€ 349,60	€ 1.293,49
8	Totaal	€ 14.181,59	€ 8.127,78	€ 22.309,37

Figuur 2.1: Voorbeeld van een tabelafbeelding. Bron: Microsoft (2020)

Indien men bij tabelafbeelding 2.1 enkel OCR voor digitalisatie zou gebruiken, dan verkrijgt men wel de tekst, zoals de tekststukken zoals "Kw 1", "Kw 2", "€744,60", "€ 162,56", en meer, maar men behoudt niet de relatie tussen de tekststukken. Hierdoor zal men enkel met OCR niet te weten komen of de verkoopbedrag van € 744,60 bij de eerste kwartaal behoort, of bij de tweede, wat essentiële informatie is voor verdere financiële analyse.

Tot heden bestaat er geen open source oplossing die tabellen in foto's transformeert naar digitale tabellen, m.a.w. naar digitale structuren waarbij de tekst, evenals de relatie tussen de verschillende teksten getransformeerd wordt. Daarom werd er voor deze bachelorproef besloten om een proof-of-concept van een tabeltransformatiesoftware te creëren die bij een foto automatisch tabellen detecteert en deze tabellen digitaliseert.

Een belangrijke professionele toepassing van digitale tabeltransformatie is het digitaliseren van ingescande medicatieschema's, door technologiebedrijven zoals Into.care die zich bezig houden met digitale gezondheidszorg. Medicatieschema's worden in de gezondheidszorg gebruikt om medicatiedata voor patiënten te bewaren en weer te geven. Volgens de definitie van Apothekersnetwerk (Apothekersnetwerk, 2013) is het medicatieschema een geheel van gestandaardiseerde informatie over de actieve medicatie van een patient, met inbegrip van de identiteit van de geneesmiddelen, hun dosering, indicatie, relevante gebruiksaanwijzingen en bijkomende informatie waar nodig. Het omvat zowel voorgeschreven als

niet-voorgeschreven geneesmiddelen en voedingssupplementen.

Deze oplijsting van de actieve medicatie van de patient is niet enkel een essentieel hulpmiddel voor de patient bij de correct inname van medicatie maar ook voor medische professionelen om bv. over- of onderdosering, dubbelmedicatie, en andere geneesmiddelgebonden problemen te voorkomen. Ook wordt het gebruikt bij de communicatie tussen zorgverstrekkers. Het medicatieschema wordt eveneens door verpleegsters geraadpleegd voor het klaarzetten van de medicatie.

<b>Apotheek Maudens</b>	<b>Tel:</b>						
<b>Brusselsesteenweg 713, 9050 GENTBRUGGE</b>	<b>GSM:</b>						
<b>Titularis: Elisabeth Maudens</b>	<b>Fax :</b>						
<b>Medicatieschema</b>							
Naam:	Geslacht	Datum:					
INSZ Nr. (of Nr. RR of Nr. ID):	Geboortedatum:	Arts :					
Pathologieën : Hypertensie	Allergieën / Intoleranties :						
Dagelijkse medicatie - gecodeerd	Eenheid	Ontbijt	10u	Middagmaal	16u	Avondmaal	20u
		Voor	Met	Na	Voor	Met	Na
RHUMAL COMPLET SACHET 90			1				
XARELTO 15 MG COMP PELL 98 X 15 MG			1				
SPIRONOLACTONE EG COMP 50X 25MG			1				
LOSARTAN EG COMP PELL 98 X 100 MG			1/2				
CARVEDILOL EG 25,00 MG COMP 98 X 25 MG			1				1/2
LYRICA CAPS HARDE - DUR 200 X 150 MG	Capsule						
Niet dagelijkse medicatie of niet gestructureerde	Eenheid	Posologie					
ALENDRONATE EG 70 MG COMP 12 X 70 MG		Chronische medicatie: wekelijks					
METATOP 2 MG COMP 30 X 2 MG		Chronische medicatie: 1 tablet bij het slapengaan dagelijks					
DAFALGAN CODEINE EFF 500MG TABL 32		Indien nodig:					

Figuur 2.2: Voorbeeld medicatieschema. Bron: Tsjoen (2020)

Zoals men in figuur 2.2 kan zien, wordt dit schema grafisch in tabulaire vorm gepresenteerd. Echter is de lay-out hiervan niet standaardiseerd; afhankelijk van de apotheker of andere zorgverstrekker worden andere kolomnamen, kolomverdeling, rand- en verdelingstijl, celgrootte en andere tabelelementen aangewend. Dit bemoeilijkt ernstig het ontwikkelen van een transformatiesysteem die ingescande medicatieschema's omzet in instanties van een uniform digitale datastructuur in bv. XML- of JSON-formaat voor digitale verwerking van de medicatiedata in gezondheidszorgplatformen.

Een open source tabeltransformatiesoftware zal automatisch medicatieschema's kunnen omzetten in een uniform digitale datastructuur. Hierdoor zal er geen manuele werk uitgevoerd moeten worden, wat tijd- en kostenreductie als positieve gevolg heeft. Daarbovenop, omdat het open source zal zijn, zal men verzekerd zijn dat Into.care niet zal te maken hebben met softwarelicentiekosten of privacyschending.

Hoewel het digitaliseren van medicatieschema's een belangrijke toepassing is, zijn er tal van andere potentiële toepassingen, aangezien tabellen zo vaak gebruikt worden. Zo zou

men tabeltransformatie eveneens kunnen gebruiken voor het inscannen van kassatickets, het analyseren van een sudokuspel, het digitaal weergeven van een - op een whiteboard gemaakte - matrix voor online leerplatformen, het verwerken van een foto van een voedingswaardetabel op de verpakking van voedsel, en meer. Het is duidelijk dat een open source tabeltransformatiesoftware een beduidende universeel meerwaarde zal aanbieden.

## **2.2 Onderzoeksvraag**

Men kan zich bij tabeltransformatie, en dus bij dit onderzoek, enkele vragen stellen.

- Is kant-en-klaar open source software voor tabeltransformatie beschikbaar?
- Uit welke processen bestaat tabeltransformatie? In welke volgorde vinden deze processen plaats?
- Is preprocessing van de afbeelding nodig om de nauwkeurigheid van de resultaten te bewaren? Indien ja, uit welke stappen bestaat deze preprocessing?
- Analoog, is postprocessing van de verkregen tabel noodzakelijk? Indien ja, uit welke stappen bestaat deze postprocessing?
- Hoe nauwkeurig kunnen de verkregen resultaten van tabeltransformatie zijn?

## **2.3 Onderzoeksdoelstelling**

Aangezien het doel van deze studie het creëren van een end-to-end tabeltransformatie-tool is, zal er niet alleen gestreefd worden subprocessen zoals OCR of preprocessing geïsoleerd te bestuderen maar evenwel de subprocessen te implementeren in code. Eveneens is het de bedoeling dat de componenten met elkaar op een geïntegreerde manier zullen kunnen functioneren.

Dit betekent dat de prototype niet enkel zal bestaan uit tabelanalysesoftware, maar alsook uit een GUI, een backend server, een preprocessing pipeline, en meer.

## **2.4 Opzet van deze bachelorproef**

De rest van deze bachelorproef is als volgt opgebouwd:

In Hoofdstuk 3 wordt een overzicht gegeven van de stand van zaken binnen het onderzoeksdomain, op basis van een literatuurstudie.

Verder wordt in Hoofdstuk 4 de methodologie toegelicht en worden de gebruikte onderzoekstechnieken besproken om een antwoord te kunnen formuleren op de onderzoeksvragen.

In Hoofdstuk 5 wordt vervolgens de architectuur van de proof of concept uitgelegd.

Eveneens worden de verschillende algoritmen in detail besproken.

Verder worden in Hoofdstuk 6 de met de proof of concept verkregen resultaten besproken en vergeleken.

En tenslotte wordt in Hoofdstuk 7, de conclusie gegeven en een antwoord geformuleerd op de onderzoeks vragen.



## 3. Stand van zaken

In dit hoofdstuk wordt de stand van zaken besproken wat tabeltransformatie van afbeeldingen betreft. Er wordt besproken wat tabulair data is, waarom tabellen belangrijk zijn in de huidige informatiewereld, wat er bedoeld wordt met tabeldetectie en structuuranalyse, waar de uitdagingen hierbij zich bevinden en tenslotte wordt er in detail de verschillende technieken besproken die ontwikkeld werden om tabellen te kunnen detecteren en analyseren, met hun voor- en nadelen.

### 3.1 Tabulair data

#### 3.1.1 Definitie

Zoals Zanibbi e.a. (2003) het aangeeft, is een tabel een vorm van visualisatie dat men gebruikt om ermee data op te zoeken en te vergelijken. Meer specifiek geeft, volgens Zanibbi e.a. (2003), een tabel indexeringsschema's weer voor relaties. Een relatie heeft een verzameling van  $\eta$  tupels, die de domeinen of dimensies van de relatie genoemd worden.

De dimensies kunnen d.m.v. verschillende combinaties van rijen en kolommen opgesteld worden, waardoor verschillende tabelopstellingen exact dezelfde informatie op verschillende manieren kunnen weergeven. Dit kan gedemonstreerd worden a.d.h.v. de volgende twee figuren.

Year	Term	Mark					
		Assignments			Examinations		Grade
		Ass1	Ass2	Ass3	Midterm	Final	
1991	Winter	85	80	75	60	75	75
	Spring	80	65	75	60	70	70
	Fall	80	85	75	55	80	75
1992	Winter	85	80	70	70	75	75
	Spring	80	80	70	70	75	75
	Fall	75	70	65	60	80	70

Figuur 3.1: Een tabel van evaluaties. Het geeft dezelfde informatie weer als tabelfiguur 3.2. Bron: Long, 2010

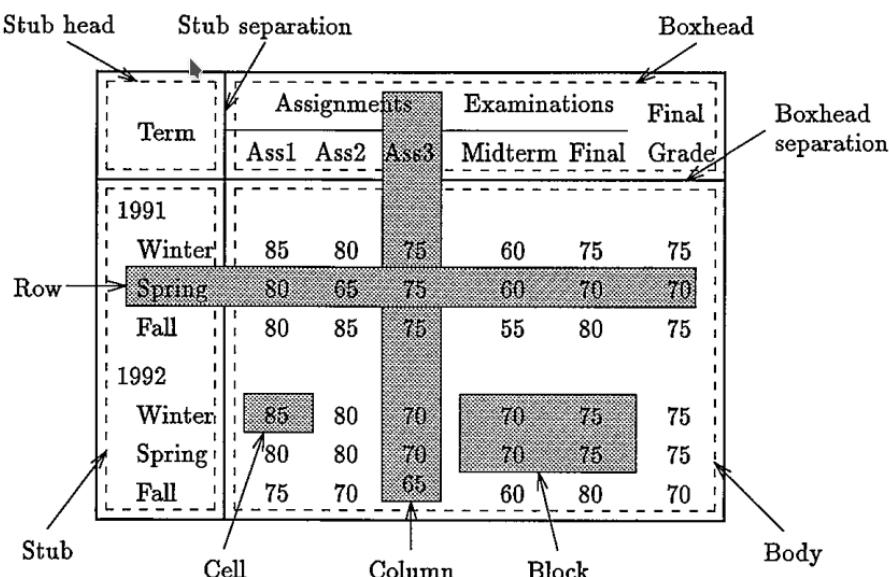
Term	Assignments	1991	1992
Winter	Ass1	85	85
	Ass2	80	80
	Ass3	75	70
	Midterm	60	70
	Final	75	75
	Grade	75	75
Spring	Ass1	80	80
	Ass2	65	80
	Ass3	75	70
	Midterm	60	70
	Final	70	75
	Grade	70	75
Fall	Ass1	80	75
	Ass2	85	70
	Ass3	75	65
	Midterm	55	60
	Final	80	80
	Grade	75	70

Figuur 3.2: Een tabel van evaluaties. Het geeft dezelfde informatie weer als tabelfiguur 3.1. Bron: Long, 2010

Hoewel beide tabellen identiek zijn wat informatieinhoud betreft, kan duidelijk gemerkt worden dat tabelfiguur 3.1 de evaluaties duidelijker weergeeft. Meestal wordt een combinatie van rijen en kolommen zodanig gekozen zodat de data van de tabel zo eenvoudig en snel mogelijk gelezen en geïnterpreteerd kan worden. Ook kunnen verschillende lettertypes, kleuren en lettergroottes gebruikt worden om de leesbaarheid te vergroten.

### 3.1.2 Anatomie

Volgens Wang (1996) is een tabel, door *stub scheiding* en *boxhead scheiding*, verdeeld in vier hoofdregio's die in onderstaande figuur 3.3 merkbaar zijn. De regio linksbeneden die de rijhoofdingen bevat en de regio rechtsboven die de kolomhoofdingen bevat, worden respectievelijk de *stub* en de *boxhead* genoemd. De regio linksboven, die de categorieën in de *stub* inhouden is gekend als de *stub head* en de *body*, tenslotte, is de regio rechts van de *sub* en onder de *boxhead* die de tabeldata-elementen bevatten. De snijpunt van een rij en een kolom wordt een *cel* genoemd; en een rechthoekig verzameling van *cellen* is gekend als een *blok*.



Figuur 3.3: De anatomie van de structurele rij-kolomvoorstelling van een tabel. Bron: Wang, 1996

Zoals men in figuur 3.3 kan zien, kunnen multidimensionele relaties in een tweedimensionale tabel gepresenteerd worden door meer dan één categorie te associeren met de *boxhead* en/of met de *stub*. Zo worden hier de rijhoofdingen niet enkel met één hoofdcategorie “Term” maar eveneens met meerdere subcategorieën, “1991” en “1992” geassocieerd. Analog zijn de kolomhoofdingen gekoppeld aan drie categorieën, namelijk “Assignments”, “Examinations” en “Finals”.

### 3.1.3 Functie

Als vermeld door Shahzad e.a. (2019), worden tabellen veelal gebruikt voor het gestructureerd vertonen van essentiële informatie in documenten. Ze worden gebruikt in boeken, artikelen, onderzoekspapers, en verschillende andere soorten media. In sectoren zoals de financiële en de administratieve sectoren wordt data veelal in tabelvorm geformuleerd omdat tabellen, volgens Coüasnon en Lemaitre (2014), veel informatie voorstellen op

een beknopte manier waardoor het begrijpbaar blijft voor de lezer; ze laten ook zo toe de belangrijke delen te benadrukken.

### 3.1.4 Creatie en representatie

Doorheen de tijd werden verschillende software applicaties ontwikkeld om digitaal tabulair data aan te maken, te beheren en voor te stellen. Een veelgebruikte software voor tabelcompositie is Microsoft Excel. Het is, zoals Wang (1996) het vermeldt, een complexe rekenbladprogramma waarbij tabulair data in een werkblad, in een tweedimensionele rooster die a.d.h.v. rij en kolomindexes geadresseerd kan worden, geplaatst wordt.

Een andere bekend software voor het creëren van tabellen is L<sup>A</sup>T<sub>E</sub>X. Het is een systeem voor het zetten van documenten. Wang (1996) geeft aan dat tabellen in L<sup>A</sup>T<sub>E</sub>X gespecificeerd kunnen worden met de “tabular”- en de “array”-omgeving. De eerste omgeving wordt meestal gebruikt voor tekstuele tabeldata, de tweede voor wiskundige uitdrukkingen.

Voor de voorstelling van tabellen op het internet, m.a.w. op internetbrowsers, wordt de opmaaktaal HTML gebruikt. Door middel van de “table”-, “tr”-, “th”- en “td”-tags kunnen tabellen gemaakt en voorgesteld worden.

## 3.2 Tabeltransformatie

Verschillende technieken voor tabeltransformatie werden reeds ontwikkeld. Echter blijft een algemeen toepasbare oplossing een moeilijk uitdaging, en dit voor diverse redenen.

- Tabellen bezitten uiteenlopende layouts en designs, zonder enige standaardisatie (Kasar e.a., 2014)
- Verschillende tabellayouts hebben verschillende features (Kasar e.a., 2014)
- De typisch kleine inter-klas variantie tussen tabellen, figuren en grafieken vermoedt de detectie van tabellen; de kleine variantie is verantwoordelijk voor de hoge hoeveelheid valse positieven bij tabeldetectie (Embley e.a., 2006)

Kasar e.a. (2014) beschreeft tabeltransformatie als een proces bestaand uit voornamelijk twee subprocessen: tabeldetectie en tabelstructuuranalyse.

Met tabeldetectie worden eerst regio's in een bepaalde document geïdentificeerd die overeenkomen met tabellen. Vervolgens wordt tabelstructuuranalyse toegepast om relationele informatie te extraheren van de geïdentificeerde tabelregio's om de logische structuur van de tabellen te achterhalen, zoals bijvoorbeeld de rijhoofdingen, kolomhoofdingen, cellen en meer.

### 3.2.1 Tabeldetectie

Tabeldetectietechnieken kan men, kijkend naar de stand van zaken, opdelen in twee klassen: klassieke, op regelgebaseerde algoritmen enerzijds en de recentere, datagedreven algoritmen die gebruik maken van machinaal leertechnieken.

#### Regelgebaseerde technieken

Watanabe e.a. (1991) waren de auteurs van één van de vroegste werken om tabellen te identificeren. De basis voor de tabelidentificatie hier is de identificatie van individuele blokken, ingesloten door horizontale en verticale lijnsegmenten. Eerst worden lijnsegmenten gedetecteerd en hiermee wordt vervolgens de positie van hoekpunten, gevormd door deze lijnen, bepaald. Hierna wordt a.d.h.v. de positie van deze hoekpunten individuele blokken geïdentificeerd. De relatie tussen de verschillende blokken wordt uiteindelijk in globale en individuele boomstructuren gebruikt om te beslissen of het over een tabel gaat of niet.

Het jaar daarop stelden Laurentini en Viada (1992) een methode voor waarbij tekstregio's op een bottom-up manier gedetecteerd worden. De gedetecteerde karakters worden samengebracht tot woorden en deze woorden worden op hun beurt aan elkaar samengevoegd tot tekstblokken. Ook worden de scheidingslijnen gedetecteerd. Voor elke tekstblok wordt diens positie vergeleken met de scheidingslijnen, om te bepalen of het tot een bepaalde tabel behoort.

TINTIN werd door Pyreddy en Croft (1997) voorgesteld, om tabellen te detecteren. Hun algoritme steunt voor de analyse op de extra PDF-metadata van de PDF-documenten.

Enkele jaren later werd het systeem T-Recs, door Kieninger en Dengel (2001), voorgesteld. Het systeem vormt rechthoeken (bounding boxes) voor woorden in het tabel en op een bottom-up manier worden deze bounding boxes gegroepeerd volgens hun logische eenheden.

#### Datagedreven technieken

Datagedreven technieken vereisen veel data om nauwkeurig te kunnen werken. Indien het om gesuperviseerde machinale leertechnieken gaat, is er daarbij nog ground truth labeling van de tabeldatasets nodig, wat een tijdrovend proces is. Wangt e.a. (2001) ontwikkelden een galabelde dataset generator die op basis van één gelabelde tabel, met bijhorende metadata, automatisch, mits kleine aanpassingen, datasets van gelabelde tabelafbeeldingen genereert en dus de ground truth labeling proces automatiseert. Deze tool zou zeer handig geweest zijn voor modeltraining van datagedreven algoritmen en voor algoritme-evaluaties. Wangt e.a. werden voor gecontaceerd om deze tool beschikbaar te maken maar helaas is de software niet meer ter beschikking.

Objectdetectie bij afbeeldingen d.m.v. machinale leertechnieken is sinds enkele tiental jaren een populair onderzoeksonderwerp geworden. De traditionele pipeline voor

objectdetectie bestaat uit een feature extraheerder (feature extractor), gevolgd door een classificatiesysteem.

Cesarini e.a. (2002) waren één van de eersten die geprobeerd hebben machinale leertechnieken te gebruiken voor tabeldetectie. Hierbij wordt, door de besproken methode Tabfinder, het document omgezet in een Modified-XY-boomvoorstelling en wordt er gezocht naar blokken omgeven door horizontale of verticale lijnen; dit gebeurt recursief. Indien deze verticale of horizontale lijnen gevonden zijn, dan bevat het document mogelijks een tabel. Om de veronderstelling te verifiëren, wordt er in diepere niveaus van de boom gezocht naar lijnen die loodrecht staan op de reeds gedetecteerde lijnen. Indien deze lijnen daarbovenop gevonden zijn, dan kan er met zekerheid vastgesteld worden dat het om een tabel in het document gaat. Na de boomanalyse worden vervolgens subtabellen behorend tot dezelfde tabel samengevoegd.

In een paper introduceerden Mandal e.a. (2006) een simpele maar efficiënte algoritme om tabellen te identificeren. De algoritme steunt namelijk op de observatie dat de hoeveelheid witruimte tussen elementen van verschillende kolommen significant groter is dan de witruimte tussen woorden in paragrafen.

Silva (2009) heeft eveneens een data-gedreven model voorgesteld, gebruikmakend van Hidden Markov Modellen en PDF-documenten. De tekst van de PDF-documenten wordt eerst omgezet in ASCII-karakters, en hierna verwerken de Hidden Markov Modellen de waarschijnlijkhedsdistributies van de samenhang van de verschillende opeenvolgende ASCII-karakters. De modellen houden voor elke horizontale lijn ook bij of het deel uitmaakt van een tabel, of niet.

Een SVM-classificatiesysteem, steunend op enkele manueel geselecteerde dimensiefeatures van de horizontale en verticale scheidingslijnen, werd door Kasar e.a. (2013) gepresenteerd. Om de donkere, dunne en lijnachtige structuren, die als scheidingslijnen van een tabel beschouwd worden, goed te kunnen detecteren, wordt de inputafbeelding eerst verzacht met een Gaussiaanse filter. Vervolgens worden op de input enkele top-hat-transformaties toegepast. Uiteindelijk wordt voor elke groep van kruising van horizontale en verticale lijnen een SVM-classifier met 26 lijnfeatures gebruikt om te bepalen of de regio deel uitmaakt van een tabel.

Fan en Kim (2015) gebruikten zowel een niet-gesuperviseerde leermodel als tekstuile informatie van een bepaalde zone voor de detectie van tabellen. De gebruikte niet-gesuperviseerde leermodel bestaat uit een ensemble van generatieve en discriminatieve modellen.

In hetzelfde jaar presenteerden Tran e.a. (2015) een methode die zich baseert op de ruimtelijke ordening van uitgehaalde tekstblokken en op ROI's. In tegenstelling tot verschillende andere traditionele algoritmen, is hun voorstel direct bruikbaar op afbeeldingen van gescande documenten.

Eén van de eerste pogingen om deep learning toe te passen, werd gerealiseerd door Hao e.a. (2016). Kandidaattabellen worden geselecteerd op basis van scheidingslijnfeatures. Hierna worden deze kandidaattabellen verwerkt door een CNN. Uiteindelijk vindt de klassificatie

“tabel” of “geen tabel” plaats.

Rashid e.a. (2017) gebruikten een bottom-up algoritme waarbij voor elk woord een feature-vector werd aangemaakt. Elke feature-vector bevat features zoals de dimensies van het woord, de afstand tot de andere woorden in de nabijheid, de hoeveelheid witruimte, etc. Deze feature-vectors werden uiteindelijk gebruikt om een AutoMLP-klassificatiessysteem te trainen.

Een andere deep learning techniek, gebaseerd op de Faster R-CNN-architectuur werd voorgesteld door Gilani e.a. (2017). Normaliter worden pixelwaarden als input gebruikt, voor een convolutionele neurale netwerk. Bij dit onderzoek is dat niet het geval. In plaats van pixelwaarden wordt de witruimte tussen tekstblokken als input verwerkt, aangezien bij tabellen witruimte tussen tekstblokken een bepaalde patroon bevat die door het model gedetecteerd wordt.

Faster R-CNN is een populaire modelkeuze geworden door de hoge nauwkeurigheid dat het aanbiedt bij objectdetectie in verschillende domeinen. Het bestaat, zoals meegedeeld door Shahzad e.a. (2019), uit twee processen die sequentieel uitgevoerd worden. Het eerste proces is Region Proposal Network (RPN) dat kandidaatregio's voor de detectienetwerk identificeert. Het tweede proces is de klassificatiesysteem die a.d.h.v. de detectienetwerk voor elke kandidaatregio besluit of het als een tabel beschouwd kan worden of niet. De RPN-module van Faster R-CNN biedt hogere performantie dan de selectieve zoekproces (selective search) van Fast R-CNN, de voorganger van Faster R-CNN.

Siddiqui e.a. (2018) introduceerden een incrementele architectuurverbetering, door Faster R-CNN te combineren met een vervormbare (deformable) CNN. De RestNet-101-model werd gebruikt voor transfer learning, aangezien er niet voldoende gelabelde data beschikbaar was. De reden waarom gebruik gemaakt werd van een vervormbare CNN i.p.v. een klassieke CNN is dat de klassieke CNN een vaste receptieve veld heeft (fixed receptive field), wat niet gewenst is door de verschillende dimensies en transformaties zoals verdraaiing, vergroting, verschuiving en meer dat tabellen kunnen hebben. Een vervormbare CNN kan zijn receptieve veld aanpassen aan de inputdata, hierdoor is het te gebruiken op elk soort tabel, ongeacht de layout ervan.

### 3.2.2 Tabelstructuuranalyse

Om de structuur van een tabel te kunnen analyseren, gebruikten Nazemi e.a. (2016) celsegmentatie. De horizontale en verticale scheidingslijnen van de tabel worden geïdentificeerd, en op basis van deze lijnen en hun kruisingen worden de tabelcellen gevormd. Voor tabellen zonder scheidingslijnen worden rijen gevormd door tekstblokken die dezelfde Y-waarde bezitten, en kolommen worden gevormd door tekstblokken waartussen weinig witruimte zit, m.a.w. door tekstblokken die relatief dicht bij elkaar staan.

Qasim e.a. (2019) meenden dat CNN, die normaliter een populaire keuze is, niet de ideale neurale netwerk architectuur is voor structuuranalyse van tabellen, op een datagedreven manier. Ze stelden een architectuur voor die gebaseerd is op GNN's. Hoewel een

tabelafbeelding verwerkt kan worden door een CNN, kan een tabel zelf beter verwerkt worden door een GNN aangezien de cellen van de tabel en de relatie tussen cellen eenvoudig gepresenteerd en gemodelleerd kunnen worden door een graaf. Een CNN wordt gebruikt voor visuele feature extractie, gevolgd door drie GNN's: één om te bepalen of twee tekstblokken tot dezelfde cel behoren, een tweede om na te gaan of twee tekstblokken deel uitmaken van dezelfde rij en tenslotte een derde GNN om te bepalen of twee tekstblokken tot dezelfde kolom behoren.

### 3.2.3 End-to-end-systemen

Tot hier werden onderzoeken besproken die ofwel het probleem van tabeldetectie ofwel het probleem van tabelstructuuranalyse aanpakten. Echter werden recent nieuwe technieken geïntroduceerd die voor beide subprocessen een oplossing aanbieden.

Zo hebben Green en Krishnamoorthy (1996) aan een layoutmodelgebaseerde top-down techniek gewerkt. Eerst worden horizontale lijnen, verticale lijnen en witruimte gebruikt om de tabelregio te detecteren. Vervolgens worden individuele cellen binnen de uitgehaalde tabel gelabeld, d.m.v. de layoutmodel; zo wordt relationele informatie van de cellen verkregen.

Even later hebben Oro en Ruffolo (2009) hun oplossing, PDF-TREX genoemd, ontwikkeld. Het is een systeem die op PDF-metadata steunt.

Een state-of-the-art deep learning algoritme, gebaseerd op afbeeldingen van ingescande documenten werd door Schreiber e.a. (2017) voorgesteld. Hierbij werd gebruik gemaakt van een Faster R-CNN-model, pretrained op de gekende Pascal-VOC-dataset, voor de detectie. De ZFNet (Zeiler & Fergus, 2014) en VGG-16 (Simonyan & Zisserman, 2014) modellen werden hierbij gebruikt voor transfer learning. Voor de tabelstructuuranalyse werd gebruik gemaakt van semantische segmentatie, m.b.v. een FCN-X-architectuur (Shelhamer e.a., 2016).

Tenslotte introduceerden Prasad e.a. (2020) recent een end-to-end-systeem voor tabeltransformatie. Hierbij werd gebruik gemaakt van één CNN model, namelijk een R-CNN-HRNet-model. Dit is een deep learning model specifiek ontworpen voor instantiesegmentatie, en niet voor de meer populaire toepassing, objectdetectie. Met dit model wordt instantiesegmentatie, en meer specifiek, tabelsegmentatie uitgevoerd op elke afbeelding waarin mogelijks tabellen zich bevinden om elke tabelinstantie te identificeren. Niet enkel wordt met dit model elke tabelinstantie geïdentificeerd, maar eveneens wordt elke tabelinstantie geklassificeerd als *bordered* of *borderless*. Een *bordered* tabel wordt gedefinieerd als een tabel waarin elke cel omgeven is door een rand. De cellen van een *borderless* tabel bezitten juist geen rand. Ook individuele tabelcellen worden gesegmenteerd. Nadat tabel- en celidentificatie uitgevoerd is, vindt de structuuranalyse plaats. Indien het om een *borderless* tabel gaat, wordt voor elke geïdentificeerde tabelcel een rij- en kolompositie toegekend. Indien de tabel van de type *bordered* is, worden de horizontale en verticale scheidingslijnen van de tabel gebruikt om de tekstblokken in tabelcellen te segmenteren; de geïdentificeerde tabelcellen worden hierbij genegeerd.

# 4. Methodologie

In dit hoofdstuk worden eerst het doel en de systeemvereisten van de proof-of-concept besproken. Vervolgens wordt de selectie van technologieën voor de proof-of-concept behandeld, inclusief de keuze van de tabeltransformatie-algoritme(n). Uiteindelijk wordt de evaluatiessysteem verduidelijkt die gebruikt wordt om de performantie van de proof-of-concept te beoordelen.

## 4.1 Systeemvereisten

### 4.1.1 Goals

Om een proof-of-concept succesvol te kunnen realiseren, moeten de volgende einddoelen hiervoor bereikt worden:

- Een end-to-end-systeem dient gecreëerd te worden. Dit betekent dat de proof-of-concept niet enkel een tabeldetectiecomponent moet bevatten, maar eveneens een tabelstructuuranalysecomponent, een GUI en andere nodige elementen om de tabeltransformatie zoveel mogelijk te automatiseren.
- De software moet modulair geïmplementeerd worden. Code geschreven voor tabeldetectie, bijvoorbeeld, mag niet afhankelijk zijn van code geschreven voor tabelstructuuranalyse en vice versa. Dit maakt het mogelijk om later één deel van de tabeltransformatiesoftware te verbeteren of opnieuw te herschrijven, zonder hierdoor een impact te hebben op de rest van de software.

- Enkel open source software en libraries, zonder commerciële restricties, mogen gebruikt worden.
- De proof-of-concept moet kunnen functioneren op verschillende eenvoudige tabellayouts. D.w.z. dat de tabeltransformatie niet mag afhangen van de fysieke dimensies van de tabel zelf of van de cellen. Nog mag het afhankelijk zijn van het aantal rijen of kolommen. Verder moet de software contextonafhankelijk zijn; een voedingswaardetabel bijvoorbeeld moet even nauwkeurig getransformeerd kunnen worden als een medicatieschema.
- Een tabeltransformatie moet kunnen uitgevoerd worden door een REST-server op te roepen. De tabeltransformatieresultaat die de server bij deze terug zal geven, moet een intuïtief, goed opgesteld JSON-object zijn.

#### 4.1.2 Non-goals

Naast einddoelen die de scope van de proof-of-concept vormen zijn er eveneens non-goals, vereisten die expliciet buiten de scope liggen:

- Het is niet de bedoeling om een volledige softwarepakket te ontwikkelen, met error handling, unit tests, integratie tests, authenticatie en meer.
- De proof-of-concept moet niet complexe tabellen kunnen verwerken. Hiermee worden voornamelijk tabellen bedoelt met meerdere niveau's van rijen en kolommen of tabellen met subtabellen.
- Preprocessing van de inputafbeeldingen wordt niet uitgevoerd. Er wordt verwacht dat de documenten correct zijn ingescand.
- Tabellen met handgeschreven tekst worden niet in beschouwing genomen.

## 4.2 Selectie technologieën

Op basis van de literatuurstudie (hoofdstuk 3) en de systeemvereisten (paragraaf 4.1) kan de selectie van algoritmes en technologieën plaatsvinden.

#### 4.2.1 Tabeldetectie en Tabelstructuuranalyse

De verschillende algoritmes voor tabeldetectie en tabelstructuuranalyse, behandeld in de literatuurstudie (hoofdstuk 3), hebben elk hun voor- en nadelen. Deze voor- en nadelen worden in de volgende tabel 4.1 weergegeven.

	Werkt zonder lijnen	Bruikbaar op afbeelding	Diverse layouts
<i>Tabeldetectie</i>			
Watanabe e.a.		✓	✓
Laurentini en Viada		✓	✓
Pyreddy en Croft	✓		✓
Kieninger en Dengel	✓	✓	
Cesarini e.a.		✓	✓
Mandal e.a.	✓	✓	
Silva			✓
Kasar e.a.		✓	✓
Fan en Kim	✓	✓	✓
Tran e.a.	✓	✓	✓
Hao e.a.			
Rashid e.a.	✓	✓	
Gilani e.a.	✓	✓	✓
Siddiqui e.a.	✓	✓	✓
<i>Tabelstructuuranalyse</i>			
Nazemi e.a.		✓	✓
Qasim e.a.	✓	✓	✓
<i>End-to-end-systemen</i>			
Green en Krishnamoorthy		✓	
Oro en Ruffolo	✓		✓
Schreiber e.a.	✓	✓	✓
Prasad e.a.	✓	✓	✓

Tabel 4.1: Voor- en nadelen van tabeltransformatiealgoritmes

Wat tabeldetectie betreft, bieden de algoritmes (Tran e.a., 2015), (Gilani e.a., 2017) en (Siddiqui e.a., 2018) alle voordelen aan. Alle drie technieken werken zonder de aanwezigheid van horizontale of verticale tabscheidingslijnen. Ze kunnen direct gebruikt worden op afbeeldingen, en zijn dus niet afhankelijk van PDF-metadata bijvoorbeeld. Tenslotte kunnen ze tabellen van verschillende layouts detecteren.

Van de technieken voor tabelstructuuranalyse heeft (Qasim e.a., 2019) alle voordelen.

Tenslotte bieden de end-to-end-technieken, (Schreiber e.a., 2017) en (Prasad e.a., 2020), eveneens alle voordelen.

Om het aantal kandidaat-algoritmes voor de proof-of-concept verder te verminderen, worden hun performanties met elkaar vergeleken. Omdat voor de tabelstructuuranalyse geen algoritmes te vinden zijn wiens performantiemeting met dezelfde methodologie uitgevoerd is, worden deze performanties niet in rekening gehouden. In onderstaande tabel 4.2 kan men de performantie van de verschillende technieken voor tabeldetectie terugvinden.

	Recall	Precisie	F1-score
Tran e.a.	0,9636	0,9521	0,9578
Gilani e.a.	0,9067	0,8230	0,8629
Siddiqui e.a.	0,996	0,996	0,996
Schreiber e.a.	0,9615	0,9740	0,9677
Prasad e.a.	1,0	1,0	1,0

Tabel 4.2: Tabeldetectieperformanties van de verschillende algoritmes

Deze performantiemetingen werden uitgevoerd op de ICDAR 2013 Table Competition dataset. Dit is één van de meest bekende datasets voor tabeldetectie en tabelstructuuranalyse. Het bevat in totaal 238 ingescande documenten. Voor de berekening van de F1-score wordt bij deze dataset een IoU threshold van 0,5 gebruikt.

Aangezien de tabeldetectiealgoritme van Prasad e.a. (2020) de meest performante is, wordt deze geselecteerd als de tabeldetectiecomponent van de proof-of-concept. Voor de tabelstructuuranalyse kan men nog kiezen tussen (Qasim e.a., 2019), (Schreiber e.a., 2017) en opnieuw (Prasad e.a., 2020). Hoewel (Qasim e.a., 2019) en (Schreiber e.a., 2017) gepaste keuzes zijn, wordt er beslist om voor de structuuranalyse eveneens (Prasad e.a., 2020) te selectioneren. Dit komt voornamelijk omdat Prasad e.a. (2020) niet enkel hun volledige modeltrainingdataset maar ook de code-implementatie van hun algoritme openbaar hebben gemaakt, met een open source licentie.

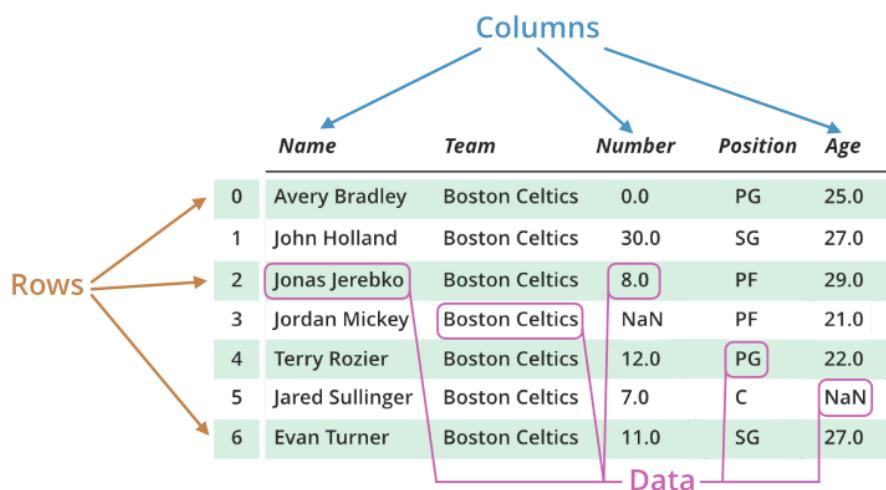
#### 4.2.2 Programmeertaal

Voor de programmeertaal van de proof-of-concept wordt Python (Van Rossum & Drake, 2009) gekozen, en dit voor verschillende redenen:

- Python is een multifunctioneel programmeertaal die toelaat om eenvoudig en snel softwareprototypes te creëren.
- Het heeft een grote bibliotheek van libaries voor onder meer statistiek en data-analyse, zoals Numpy (Oliphant, 2006) en Pandas (McKinney, 2010).
- De open source code van Prasad e.a. (2020) is reeds in Python geschreven. Door met Python verder te werken, kan code herbruikt worden en zal een een volledig nieuwe reïmplementatie van de software dus niet nodig worden.

### 4.2.3 Interne tabelmodel

Om de getransformeerd tabel te kunnen verwerken en verbeteren, moet de software van de proof-of-concept deze d.m.v. een datastructuur bijhouden. Hoewel voor de datastructuur JSON of XML gekozen kan worden, wordt er beslist om intern de getransformeerd tabel te presenteren en te verwerken d.m.v. een Pandas Dataframe. Een Pandas Dataframe is een twee dimensionale tabulaire datastructuur van de Pandas library. Deze datastructuur bestaat voornamelijk uit drie hoofdcomponenten, die in onderstaande figuur worden weergegeven: rijen, kolommen en data.



Figuur 4.1: Anatomie van een Pandas Dataframe. Bron: GeeksforGeeks (2020)

Niet enkel modelleert een Pandas Dataframe, met zijn twee dimensionele structuur, op een gepaste manier een tabel maar het bezit eveneens andere voordelen. Zo kan de opgeslagen data heterogeen zijn, d.w.z. dat de datatypes van de data-elementen niet identiek hoeven te zijn. Verder bezit het meerdere functionaliteiten om data eenvoudig te manipuleren. Tenslotte kan een Pandas dataframe zeer efficiënt met grote hoeveelheden data werken.

### 4.2.4 OCR

Voor de OCR-component zijn er niet veel keuzes. Tesseract (Kay, 2007) is momenteel de enige OCR-software die open source is en nauwkeurig tekst in afbeeldingen detecteert en transformeert. Bovendien kan de Python library, Pytesseract (Lee, 2009), gebruikt worden om Tesseract-functionaliteiten met Python te gebruiken. Zo kan men Tesseract functionaliteiten d.m.v. Pytesseract in Python code oproepen en kunnen de OCR-resultaten als een Pandas Dataframe verkregen worden.

#### **4.2.5 Back end server**

Voor de REST-server wordt Flask (Grinberg, 2018) gekozen. Flask is een eenvoudige micro web framework, bedoeld voor Python-softwareontwikkeling, met REST-ondersteuning.

#### **4.2.6 Front end**

Er bestaan verschillende front end frameworks om snel en eenvoudig GUI's te ontwerpen die via de browser gebruikt zullen kunnen worden, zoals Angular (Jain e.a., 2014) en React (Fedosejev, 2016). Voor deze proof-of-concept wordt React gebruikt, in combinatie met de UI framework Ant Design (Financial, 2020). De UI framework Ant Design biedt herbruikbare UI-componenten, onder meer voor tabulair data, die de ontwikkeling van de proof-of-concept vereenvoudigen en versnellen.

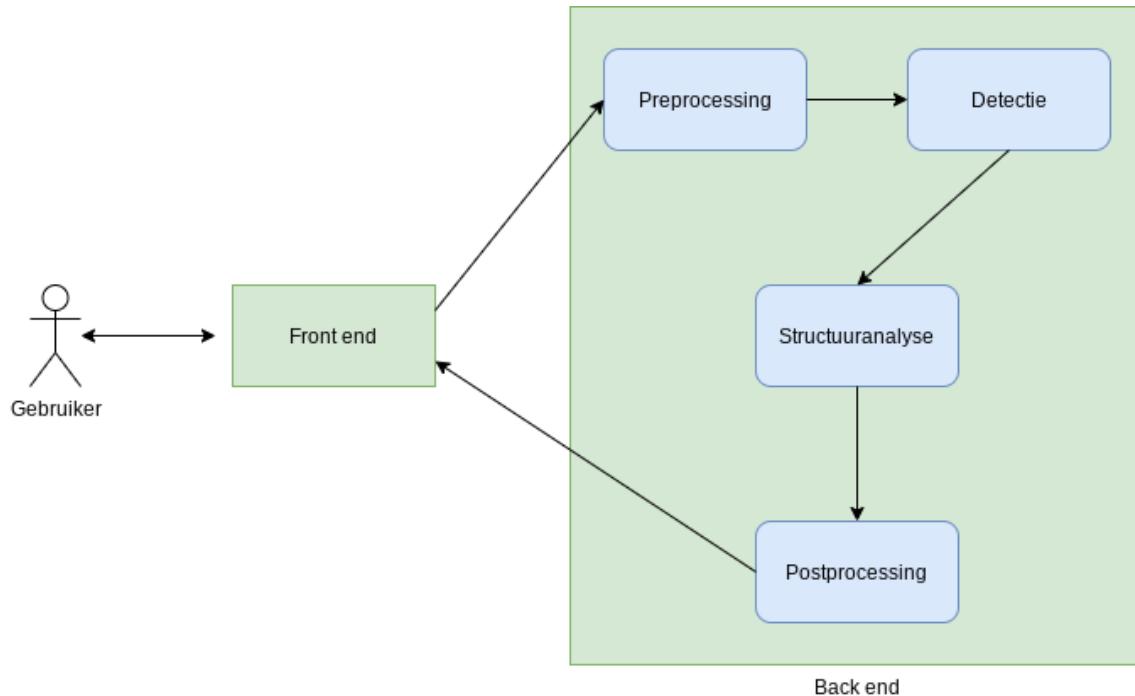
### **4.3 Performantiemeting**

De architectuur van de proof-of-concept wordt in de volgende hoofdstuk 5 meer in detail besproken. Om de performantie van de software te kunnen meten, wordt de software van de proof-of-concept uitgevoerd op dertig willekeurige ingescande documenten waarin in elk document één of meerdere tabelafbeeldingen te vinden zijn. Het resultaat van de tabeltransformatie op deze testdocumenten worden in hoofdstuk 6 behandeld.

## **5. Proof of concept**

De proof-of-concept wordt in dit hoofdstuk in detail behandeld. De verantwoordelijkheid en werking van de verschillende componenten wordt uitgelegd. Hiernaast wordt de flow (stroom) van de data verduidelijkt. Tenslotte wordt de installatie van de proof-of-concept-software uitgelegd.

## 5.1 Architectuur

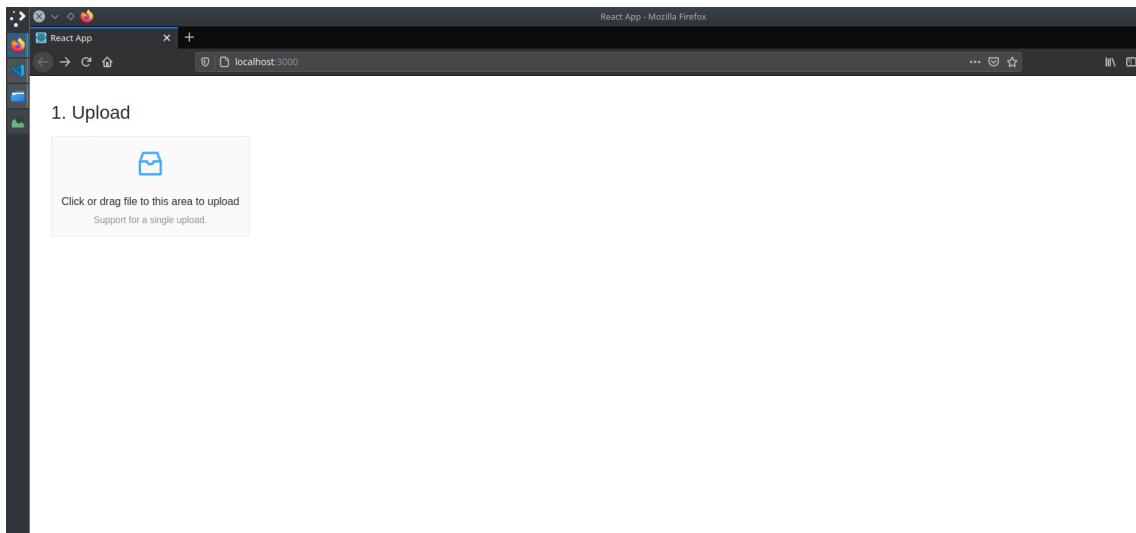


Figuur 5.1: Architectuur van de proof-of-concept.

In figuur 5.1 wordt de architectuur van de proof-of-concept weergegeven, inclusief de verschillende componenten zoals OCR, tabeldetectie en meer. De datastream, samen met de werking van de componenten, wordt in de volgende subsections verder toegelicht.

### 5.1.1 Documentinput

Wanneer de gebruiker de software wil gebruiken, zal hij/zij een GUI zien die in de volgende figuur weergegeven wordt.



Figuur 5.2: GUI van de software bij eerste gebruik.

De gebruiker klikt op de upload-zone, waar de tekst “Click or drag file to this area to upload” vermeld staat. Vervolgens selecteert de gebruiker het document die getransformeerd moet worden. Bij de bevestiging van de selectie van het bestand wordt deze geüpload naar de back end, waar preprocessing plaatsvindt.

### 5.1.2 Preprocessing

Bij de preprocessing wordt extra witruimte toegevoegd aan de afbeelding. Dit is nodig omdat de tabeldetectiealgoritme, hoewel het zeer goed werkt op afbeeldingen van ingescande documenten, moeilijkheden heeft met tabeldetectie wanneer de afbeelding enkel uit de tabel zelf bestaat, zonder extra tekstelementen, witruimte, grafieken, figuren of dergelijke. Dit komt omdat de deep learning model getrained werd op documentafbeeldingen en niet op tabelafbeeldingen zelf. Eens de preprocessing uitgevoerd is, wordt tabeldetectie op de nieuwe afbeelding, waar extra witruimte aanwezig is, uitgevoerd.

### 5.1.3 Tabeldetectie

Voor de detectie wordt een drempel waarde (threshold) van 0,85 gekozen. Dit heeft als gevolg dat indien de berekende kans op de aanwezigheid van een tabel door de deep learning model gelijk is aan, of groter is dan 0,85, dan wordt de aanwezigheid van één of meerdere tabellen in de afbeelding als waar beschouwd; anders niet.

In onderstaande figuren 5.3 en 5.4 kan men respectievelijk links een origineel ingescande document zien en rechts de gedetecteerde tabellen in het document.

UNITED NATIONS E-GOVERNMENT SURVEY 2014

months. Ease of use was further supported by access to content in more than one language in 74 per cent of cases (142 countries), availability of a site map or index 68 per cent of the time (131 countries) and the online publication of a help or frequently asked questions document in 46 per cent of cases (89 countries).<sup>1</sup>

The Survey shows continued effort of most countries in building and maintaining site-specific tools, notwithstanding the ubiquity and convenience of commercial search engines. There is also a growing recognition of the importance of providing content in different languages. In 2012, just over half of all countries had multilingual sites whereas according to the 2014 assessment almost three quarters had expanded language options in some form.

Enhanced (Stage 2) features are less common as Table 2.4 suggests. Roughly half of the United Nations Member States maintain an advanced search engine or publish a statement setting out a privacy policy in connection with the use of the government website. User opinion features such as tag clouds and ‘hot topics’ lists could be found on only 41 per cent of websites. Less than one third of national portals surveyed showed availability of a secure connection.

Table 2.3. Availability of selected basic features

	Number of countries	Per centage of countries
Find website using search tools	168	87%
Site updated within past three months	148	77%
Access in more than one language	142	74%
Availability of map/index	131	68%
Help/FAQ feature	89	46%

Table 2.4. Availability of selected enhanced features

	Number of countries	Per centage of countries
Advanced search option	101	52%
Privacy statement	97	50%
Tag cloud or ‘hot topics’	80	41%
Secure website	53	27%

Figure 2.4 provides a breakdown of typical transactional services and the number of countries for which these services could be readily identified through the national website. Of the transactional services included in the Survey instrument, the most commonly found were setting up of personal online accounts (101 countries), income tax filing (73 countries) and business registration (60 countries). An open-ended ‘other’ category also scored well (76 countries) reflecting a diversity of priorities in building and expanding online services at national level.

months. Ease of use was further supported by access to content in more than one language in 74 per cent of cases (142 countries), availability of a site map or index 68 per cent of the time (131 countries) and the online publication of a help or frequently asked questions document in 46 per cent of cases (89 countries).

The Survey shows continued effort of most countries in building and maintaining site-specific tools, notwithstanding the ubiquity and convenience of commercial search engines. There is also a growing recognition of the importance of providing content in different languages. In 2012, just over half of all countries had multilingual sites whereas according to the 2014 assessment almost three quarters had expanded language options in some form.

Enhanced (Stage 2) features are less common as Table 2.4 suggests. Roughly half of the United Nations Member States maintain an advanced search engine or publish a statement setting out a privacy policy in connection with the use of the government website. User opinion features such as tag clouds and ‘hot topics’ lists could be found on only 41 per cent of websites. Less than one third of national portals surveyed showed availability of a secure connection.

Table 2.3. Availability of selected basic features

	Number of countries	Per centage of countries
Find website using search tools	168	87%
Site updated within past three months	148	77%
Access in more than one language	142	74%
Availability of map/index	131	68%
Help/FAQ feature	89	46%

Table 2.4. Availability of selected enhanced features

	Number of countries	Per centage of countries
Advanced search option	101	52%
Privacy statement	97	50%
Tag cloud or ‘hot topics’	80	41%
Secure website	53	27%

Figure 2.4 provides a breakdown of typical transactional services and the number of countries for which these services could be readily identified through the national website. Of the transactional services included in the Survey instrument, the most commonly found were setting up of personal online accounts (101 countries), income tax filing (73 countries) and business registration (60 countries). An open-ended ‘other’ category also scored well (76 countries) reflecting a diversity of priorities in building and expanding online services at national level.

51

Figuur 5.3: Origineel document

Figuur 5.4: Gedetecteerde tabellen.

Zoals men kan zien in figuur 5.4, worden niet enkel de tabellen zelf maar eveneens cellen gedetecteerd. De detectie van tabellen is zeer accuraat, hoewel de detectie van cellen niet even nauwkeurig is. In de figuur bijvoorbeeld werden sommige cellen niet gedetecteerd.

Nadat de tabellen gedetecteerd zijn, worden ze bijgesneden om enkel de tabellen zelf te behouden en niet meer de rest van het document. Op elke geïsoleerd tabel worden de volgende stappen, tabelstructuuranalyse (5.1.4) en postprocessing (5.1.5) uitgevoerd.

## 5.1.4 Tabelstructuuranalyse

### Origineel algoritme

Bij de tabelstructuuranalyse-algoritme van Prasad e.a. (2020), die vanaf dit punt als “algoritme A” genoemd zal worden, wordt eerst gekeken naar de soort gedetecteerd tabel die bepaald en meegegeven werd door de tabeldetectie-component. Indien het om een bordered tabel gaat, wordt lijndetectie toegepast en wordt vervolgens d.m.v. cellsegmentatie aan elke cel een rij- en kolomwaarde toegekend. Anders, indien het dus als een borderless tabel gedetecteerd werd, wordt de positie van mogelijks gemiste, niet-gedetecteerde cellen voorspeld. De voorspelde cellen, samen met de reeds gedetecteerde cellen, worden vervolgens gebruikt om de rijen en kolommen te vormen.

Eens de tabel getransformeerd is, wordt het resultaat opgeslagen in een XML-bestand. Een

voorbeeld hiervan wordt in de volgende code 5.1 weergegeven.

```

<?xml version="1.0" encoding="UTF-8"?>
<document>
  <table>
    <Coords points="387,531_387,738_1009,738_1009,531"/>
    <cell end-col="0" end-row="0" start-col="0" start-row="0">
      <Coords points="411,540_411,555_742,555_742,540"/>
    </cell>
    <cell end-col="0" end-row="1" start-col="0" start-row="1">
      <Coords points="573,563_573,580_819,580_819,563"/>
    </cell>
    <cell end-col="1" end-row="1" start-col="1" start-row="1">
      <Coords points="947,580_947,593_983,593_983,580"/>
    </cell>
    <cell end-col="0" end-row="2" start-col="0" start-row="2">
      <Coords points="409,608_409,651_549,651_549,608"/>
    </cell>
    <cell end-col="1" end-row="2" start-col="1" start-row="2">
      <Coords points="942,606_942,622_974,622_974,606"/>
    </cell>
    <cell end-col="0" end-row="3" start-col="0" start-row="3">
      <Coords points="410,689_410,709_543,709_543,689"/>
    </cell>
    <cell end-col="1" end-row="3" start-col="1" start-row="3">
      <Coords points="921,691_921,708_973,708_973,691"/>
    </cell>
  </table>
</document>

```

Listing 5.1: Voorbeeld van een XML-resultaat-bestand.

Daaropvolgend wordt de XML-resultaat-bestand geïtereerd. Bij een “cell“-XML-tag wordt gekeken naar de coördinaten en dimensies (gespecificeerd door de “Coords“-XML-tag van de “cell“-tag) en op basis hiervan wordt OCR op enkel die regio toegepast. Hierdoor krijgt men de tekst te vinden binnen de celregio, gedefinieerd door de XML-resultaat. Dit proces wordt iteratief uitgevoerd op elke “cell“-tag. Door middel van de tekst en rij- en kolomwaarden (gespecificeerd door respectievelijk de “start/end-row“- en “start/end-col“-attribuut) van elke cel wordt tenslotte een Pandas Dataframe aangemaakt die de getransformeerd tabel bevat.

### Voorgesteld verbeterd algoritme

In tabelafbeeldingen waarin voldoende ruimte aanwezig is tussen de cellen (in geval van borderless tabellen) of in tabelafbeeldingen waarin de horizontale en verticale scheidingslijnen goed te onderscheiden zijn van de rest van de tabel (in geval van bordered tabellen), is algoritme A bruikbaar. Echter, zoals reeds aangegeven, worden de cellen door de tabeldetectiemodel niet altijd gedetecteerd. Ook zijn tabscheidingslijnen soms moeilijk te onderscheiden van andere lijnen, zoals de lijnen die gevormd worden door letters zoals de

'l' of 'p'. Door deze complicaties werkt algoritme A meestal niet.

Daarom worden enkele aanpassingen voorgesteld voor de tabelstructuuranalyse, die gezamenlijk als “algoritme B“ genoemd zullen worden.

Voor bordered tabellen wordt een verbetering van de lijndetectie-algoritme voorgesteld.

De verschillende stappen die genomen worden bij lijndetectie worden gedemonstreerd d.m.v. de volgende figuur 5.5.

ASAFLOW 80MG MAAGSPRES COMP BLI 168X 80MG	Dagelijks	04/09/2015	Tablet		<b>1</b>									Indicatie: Bloedverdunner Instructie: Bij begin van de maaltijd
AZILECT 1 MG TABL 28 X 1 MG	Dagelijks	04/09/2015	Tablet			<b>1</b>								Indicatie: Parkinson
FORLAX PI PHARMA 20 ZAK 10 G	Dagelijks	04/09/2015	Zakje				<b>1</b>							Indicatie: Stoelgang Instructie: Oplossen in glas water
INDERAL COMP 50X10MG	Dagelijks	04/09/2015	Tablet		<b>1</b>		<b>1</b>		<b>1</b>					Indicatie: Beven/parkinson
MIRAPEXIN 3,15 MG COMP VERLENGDE AFGIFTE 100	Dagelijks	04/09/2015	Tablet			<b>1</b>								Indicatie: Parkinson
STALEVO 100/25/200 100 TABL	Dagelijks	04/09/2015	Tablet								<b>1</b>	<b>1</b>	<b>1</b>	Indicatie: Parkinson

Figuur 5.5: Origineel geïsoleerd tabel.

In algoritme A, om de horizontale en verticale scheidingslijnen van de tabel te detecteren, wordt in een eerste stap, adaptive thresholding toegepast. Bij simpele thresholding wordt een afbeelding in kleur gesegmenteerd, meestal tot een zwart-wit-versie van de afbeelding. Indien de intensiteit van een pixel, bij simpele thresholding, een vaste threshold (drempelwaarde) overschrijdt, wordt de pixel als zwart gesegmenteerd, anders als wit. Bij adaptive thresholding is de waarde van de threshold afhankelijk van de regio waarin de pixel zich bevindt. In de volgende figuur werd adaptive thresholding toegepast op de origineel figuur 5.5.

Chronische medicatie	Frequentie	Begin	Einde	Inname/ Eenhed	Ontbijt			Middagmaal			Avondmaal			Opmeringen				
					% morgens	Voor	Tijdens	Na	Voor	Tijdens	Na	Slaap	08 u	12 u	16 u	20 u		
ASAFLOW 80MG MAAGSAPRES COMP BLI 16X 80MG	Dagelijks	04/09/2015		Tablet		1												Indicatie: Bloedverdunner Instructie: Bij begin van de maaltijd
AZILECT 1 MG TABL 28 X 1 MG	Dagelijks	04/09/2015		Tablet			1											Indicatie: Parkinson
FORLAX PI PHARMA 20 ZAK 10 G.	Dagelijks	04/09/2015		Zakje				1										Indicatie: Stoelgang Instructie: Oplossen in glas water
INDERAL COMP 50X10MG	Dagelijks	04/09/2015		Tablet		1			1			1						Indicatie: Beven/parkinson
MIRAPEXIN 3,15 MG COMP VERLENDE AFGIFTE 100	Dagelijks	04/09/2015		Tablet			1											Indicatie: Parkinson
STALEVO 100/25/200 100. TABL.	Dagelijks	04/09/2015		Tablet										1	1	1	1	Indicatie: Parkinson

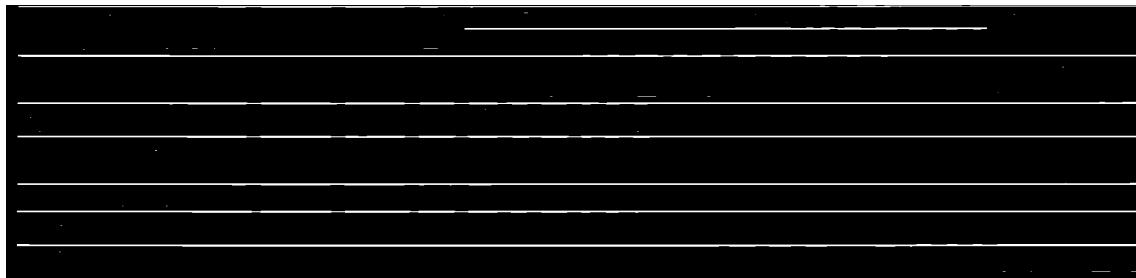
Figuur 5.6: Adaptive thresholding.

Vervolgens wordt een bitwise-not-operatie uitgevoerd. Dit heeft als gevolg dat de kleuren geïnverteerd worden.

Chronische medicatie	Frequentie	Begin	Einde	Inname/ Eenheid 's morgens	Ontbijt		Middagmaal		Avondmaal		Slapen	08 u	12 u	16 u	20 u	Opmerkingen	
					Voor	Tijdens	Na	Voor	Tijdens	Na							
ASAFLOW 80MG MAAGSAPRES COMP BLI 168X 80MG	Dagelijks	04/09/2015		Tablet		1											Indicatie: Bloedverdunner Instructie: Bij begin van de maaltijd
AZILECT 1 MG TABL 28 X 1 MG	Dagelijks	04/09/2015		Tablet			1										Indicatie: Parkinson
FORLAX PI PHARMA 20 ZAK 10 G	Dagelijks	04/09/2015		Zakje				1									Indicatie: Stoelgang Instructie: Oplossen in glas water
INDERAL COMP 50X10MG	Dagelijks	04/09/2015		Tablet		1			1		1						Indicatie: Beven/parkinson
MRAPEXIN 3,16 MG COMP VERLENGDE AFGIFTE 100	Dagelijks	04/09/2015		Tablet			1										Indicatie: Parkinson
STALEVO 100/25/200 100 TABL	Dagelijks	04/09/2015		Tablet								1	1	1	1		Indicatie: Parkinson

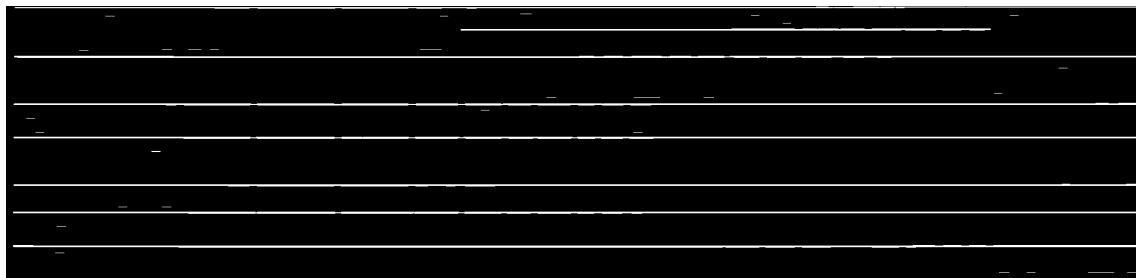
Figuur 5.7: Bitwise-not-operatie.

Hierna, om horizontale lijnen te detecteren, worden de pixels geërodeerd d.m.v. een horizontale structuur.



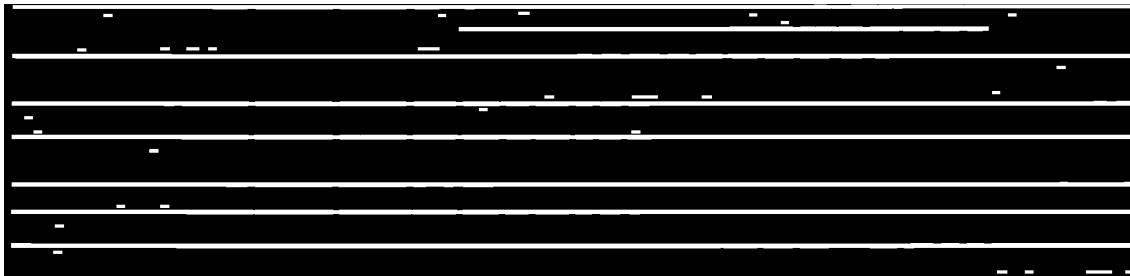
Figuur 5.8: Erosie d.m.v. een horizontale structuur.

Nadien vindt een dilatatie van de pixels plaats, d.m.v. de horizontale structuur.

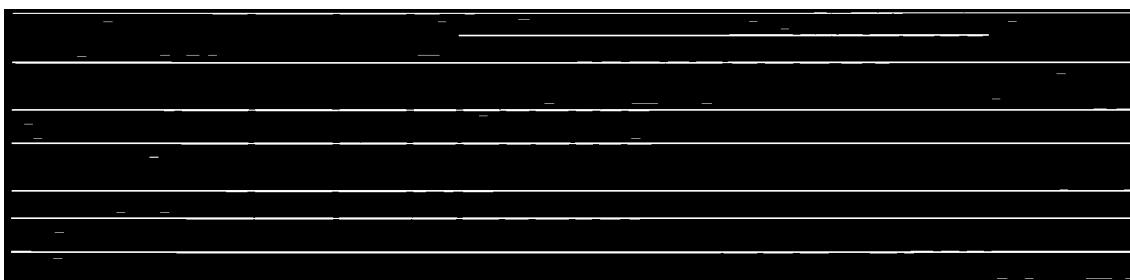


Figuur 5.9: Dilatatie d.m.v. de horizontale structuur.

Hierna wordt een simpele dilatatie uitgevoerd, gevolgd door een simpele erosie.



Figuur 5.10: Simpele dilatatie.



Figuur 5.11: Simpele erosie.

Uiteindelijk worden horizontale structuren met een te kleine lengte buiten beschouwing gebracht en wordt een Hough Line-transformatie uitgevoerd. Zo worden de horizontale lijnen verkregen. Op een analoog manier worden de verticale lijnen gedetecteerd. De gedetecteerde lijnen zijn in figuur 5.12 in rood en groen voorgesteld.

Figuur 5.12: Lijnen gedetecteerd door algoritme A.

Met algoritme B wordt een andere aanpak voor de lijndetectie voorgesteld.

Zo worden eerst de tekstblokken, gedetecteerd door OCR, gemaskeerd door witte rechthoeken om de tekstelementen binnen de tabel te verbergen. In de volgende figuur 5.13 wordt het resultaat van deze operatie weergegeven. Men kan merken dat er nog tekstelementen aanwezig zijn. Dit komt omdat de resolutie van de ingescande document laag is, waardoor de OCR-software niet alle tekstblokken heeft kunnen detecteren.

		Begin	Einde	Inname/ Eenheid	's morgens	Voor	Tijdens	Na	Voor	Tijdens	Na	Voor	Tijdens	Na	Slaap	08 u	12 u	16 u	20 u	
				Tablet		1														
				Tablet			1													
				Zakje				1												
	Dagelijks			Tablet		1			1			1								
				Tablet			1										1	1	1	1
				Tablet													1	1	1	1

Figuur 5.13: Verwijdering van tekstelementen.

Vervolgens wordt een dilatatie, gevolgd door een erosie uitgevoerd.

		Begin	Einde	Inname/ Eenheid	's morgens	Voor	Tijdens	Na	Voor	Tijdens	Na	Voor	Tijdens	Na	Slaap	08 u	12 u	16 u	20 u	
				Tablet		1														
				Tablet			1													
				Zakje				1												
	Dagelijks			Tablet		1			1			1								
				Tablet			1										1	1	1	1
				Tablet													1	1	1	1

Figuur 5.14: Dilatatie.

		Begin	Einde	Inname/ Eenheid	's morgens	Voor	Tijdens	Na	Voor	Tijdens	Na	Voor	Tijdens	Na	Slaap	08 u	12 u	16 u	20 u	
				Tablet		1														
				Tablet			1													
				Zakje				1												
	Dagelijks			Tablet		1			1			1								
				Tablet			1										1	1	1	1
				Tablet													1	1	1	1

Figuur 5.15: Erosie.

Uiteindelijk worden een Canny edge-operatie en een Hough-Line-transformatie toegepast om de horizontale en verticale tegelijk te detecteren. Deze gedetecteerde lijnen zijn in figuur 5.16 in rood en groen voorgesteld.

Figuur 5.16: Lijnen gedetecteerd door algoritme B.

Met de volgende figuren 5.12 en 5.16 wordt het resultaat van beide algoritmes opnieuw weergegeven.

Figuur 5.17: Lijnen gedetecteerd door algoritme A.

Figuur 5.18: Lijnen gedetecteerd door algoritme B.

Zoals men kan zien, heeft algoritme A last van valse positieven. Er worden namelijk teveel lijnen gedetecteerd. Dit komt enerzijds omdat het de streep van de tekstelementen "1" als verticale lijnen detecteert. Anderzijds wordt door afbeeldingruis een extra horizontaal lijn, rechts onderaan, gedetecteerd. Met algoritme B worden alle lijnen correct gedetecteerd.

Voor borderless tabellen wordt een clusteringalgoritme voorgesteld, in combinatie met lijndetectie. Bij borderless tabellen kan men een bepaalde patroon herkennen. Het verschil in X-positie tussen tekstblokken van verschillende kolommen is namelijk veel groter dan het verschil in X-positie tussen tekstblokken binnen dezelfde kolom. Overeenkomstig is het verschil in Y-positie tussen tekstblokken van verschillende rijden veel groter dan het verschil in Y-positie tussen tekstblokken binnen dezelfde rij. Dit betekent dat men cellen kan clusteren (groeperen) op basis van hun X-Y-positie in rij- en kolomclusters. Deze rij-

en kolomclusters worden in relatie gebracht met bepaalde rij- en kolomindexen. Zo kan men door clustering aan elk cel een rij- en kolomindex toekennen. Een populair clusteringalgoritme is K-means-clustering. Echter moet hiervoor het aantal clusters meegegeven worden, wat niet mogelijk is aangezien het aantal rijen en kolommen van een willekeurig tabel niet op voorhand gekend is. Daarom wordt voor Algoritme B, voor de clustering van cellen van borderless tabellen, gebruikgemaakt van hiërarchische clustering die op basis van een maximum afstand werkt.

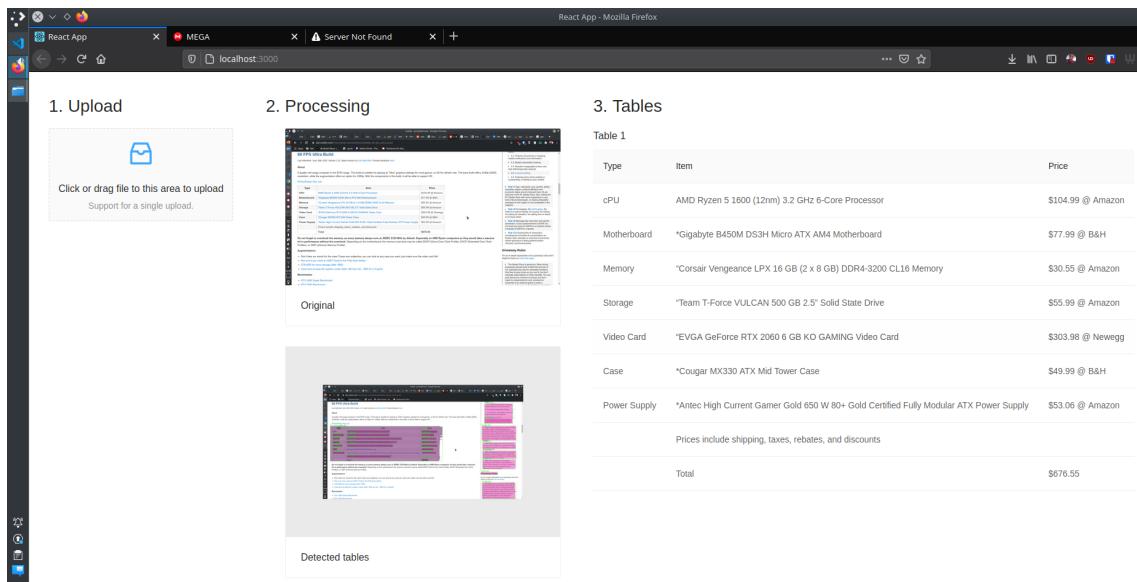
### 5.1.5 Postprocessing

Aangezien algoritme B op OCR steunt voor tabelstructuuranalyse, is een postprocessing van het resultaat noodzakelijk. Postprocessing is niet nodig indien gebruik wordt van algoritme A.

Eens een Pandas Dataframe, met daarin tekstelementen met bijhorende rij- en kolomwaarden, verkregen is d.m.v. algoritme B, vindt concatenatie plaats. Hierbij worden tekstblokken die dezelfde rij- en kolomwaarden hebben, aan elkaar geconcateneerd. Bovendien worden volledige lege rijen en kolommen verwijderd van de Pandas Dataframe.

### 5.1.6 Resultaatweergave

Na de strucuuranalyse, en postprocessing indien algoritme A gebruikt werd, wordt de Pandas Dataframe omgezet in een JSON-object die teruggestuurd wordt van de back end naar de GUI van de gebruiker. De gebruiker krijgt dan het resultaat van de tabeltransformatie te zien.



Figuur 5.19: GUI van de software wanneer de tabeltransformatie uitgevoerd is.

## 5.2 Praktisch gebruik

### 5.2.1 Hardware- en software-vereisten

Wat hardware betreft, is een grafische kaart van het merk NVIDIA nodig. Dit komt omdat, voor de tabeldetectie, Prasad e.a. (2020) gebruik maakt van een verouderd versie van MMDetection (Chen e.a., 2019), een objectdetectie-library. Bij de meest recente versie van MMDetection is objectdetectie d.m.v. enkel de CPU mogelijk.

Wat software betreft is een Linux-distributie of MacOS nodig voor de tabeldetectie, aangezien MMDetection niet ondersteund wordt op het besturingssysteem Windows. Indien de tabellen reeds geïsoleerd zijn en dus enkel tabelstructuuranalyse nodig is, dan is tabeldetectie niet nodig en kan de software eveneens op Windows gebruikt worden.

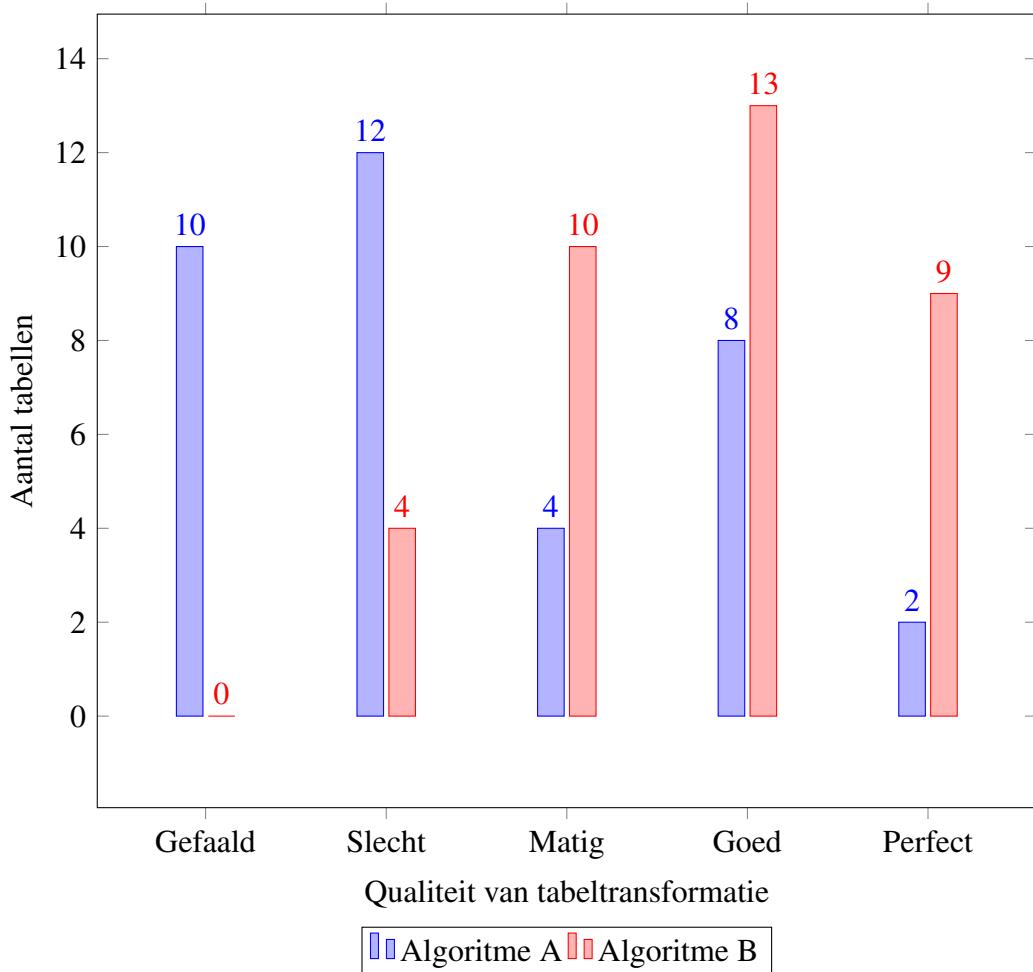
### 5.2.2 Installatie en gebruik

De open source broncode, inclusief gedetailleerde installatie- en gebruiksinstructies zijn te vinden in de repository Nazari (2020).



## **6. Resultaten**

Kijkend naar de resultaten van de tabeltransformatie van de dertig willekeurige documenten, die in appendix-deel A bekijken kunnen worden, kan men enkele waarnemingen maken. Zo ziet men dat OCR niet altijd nauwkeurig werkt, zeker niet wanneer de resolutie van de afbeelding laag is. Verder kan men, door de grafiek 6 te bestuderen, merken dat de nauwkeurigheid van de tabeltransformatie afhankelijk is van de algoritme die gebruikt wordt voor tabelstructuuranalyse.



Bij gebruik van de voorgestelde algoritme B is de slaagkans op een succesvolle tabeltransformatie merkbaar groter dan bij gebruik van algoritme A. Verder kan men zien dat, hoewel een groot aantal goede en perfecte tabeltransformaties verkregen zijn, de hoeveelheid slechte en matige transformaties niet onbeduidend is.

## 7. Conclusie

Uit dit onderzoek kan men enkele conclusies trekken. Zo kan men, uit de literatuurstudie (hoofdstuk 3) concluderen dat tabeltransformatie een complex domein is. Kant en klaar software-pakketten voor tabeltransformatie bestaan, echter zijn deze niet open source en betalend. Een open source versie bestaat momenteel niet.

Verder kan men besluiten dat tabeltransformatie niet een simpel eenvoudig proces is, maar een complex procedure die uit verschillende subprocessen bestaat. Zo vindt preprocessing eerst plaats. Hierna wordt met tabeldetectie de tabellen van de rest van het document geïsoleerd. Vervolgens vindt de transformatie plaats d.m.v. structuuranalyse en OCR. Hierna wordt het resultaat verder behandeld door postprocessing. Uiteindelijk wordt de getransformeerd tabel terug naar de gebruiker gestuurd.

Bovendien kan men de gevolgtrekking maken dat, afhankelijk van de gebruikte algoritmes, pre- en postprocessing ofwel noodzakelijk ofwel niet nodig zijn. Voor de proof-of-concept, bij gebruik van de voorgestelde algoritme, is postprocessing bijvoorbeeld noodzakelijk. Postprocessing bleek in het algemeen nodig te zijn voor de proof-of-concept, om nauwkeurig tabellen te kunnen detecteren.

Tenslotte kan men concluderen dat tabeldetectie enerzijds zeer nauwkeurig is, terwijl anderzijds structuuranalyse minder optimale resultaten kan leveren. De algoritme voorgesteld in dit onderzoek, die een niet onbelangrijke verbetering van de tabeltransformaties heeft teweeggebracht, toont aan dat optimalisatiemogelijkheden zeker nog mogelijk zijn.



## **A. Details van resultaten**

## A.1 Document 1

UNITED NATIONS E-GOVERNMENT SURVEY 2014

A channel can change a user's perception and confidence of a public service. For instance, in the context of value proposition of a public service, a channel can add value with a positive user experience, and in the same vein, it can subtract value through a negative user experience. Therefore, it is extremely important to select the right channel for the right service targeting the specific audience. At the same time, channel selection is a deciding factor to effectively reach out to specific groups of citizens, for example, rural population with limited ICT access.

Table 5.1. List of channels (non-exhaustive)

1. Counter (face-to-face) service
2. Telephone (voice) service and call centres
3. Web portal
4. Email
5. SMS and other messaging services
6. Mobile portal (mobile website)
7. Mobile app
8. Social media
9. Public kiosks
10. Intermediaries through public-private partnership

Digital channels, with both their diversity and spread, are increasingly embraced by almost all countries, while counter (face-to-face service) and telephone (voice) services, have continued to serve as fundamental channels as preferred by some citizen groups. Through strategic and optimised mixed use of channels, governments will be able to provide ubiquitous 24 x 7 access to information and services to different user groups.

### 5.2.1. Web portal

The 2014 Survey concludes that all 193 United Nations Member States have some form of online presence, as compared to 18 countries with no online presence in 2003 and 3 countries in 2012 (see Figure 2.1 in Chapter 2).

Research shows that offering more online services and aiming for increased usage of these services improves efficiency and results in cost reduction. According to one research study,<sup>10</sup> the Government of the United Kingdom could save between GBP 3.30 and GBP 12 per transaction by moving public services online. Denmark has made mandatory the use of online public services in its e-government strategy and action plan covering the period 2011–2015.<sup>11</sup> When fully implemented, it will go "digital by default" for citizens and businesses with the aim of making public service delivery more cost-efficient. Similarly, the Government Digital Strategy<sup>12</sup> of the United Kingdom published in November 2012, stresses that all services should be "digital by default". The strategy states 11 principles and 14 actions to shape how central government departments and agencies will embrace digitalization of their services and improve usage by citizens and businesses. The Government of the United Kingdom has since committed to the

97

Figuur A.1: Origineel document.

## UNITED NATIONS E-GOVERNMENT SURVEY 2014

A channel can change a user's perception and confidence of a public service. For instance, in the context of value proposition of a public service, a channel can add value with a positive user experience, and in the same vein, it can subtract value through a negative user experience. Therefore, it is extremely important to select the right channel for the right service targeting the specific audience. At the same time, channel selection is a deciding factor to effectively reach out to specific groups of citizens, for example, rural population with limited ICT access.

seeID 756

**Table 5.1** List of channels (non-exhaustive)

1. Counter (face-to-face service)
2. Telephone (voice) service and call centres
3. Web portal
4. Email
5. SMS and other messaging service
6. Mobile portal (mobile website)
7. Mobile app
8. Social media
9. Public kiosk
10. Intermediaries through public-private partnership

Digital channels, with both their diversity and spread, are increasingly embraced by almost all countries, while counter (face-to-face service) and telephone (voice) services, have continued to serve as fundamental channels as preferred by some citizen groups. Through strategic and optimised mixed use of channels, governments will be able to provide ubiquitous 24 x 7 access to information and services to different user groups.

### 5.2.1. Web portal

The 2014 Survey concludes that all 193 United Nations Member States have some form of online presence, as compared to 18 countries with no online presence in 2003 and 3 countries in 2012 (see Figure 2.1 in Chapter 2).

Research shows that offering more online services and aiming for increased usage of these services improves efficiency and results in cost reduction. According to one research study,<sup>10</sup> the Government of the United Kingdom could save between GBP 3.30 and GBP 12 per transaction by moving public services online. Denmark has made mandatory the use of online public services in its e-government strategy and action plan covering the period 2011–2015.<sup>11</sup> When fully implemented, it will go "digital by default" for citizens and businesses with the aim of making public service delivery more cost-efficient. Similarly, the Government Digital Strategy<sup>12</sup> of the United Kingdom published in November 2012, stresses that all services should be "digital by default". The strategy states 11 principles and 14 actions to shape how central government departments and agencies will embrace digitalization of their services and improve usage by citizens and businesses. The Government of the United Kingdom has since committed to the

97

Figuur A.2: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

---

**Counter (face-to-face) service**

---

- 0 Telephone (voice) service and call centres
  - 1 Web portal
  - 2 Email
  - 3 SMS and other messaging service:
  - 4 Mobile portal (mobile website)
  - 5 7. Mobile app
  - 6 Social media
  - 7 Public kiosks
  - 8 Intermediaries through public-private partnershir
- 

Getransformeerde tabellen met algoritme B:

---

**1. Counter (face-to-face) service**

---

- 0 2. Telephone (voice) service and call centres
  - 1 3. Web portal
  - 2 4. Email
  - 3 5. SMS and other messaging services
  - 4 6. Mobile portal (mobile website)
  - 5 7. Mobile app
  - 6 8. Social media
  - 7 9. Public kiosks
  - 8 10. Intermediaries through public-private part...
-

## A.2 Document 2

DRAFT BUDGETARY PLAN 2019

on the value of 3.539e-007, the mean of the drift has been set at 0.0175 and its standard deviation at 0.004591, the mean of the slope at 0.72 and its standard deviation at 0.0781. For what concerns the second equation, the mean and variance of the variance of the error component have been revised to 0.00555, the mean of the coefficient to at 1.4 and its standard deviation to 0.705<sup>7</sup>.

**TABLE IV.2-1 INITIAL PARAMETERS FOR THE NAWRU ESTIMATE**

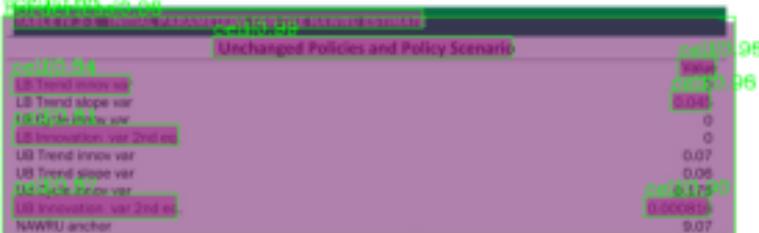
Unchanged Policies and Policy Scenario	Value
LB Trend innov var	0
LB Trend slope var	0.045
LB Cycle innov var	0
LB Innovation var 2nd eq.	0
UB Trend innov var	0.07
UB Trend slope var	0.06
UB Cycle innov var	0.175
UB Innovation var 2nd eq.	0.000816
NAWRU anchor	9.07

### IV.3 METHODOLOGICAL NOTE ON THE CRITERIA FOR FORMULATING MACROECONOMIC AND BUDGETARY PROJECTIONS

See the document "Nota metodologica sui criteri di formulazione delle previsioni tendenziali" (in Italian only)

<sup>7</sup> For details on the methodology and the parameters employed, see the document "Nota metodologica sui criteri di formulazione delle previsioni tendenziali" (in Italian only).

Figuur A.3: Origineel document.

DRAFT BUDGETARY PLAN 2019	
<b>on the value of 3.539e-007, the mean of the drift has been set at 0.0175 and its standard deviation at 0.004591, the mean of the slope at 0.72 and its standard deviation at 0.0781. For what concerns the second equation, the mean and variance of the variance of the error component have been revised to 0.00555, the mean of the coefficient to at 1.4 and its standard deviation to 0.705<sup>7</sup>.</b>	
	
<b>IV.3 METHODOLOGICAL NOTE ON THE CRITERIA FOR FORMULATING MACROECONOMIC AND BUDGETARY PROJECTIONS</b>	
See the document "Nota metodologica sui criteri di formulazione delle previsioni tendenziali" (in Italian only)	

<sup>7</sup> For details on the methodology and the parameters employed, see the document "Nota metodologica sui criteri di formulazione delle previsioni tendenziali" (in Italian only).

Figuur A.4: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

	ate haath alee eid	NaN
0	Unchanged Policies and Policy Scenario	None
1	None	Vane
2	L8 Trend innov var	90
3	LB Trend slope var	0.045
4	LB Cycle inno var	0
5	LB Innovation var 2nd ea,	0
6	UB Trend Innov var	0.07
7	UB Trend stope var	0.08
8	UB Cycle innov var	0.178
9	UB Innovation var 2nd ea,	0.000816
10	NAWRU anchor	'9.07

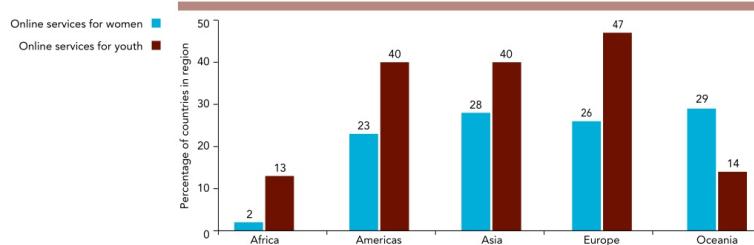
### A.3 Document 3

#### Chapter 6

##### CHAPTER 6 • BRIDGING THE DIGITAL DIVIDE

Despite the fact that young people are far more likely to be online, as compared to women as a group as well as the overall population, national government online services geared specifically towards the youth are still taking off. Disparities abound across all regions and are likely to be linked to the policy, level of development and national income of a country. In the more advanced regions, such as Europe and the Americas, 47 per cent and 40 per cent of the countries provide specific services geared to the young on their national websites (see Figure 6.13).

Figure 6.13. Countries providing online services for women and youth, by region



Among upper middle income countries Belize, China, Islamic Republic of Iran, Kazakhstan are among those providing specific services for women while Argentina, Belize, Colombia and Cuba are among those providing services for youth. Around a quarter of the countries of Europe and the Americas had specific information aimed at these groups on their websites.

It is also notable that a large number of countries are now providing more archival information related to government policies and programs (see Table 6.5). In 2014 more than 50 per cent of the countries in the world regions, except for Africa, provided some data pertaining to disadvantaged and vulnerable groups. Despite not keeping up with the world averages in general, in terms of data, Africa was behind other regions with 28 per cent of the countries providing data pertaining to disadvantaged and vulnerable groups. Among these were Egypt, Gabon, Kenya, Madagascar, Mauritania, Mauritius, Morocco, Niger, Nigeria, Rwanda and Zimbabwe.

Table 6.5. Online archived information and data for disadvantaged and vulnerable groups

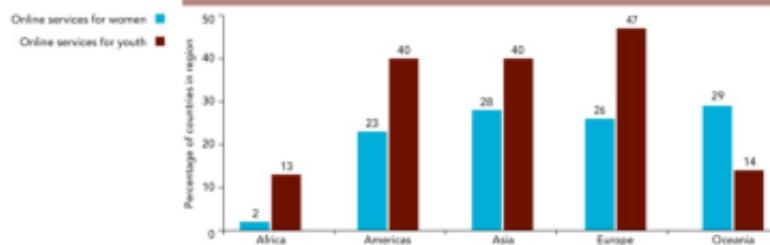
Region	Archived sources of information		Data % of countries
	% of countries	% of countries	
Africa	41	28	
Americas	69	69	
Asia	68	51	
Europe	86	60	
Oceania	57	57	

Figuur A.5: Origineel document.

## CHAPTER 6 • BRIDGING THE DIGITAL DIVIDE

Despite the fact that young people are far more likely to be online, as compared to women as a group as well as the overall population, national government online services geared specifically towards the youth are still taking off. Disparities abound across all regions and are likely to be linked to the policy, level of development and national income of a country. In the more advanced regions, such as Europe and the Americas, 47 per cent and 40 per cent of the countries provide specific services geared to the young on their national websites (see Figure 6.13).

Figure 6.13. Countries providing online services for women and youth, by region



Among upper middle income countries Belize, China, Islamic Republic of Iran, Kazakhstan are among those providing specific services for women while Argentina, Belize, Colombia and Cuba are among those providing services for youth. Around a quarter of the countries of Europe and the Americas had specific information aimed at these groups on their websites.

It is also notable that a large number of countries are now providing more archival information related to government policies and programs (see Table 6.5). In 2014 more than 50 per cent of the countries in the world regions, except for Africa, provided some data pertaining to disadvantaged and vulnerable groups. Despite not keeping up with the world averages in general, in terms of data, Africa was behind other regions with 28 per cent of the countries providing data pertaining to disadvantaged and vulnerable groups. Among these were Egypt, Gabon, Kenya, Madagascar, Mauritania, Mauritius, Morocco, Niger, Nigeria, Rwanda and Zimbabwe.

Table 6.5. Online archived information and data for disadvantaged and vulnerable groups

Region	Archived sources of information		Data
	% of countries	# of countries	
Africa	41	38	28
Americas	69	63	69
Asia	68	51	51
Europe	86	60	60
Oceania	57	62	57

138

Figuur A.6: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Region	Archived sources of information	Data
0 % of countries	% of countries	None
1 Africa	Al	28
2 Americas	69	69
3 Asia	68	51
4 Europe	Bb	60
5 Oceania	57	57

Getransformeerde tabellen met algoritme B:

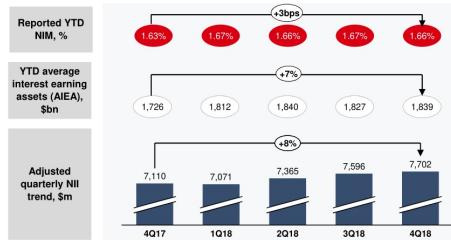
Region	Archived sources of information	Data
0 None	% of countries	% of countries
1 Africa	A	28
2 Americas	69	69
3 Asia	68	51
4 Europe	86	60
5 Oceania	57	a7

## A.4 Document 4

FY18 financial performance

**FY18 NIM of 1.66% up 3bps vs. FY17**

**Net interest margin analysis**



**Net interest margin by key legal entity\*, %**

	FY17	9M18	FY18	NII contribution to Group	AIEA contribution to Group
The Hongkong and Shanghai Banking Corporation (HBAP)	1.88%	2.05%	2.06%	53%	43%
HSBC Bank plc (NRFB) + HSBC UK Bank plc (RFB)	1.35%	1.19%	1.16%	27%	38%
HSBC Bank plc (NRFB)	n/a	0.46%	0.37%	5%	24%
HSBC UK Bank plc (RFB)*	n/a	2.15%	2.16%	21%	16%
HSBC Bank USA	0.98%	1.07%	1.08%	8%	12%

\* Further analysis can be found in the HSBC Holdings plc Annual Report and Accounts 2018 and 20-F; the HSBC UK Bank plc (RFB) started operations on 1st July 2018

**FY18 NII of \$30.5bn up \$2.3bn or 8% vs. FY17**

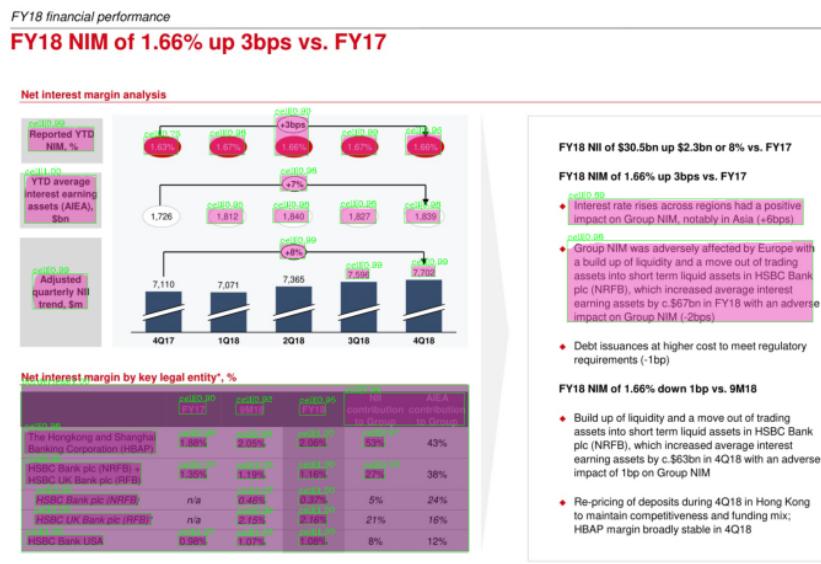
**FY18 NIM of 1.66% up 3bps vs. FY17**

- ◆ Interest rate rises across regions had a positive impact on Group NIM, notably in Asia (+6bps)
- ◆ Group NIM was adversely affected by Europe with a build up of liquidity and a move out of trading assets into short term liquid assets in HSBC Bank plc (NRFB), which increased average interest earning assets by c.\$67bn in FY18 with an adverse impact on Group NIM (-2bps)

**FY18 NIM of 1.66% down 1bp vs. 9M18**

- ◆ Build up of liquidity and a move out of trading assets into short term liquid assets in HSBC Bank plc (NRFB), which increased average interest earning assets by c.\$63bn in 4Q18 with an adverse impact of 1bp on Group NIM
- ◆ Re-pricing of deposits during 4Q18 in Hong Kong to maintain competitiveness and funding mix; HBAP margin broadly stable in 4Q18

Figuur A.7: Origineel document.



Figuur A.8: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gedaald.

Getransformeerde tabellen met algoritme B:

The Hongkong and Shanghai		1.88%	2.05%	2.06%	53%	43%
0	Banking Corporation (HBAP)	None	None	None	None	None
1	HSBC Bank ple (NRFB) + HSBC UK Bank ple (RFB)	1.35%	1.19%	1.16%	27%	38%
2	HSBC Bank plc (NRFB)	Wa	0.46%	0.37%	5%	24%
3	HSBC UK Bank plo (RFB)*	na	2.15%	2.16%	21%	16%
4	HSBC Bank USA	0.98%	1.07%	1.08%	8%	12%
5	None	FY17	ey)	aly	Pele	mr hel)
6	None	None	None	None	ii}	rN
7	None	None	None	None	eect)	feeds)

## A.5 Document 5

### Annexes

#### ANNEX

##### Regional and economic groupings

	HCI	Adult literacy (%)	Gross enrolment ratio (%)	Expected years of schooling	Mean years of schooling
Africa	0.4492	64.65	59.74	9.87	4.65
Americas	0.7202	90.83	79.85	13.29	8.33
Asia	0.6615	86.79	74.08	12.38	7.58
Europe	0.8434	98.63	88.54	15.25	10.67
Oceania	0.7073	91.20	75.15	12.63	8.94
World	0.6566	84.29	74.41	12.50	7.68

	HCI	Adult literacy (%)	Gross enrolment ratio (%)	Expected years of schooling	Mean years of schooling
Small island developing States	0.6654	87.51	74.29	12.26	7.72
Land Locked Developing Countries	0.5368	73.86	64.92	10.74	6.14
Least Developed Countries	0.3960	59.98	56.26	9.04	3.87

	HCI	Adult literacy (%)	Gross enrolment ratio (%)	Expected years of schooling	Mean years of schooling
High Income	0.8343	97.58	88.87	15.30	10.25
Upper Middle Income	0.7253	92.17	78.54	13.29	8.58
Lower Middle Income	0.5787	80.19	66.84	11.08	6.56
Low Income	0.3884	57.28	56.69	9.06	3.95

Figuur A.9: Origineel document.

## Annexes

## ANNEX

Regional and economic groupings					
	HDI (0)	Adult literacy (%)	Gross enrollment ratio (%)	Expected years of schooling	Mean years of schooling
Africa	0.449	84.65	59.74	9.87	4.65
Americas	0.720	90.83	79.85	13.29	8.33
Asia	0.615	86.79	74.08	12.38	7.58
Europe	0.841	98.63	88.54	15.25	10.67
Oceania	0.807	91.20	75.15	12.63	8.94
World	0.606	84.95	74.41	12.50	7.68

Borderless(0.99)					
	HDI (0)	Adult literacy (%)	Gross enrollment ratio (%)	Expected years of schooling	Mean years of schooling
Small Least Developed States	0.345	87.51	74.29	12.26	7.72
Least Developed Countries	0.334	73.86	64.92	10.74	6.14
Least Developed Countries	0.396	59.98	56.24	9.04	3.87

Borderless(1.05)					
	HDI (0)	Adult literacy (%)	Gross enrollment ratio (%)	Expected years of schooling	Mean years of schooling
High Income	0.841	97.41	88.87	15.30	10.25
Upper Middle Income	0.704	92.17	78.54	13.29	8.58
Lower Middle Income	0.570	80.19	54.84	11.08	6.56
Low Income	0.389	57.28	56.69	9.06	3.95

236

Figuur A.10: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Adult\nliteracy (%)	Gross\lenrolmen\lratio (%)	Expected years of\nschooling	Mean years\nof s
0 Africa	59.74		987	465
1 Americas	79.85		13.29	8.33
2 06615	74.08		12.38	7.58
3 Europe	88.54		15.25	10.67
4 Oceania	75.15		12.63	894
5 World	74.41		12.50	7.68

	Adult\nliteracy (%)	Gross\lenrolment\nratio (%)	Expected years of\nschooling	Mean years\ln
0	High Income	88.87	15.30	10.25
1	Upper Middle Income	78.54	13.29	8.58
2	Lower Middle Income	66.84	11.08	6.56
3	Low Income	56.69	9.06	3.95

	Adult\nliteracy (%)	enrolment\lratio (%)	Expected years of\lnschooling	Mean ye
0	Small island developing States	74.29	12.26	7.72
1	Land Locked Developing Countries	64.92	10.74	6.14
2	Least Developed Countries	56.26	9.04	3.95

Getransformeerde tabellen met algoritme B:

	Gross	NaN	NaN	NaN	NaN	NaN
0	enrolment	Adult	Expected yearsof	Mean years	None	None
1	__ratio (%)	literacy (%)	schooling	of schooling	HCI	None
2	59.74	64.65	9.87	4.65	0.4492	Africa
3	79.85	90.83	13.29	8.33	0.7202	Americas
4	74.08	86.79	12.38	7.58	0.6615	Asia
5	88.54	98.63	18.25	10.67	0.8434	Europe
6	75.15	91.20	12.63	8.94	0.7073	Oceania
7	None	84.29	12.50	7.68	0.6566	World

	Gross	NaN	NaN	NaN	NaN	NaN
0	Adult enrolment	Expected yearsof	Mean years	None	None	None
1	literacy (%) _ ratio (%)	schooling	of schooling	None	None	None
2	97.58 88.87	15.30	10.25	High Income	0.8343	None
3	92.17 78.54	13.29	8.58	Upper Middle	0.7253	Income
4	80.19 66.84	11.08	6.56	Lower Middle	0.5787	Income
5	57.28 56.69	9.06	3.95	Low Income	0.3884	None

	Adult	enrolment	Expected yearsaf © Mean years	NaN	NaN
0	literacy (%)	_ratio (%)	schooling of schooling	Het	None
1	87.51	74.29	12.26 7.72	0.6654	Small island developing States
2	73.86	64.92	10.74 6.14	0.5368	Land Locked Developing Countries
3	59.98	56.26	9.04 3.87	0.3960	Least Developed Countries

## A.6 Document 6

### Federal Communications Commission

**NG59** The frequencies 37.60 and 37.85 MHz may be authorized only for use by base, mobile, and operational fixed stations participating in an interconnected or coordinated power service utility system.

**NG60** In the band 31-31.3 GHz, for stations in the fixed service authorized after August 6, 2018, the unwanted emissions power in any 100 MHz of the 31.3-31.5 GHz Earth exploration-satellite service (passive) band shall be limited to -38 dBW (-38 dBW/100 MHz), as measured at the input to the antenna.

**NG62** In the bands 28.5-29.1 GHz and 29.25-29.5 GHz, stations in the fixed-satellite service shall not cause harmful interference to, or claim protection from, stations in the fixed service operating under the following call signs: KEB35, KGB72, KGC79, KIL20, KME49, KQG58, KQH74, KSA96, KSE73, KVH83, KYJ33, KZS88, WAX78, WLT380, WMK817, WML443, WMP367, and WSL69.

**NG63** In the band 37.5-40 GHz, earth station operations in the fixed-satellite service (space-to-Earth) shall not claim protection from stations in the fixed and mobile services, except where individually licensed earth stations are authorized pursuant to 47 CFR § 25.136.

**NG65** In the bands 24.75-25.25 GHz and 47.2-48.2 GHz, stations in the fixed and mobile services may not claim protection from individually licensed earth stations authorized pursuant to 47 CFR 25.136. However, nothing in this footnote shall limit the right of UMFUS licensees to operate in conformance with the technical rules contained in 47 CFR part 30. The Commission reserves the right to monitor developments and to undertake further action concerning interference between UMFUS and FSS, including aggregate interference to satellite receivers, if appropriate.

**NG66** The band 470-512 MHz (TV channels 14-20) is allocated to the broadcasting service on an exclusive basis throughout the United States and its insular areas, except as described below:

(a) In the urbanized areas listed in the table below, the indicated frequency bands are allocated to the land mobile service on an exclusive basis for assignment to eligibles in the Public Mobile Services, the Public Safety Radio Pool, and the Industrial/Business Radio Pool, except that:

(1) Licensees in the land mobile service that are regulated as Commercial Mobile Radio Service (CMRS) providers may also use their assigned spectrum to provide fixed service on a primary basis.

(2) The use of the band 482-488 MHz (TV channel 16) is limited to eligibles in the Public Safety Radio Pool in or near (i) the Los Angeles urbanized area; and (ii) New York City; Nassau, Suffolk, and Westchester Counties in New York State; and Bergen County, NJ.

Urbanized area	Bands (MHz)	TV channels
Boston, MA.....	470-476, 482-488.....	14, 16
Chicago, IL-Northwestern IN.....	470-476, 476-482.....	14, 15
Cleveland, OH.....	470-476, 476-482.....	14, 15
Dallas-Fort Worth, TX.....	482-488.....	16
Detroit, MI.....	476-482, 482-488.....	15, 16
Houston, TX.....	488-494.....	17
Los Angeles, CA.....	470-476, 482-488, 506-512.....	14, 16, 20
Miami, FL.....	470-476.....	14
New York, NY-Northeastern NJ.....	470-476, 476-482, 482-488.....	14, 15, 16
Philadelphia, PA-NJ.....	500-506, 506-512.....	19, 20
Pittsburgh, PA.....	470-476, 494-500.....	14, 18
San Francisco-Oakland, CA.....	482-488, 488-494.....	16, 17
Washington, DC-MD-VA.....	488-494, 494-500.....	17, 18

(b) In the Gulf of Mexico offshore from the Louisiana-Texas coast, the band 476-494 MHz (TV channels 15-17) is allocated to the fixed and mobile services on a primary basis for assignment to eligibles in the Public Mobile and Private Land Mobile Radio Services.

Figuur A.11: Origineel document.

Federal Communications Commission																																												
<b>NG59</b> The frequencies 37.60 and 37.85 MHz may be authorized only for use by base, mobile, and operational fixed stations participating in an interconnected or coordinated power service utility system.																																												
<b>NG60</b>	In the band 31.3-31.5 GHz, for stations in the fixed service authorized after August 6, 2018, the unwanted emissions power in any 100 MHz of the 31.3-31.5 GHz Earth exploration-satellite service (passive) band shall be limited to -38 dBW (-38 dBW/100 MHz), as measured at the input to the antenna.																																											
<b>NG62</b>	In the bands 28.5-29.1 GHz and 29.25-29.5 GHz, stations in the fixed-satellite service shall not cause harmful interference to, or claim protection from, stations in the fixed service operating under the following call signs: KEB35, KGBT2, KGC79, KIL20, KME49, KQQ58, KQH74, KSA96, KSE73, KVH83, KYJ33, KZ588, WAX78, WLT380, WMK817, WML443, WMP367, and WSL69.																																											
<b>NG63</b>	In the band 37.5-40 GHz, earth station operations in the fixed-satellite service (space-to-Earth) shall not claim protection from stations in the fixed and mobile services, except where individually licensed earth stations are authorized pursuant to 47 CFR § 25.136.																																											
<b>NG65</b>	In the bands 24.75-25.25 GHz and 47.2-48.2 GHz, stations in the fixed and mobile services may not claim protection from individually licensed earth stations authorized pursuant to 47 CFR 25.136. However, nothing in this footnote shall limit the right of UMFUS licensees to operate in conformance with the technical rules contained in 47 CFR part 30. The Commission reserves the right to monitor developments and to undertake further action concerning interference between UMFUS and FSS, including aggregate interference to satellite receivers, if appropriate.																																											
<b>NG66</b>	The band 470-512 MHz (TV channels 14-20) is allocated to the broadcasting service on an exclusive basis throughout the United States and its insular areas, except as described below:																																											
(a)	In the urbanized areas listed in the table below, the indicated frequency bands are allocated to the land mobile service on an exclusive basis for assignment to eligibles in the Public Mobile Services, the Public Safety Radio Pool, and the Industrial/Business Radio Pool, except that:																																											
(1)	Licenses in the land mobile service that are regulated as Commercial Mobile Radio Service (CMRS) providers may also use their assigned spectrum to provide fixed service on a primary basis.																																											
(2)	The use of the band 482-488 MHz (TV channel 16) is limited to eligibles in the Public Safety Radio Pool in or near (i) the Los Angeles urbanized area; and (ii) New York City; Nassau, Suffolk, and Westchester Counties in New York State; and Bergen County, NJ.																																											
<table border="1"> <thead> <tr> <th>Urbanized area</th> <th>Allocated bands</th> <th>TV channels</th> </tr> </thead> <tbody> <tr> <td>Boston, MA.....</td> <td>470-476, 482-488</td> <td>14, 16</td> </tr> <tr> <td>Chicago, IL-Northwestern IN.....</td> <td>470-476, 476-482</td> <td>14, 15</td> </tr> <tr> <td>Cleveland, OH.....</td> <td>470-476, 476-482</td> <td>14, 15</td> </tr> <tr> <td>Dallas-Fort Worth, TX.....</td> <td>470-488, 482</td> <td>16</td> </tr> <tr> <td>Detroit, MI.....</td> <td>476-482, 482-488</td> <td>15, 16</td> </tr> <tr> <td>Houston, TX.....</td> <td>482-494, 502</td> <td>17, 18, 84</td> </tr> <tr> <td>Los Angeles, CA.....</td> <td>470-476, 482-488, 506-512</td> <td>14, 16, 20</td> </tr> <tr> <td>Miami, FL.....</td> <td>470-476</td> <td>14, 16, 20</td> </tr> <tr> <td>New YorC, NY-Northeastern NJ.....</td> <td>470-476, 476-482, 482-488</td> <td>16, 17, 18, 1</td> </tr> <tr> <td>Philadelphia, PA-NJ.....</td> <td>506-508, 506-512</td> <td>16, 17, 71</td> </tr> <tr> <td>Pittsburgh, PA.....</td> <td>470-476, 494-500</td> <td>14, 18</td> </tr> <tr> <td>San Francisco-Oakland, CA.....</td> <td>482-488, 488-494</td> <td>17, 18</td> </tr> <tr> <td>Washington, DC-MD-VA.....</td> <td>488-494, 494-500</td> <td>17, 18</td> </tr> </tbody> </table>			Urbanized area	Allocated bands	TV channels	Boston, MA.....	470-476, 482-488	14, 16	Chicago, IL-Northwestern IN.....	470-476, 476-482	14, 15	Cleveland, OH.....	470-476, 476-482	14, 15	Dallas-Fort Worth, TX.....	470-488, 482	16	Detroit, MI.....	476-482, 482-488	15, 16	Houston, TX.....	482-494, 502	17, 18, 84	Los Angeles, CA.....	470-476, 482-488, 506-512	14, 16, 20	Miami, FL.....	470-476	14, 16, 20	New YorC, NY-Northeastern NJ.....	470-476, 476-482, 482-488	16, 17, 18, 1	Philadelphia, PA-NJ.....	506-508, 506-512	16, 17, 71	Pittsburgh, PA.....	470-476, 494-500	14, 18	San Francisco-Oakland, CA.....	482-488, 488-494	17, 18	Washington, DC-MD-VA.....	488-494, 494-500	17, 18
Urbanized area	Allocated bands	TV channels																																										
Boston, MA.....	470-476, 482-488	14, 16																																										
Chicago, IL-Northwestern IN.....	470-476, 476-482	14, 15																																										
Cleveland, OH.....	470-476, 476-482	14, 15																																										
Dallas-Fort Worth, TX.....	470-488, 482	16																																										
Detroit, MI.....	476-482, 482-488	15, 16																																										
Houston, TX.....	482-494, 502	17, 18, 84																																										
Los Angeles, CA.....	470-476, 482-488, 506-512	14, 16, 20																																										
Miami, FL.....	470-476	14, 16, 20																																										
New YorC, NY-Northeastern NJ.....	470-476, 476-482, 482-488	16, 17, 18, 1																																										
Philadelphia, PA-NJ.....	506-508, 506-512	16, 17, 71																																										
Pittsburgh, PA.....	470-476, 494-500	14, 18																																										
San Francisco-Oakland, CA.....	482-488, 488-494	17, 18																																										
Washington, DC-MD-VA.....	488-494, 494-500	17, 18																																										
(b)	In the Gulf of Mexico offshore from the Louisiana-Texas coast, the band 476-494 MHz (TV channels 15-17) is allocated to the fixed and mobile services on a primary basis for assignment to eligibles in the Public Mobile and Private Land Mobile Radio Services.																																											

170

Figuur A.12: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

‘Urbanized area	TV channels
0 Dallas-Fort Worth, TX.\n\nDetroit, MI\n\nHoust...	16\n\n15, 16\n7\n\n14, 16.\n14\n\n14,15\n19, 2...

Getransformeerde tabellen met algoritme B:

	Urbanized area	Bands (MH)	TV char
0	Boston, MA...	AT0-AT6, 482-488,	14,16
1	Chicago, IL-Northwestern IN.	410-476, 476-482	141s
2	Cleveland, OH...	470-476, 476-482...	14,15,
3	Dallas-Fort Worth, TX.	482-488,	16
4	Detroit, MI	476-482, 482-488.	15,16
5	Houston, TX.	488-494,	17
6	Los Angeles, CA.	470-476, 482-488, 506-512.	14, 16, 2
7	Miami. EL..... New York, NY-1 Northeastem NI	AT0AT6, AT0AT6, 476-482, 482-488.	14 14, 1
8	Philadelphia, PA-NJ.	500-506, 506-512.	19, 20
9	Pittsburgh, PA.	470-476, 494-500,	14.18
10	San Francisco-Oakland, CA.	482-488, 488-494.	16,17,
11	Washington, DC-MD-VA.	488-494, 494-500.	17,18,

## A.7 Document 7

### INDUSTRY OVERVIEW

The IDH1m inhibitor market is primarily driven by the number of addressable patients with AML, cholangiocarcinoma and glioma. IDH1 mutations are observed in approximately 5.5% of AML patients, 8% of cholangiocarcinoma patients and 77% of glioma patients. The number of IDH1m inhibitor addressable patients in China has been growing, and is expected to grow steadily as shown in the table below.

**IDH1m Inhibitor Market in China**

Indication	Historical			Predictive		Addressable Patients (2017)	
	2013	2017	2022E	CAGR (2013-2017)	CAGR (2017-2030E)		
AML	17,400	18,800	20,100	22,500	1.9%	1.4%	1,600 <sup>(1)</sup>
Cholangiocarcinoma	79,900	85,800	92,900	103,700	1.8%	1.5%	6,900 <sup>(2)</sup>
Glioma	29,600	32,600	35,900	41,000	2.5%	1.8%	13,600 <sup>(3)</sup>

Source: Frost & Sullivan Analysis

(1) Includes all patients eligible for first-line treatment of chemo-ineligible IDH1m AML (600) and R/R IDH1m AML (1,000)

(2) Includes all patients eligible for second-line treatment of IDH1m cholangiocarcinoma

(3) Includes all patients eligible for treatment of recurrent IDH1m Grade 2/3 glioma

### KIT & PDGFR $\alpha$ Specific Inhibitors

KIT is a receptor tyrosine kinase that is expressed on the surface of hematopoietic stem cells and certain cancer cells. It binds to stem cell factor (SCF) and activates downstream signaling pathways for cell survival, proliferation, and differentiation. Similarly, platelet-derived growth factor receptor  $\alpha$  (PDGFR $\alpha$ ) is also a receptor tyrosine kinase, but it binds to the PDGF family of proteins, and it also plays a role in cellular survival, growth, and differentiation. PDGFR $\alpha$  and KIT are both commonly mutated in gastrointestinal stromal tumors (GIST). Globally, there is currently no marketed KIT or PDGFR $\alpha$  specific inhibitor. In China, avapritinib (CS3007) of CStone is the only KIT or PDGFR $\alpha$  specific inhibitor under clinical development.

The KIT and PDGFR $\alpha$  specific inhibitor market is primarily driven by the number of addressable patients with GIST. The PDGFR $\alpha$  D842V mutation is found in approximately 5% of frontline unresectable or metastatic GIST patients, according to the Frost & Sullivan Report. Mutation in the PDGFR $\alpha$  gene are also observed in approximately 65.0% of aggressive SM patients. The number of KIT and PDGFR $\alpha$  addressable GIST patients in China has been growing, and is expected to grow steadily as shown in the table below.

- 189 -

Figuur A.13: Origineel document.

---

## INDUSTRY OVERVIEW

---

The IDH1m inhibitor market is primarily driven by the number of addressable patients with AML, cholangiocarcinoma and glioma. IDH1 mutations are observed in approximately 5.5% of AML patients, 8% of cholangiocarcinoma patients and 77% of glioma patients. The number of IDH1m inhibitor addressable patients in China has been growing, and is expected to grow steadily as shown in the table below.

Breedte/Lead 1.00	IDH1m Inhibitor Market in China				
	Historical CAGR (2013-2017)	Predictive CAGR (2017-2030E)	Addressable Patients (2017)	cel11.00	cel10.00
cel11.00 Indication	cel10.00 cel10.00 cel10.00 2013 2017 2022E 2030E	cel11.00 cel10.00 cel10.00 (2013-2017) (2017-2030E)	cel11.00 Patients (2017)	cel11.00	cel10.00
cel10.75 AML	cel10.00 cel10.00 cel10.00 1,540 16,800 20,000 22,500	cel11.00 cel10.00 cel10.00 15% 15% 15%	cel11.00 Patients (2017)	cel10.00	cel10.00
cel10.00 Cholangiocarcinoma	cel10.00 cel10.00 cel10.00 1,540 16,800 20,000 22,500	cel11.00 cel10.00 cel10.00 15% 15% 15%	cel11.00 Patients (2017)	cel10.00	cel10.00
Glioma	cel10.00 cel10.00 cel10.00 1,540 16,800 20,000 21,000	cel11.00 cel10.00 cel10.00 15% 15% 15%	cel11.00 Patients (2017)	cel10.00	cel10.00

Source: Frost & Sullivan Analysis

- (1) Includes all patients eligible for first-line treatment of chemo-ineligible IDH1m AML (600) and R/R IDH1m AML (1,000)
- (2) Includes all patients eligible for second-line treatment of IDH1m cholangiocarcinoma
- (3) Includes all patients eligible for treatment of recurrent IDH1m Grade 2/3 glioma

### KIT & PDGFR $\alpha$ Specific Inhibitors

KIT is a receptor tyrosine kinase that is expressed on the surface of hematopoietic stem cells and certain cancer cells. It binds to stem cell factor (SCF) and activates downstream signaling pathways for cell survival, proliferation, and differentiation. Similarly, platelet-derived growth factor receptor  $\alpha$  (PDGFR $\alpha$ ) is also a receptor tyrosine kinase, but it binds to the PDGF family of proteins, and it also plays a role in cellular survival, growth, and differentiation. PDGFR $\alpha$  and KIT are both commonly mutated in gastrointestinal stromal tumors (GIST). Globally, there is currently no marketed KIT or PDGFR $\alpha$  specific inhibitor. In China, avapritinib (CS3007) of CStone is the only KIT or PDGFR $\alpha$  specific inhibitor under clinical development.

The KIT and PDGFR $\alpha$  specific inhibitor market is primarily driven by the number of addressable patients with GIST. The PDGFR $\alpha$  D842V mutation is found in approximately 5% of frontline unresectable or metastatic GIST patients, according to the Frost & Sullivan Report. Mutation in the PDGFR $\alpha$  gene are also observed in approximately 65.0% of aggressive SM patients. The number of KIT and PDGFR $\alpha$  addressable GIST patients in China has been growing, and is expected to grow steadily as shown in the table below.

– 189 –

Figuur A.14: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Indication	2013	2017	2022E	2030E	Historical CAGR (2013-2017)
0	AML	17,400	18,800	20,100 22,500	22,500	1.9%
1	Cholangiocarcinoma	79,900	85,800	92,900 103,700	1.8%	1.5%
2	Glioma	29,600	32,600	35,000 ALONO	41,000	25%

Getransformeerde tabellen met algoritme B:

	Historical	Predictive	Addressable	NaN	NaN	NaN	NaN	NaN
0	CAGR	CAGR	Patients	None	None	None	None	None
1	(2013-2017)	(2017-2030E)	(2017)	Indication	2013	2017	None	None
2	1.9%	1.4%	1,600	AML	17,400	18,800	20,100	22,500
3	1.8%	1.5%	6,900	Cholangiocarcinoma	79,900	85,800	92,900	103,700
4	2.5%	1.8%	13,600	Glioma	29,600	32,600	35,900	41,000

## A.8 Document 8

<b>Global Wealth and Asset Management</b>		
For the years ended December 31, (\$ millions)	2018	2017
<b>Core EBITDA</b>	<b>\$ 1,498</b>	<b>\$ 1,426</b>
Amortization of deferred acquisition costs and other depreciation	(301)	(344)
Amortization of deferred sales commissions	(98)	(99)
Core earnings before income taxes	1,099	983
Core income tax (expense) recovery	(113)	(167)
<b>Core earnings</b>	<b>\$ 986</b>	<b>\$ 816</b>

**Expense efficiency ratio** is a non-GAAP measure which Manulife uses to measure progress towards our target to be more efficient. Efficiency ratio is defined as pre-tax general expenses included in core earnings divided by the sum of pre-tax core earnings and pre-tax general expenses included in core earnings.

**Embedded value ("EV")** is a measure of the present value of shareholders' interests in the expected future distributable earnings on in-force business reflected in the Consolidated Statements of Financial Position of Manulife, excluding any value associated with future new business. EV is calculated as the sum of the adjusted net worth and the value of in-force business. The adjusted net worth is the IFRS shareholders' equity adjusted for goodwill and intangibles, fair value of surplus assets, the carrying value of debt and preferred shares, and local statutory balance sheet, regulatory reserve, and capital for Manulife's Asian business. The value of in-force business in Canada and the U.S. is the present value of expected future IFRS earnings of in-force business less the present value of the cost of holding capital to support the in-force business under the MCCSR framework. The MCCSR framework was replaced by the LICAT framework on January 1, 2018 and LICAT was used to calculate EV as at December 31, 2018. It has been used to calculate quarterly NBV starting January 1, 2018. The value of in-force business in Asia reflects local statutory earnings and capital requirements. The value of in-force excludes our Global WAM, Manulife Bank and Property and Casualty Reinsurance businesses.

**New business value ("NBV")** is the change in embedded value as a result of sales in the reporting period. NBV is calculated as the present value of shareholders' interests in expected future distributable earnings, after the cost of capital, on actual new business sold in the period using assumptions that are consistent with the assumptions used in the calculation of embedded value. NBV excludes businesses with immaterial insurance risks, such as the Company's Global WAM, Manulife Bank and the short-term Property and Casualty Reinsurance businesses. NBV is a useful metric to evaluate the value created by the Company's new business franchise.

**New business value margin** is calculated as NBV divided by APE excluding non-controlling interests. APE is calculated as 100% of annualized first year premiums for recurring premium products, and as 10% of single premiums for single premium products. Both NBV and APE used in the NBV margin calculation are after non-controlling interests and exclude our Global WAM, Manulife Bank and Property and Casualty Reinsurance businesses. The NBV margin is a useful metric to help understand the profitability of our new business.

### Sales are measured according to product type:

For individual insurance, sales include 100% of new annualized premiums and 10% of both excess and single premiums. For individual insurance, new annualized premiums reflect the annualized premium expected in the first year of a policy that requires premium payments for more than one year. Single premium is the lump sum premium from the sale of a single premium product, e.g. travel insurance. Sales are reported gross before the impact of reinsurance.

For group insurance, sales include new annualized premiums and administrative services only premium equivalents on new cases, as well as the addition of new coverages and amendments to contracts, excluding rate increases.

APE sales is comprised of 100% of regular premiums/deposits and 10% of single premiums/deposits for both insurance and insurance-based wealth accumulation products.

Insurance-based wealth accumulation product sales include all new deposits into variable and fixed annuity contracts. As we discontinued sales of new Variable Annuity contracts in the U.S. in 1Q13, subsequent deposits into existing U.S. Variable Annuity contracts are not reported as sales. Asia variable annuity deposits are included in APE sales.

Bank new lending volumes include bank loans and mortgages authorized in the period.

**Gross flows** is a new business measure presented for our Global WAM business and includes all deposits into mutual funds, college savings 529 plans, group pension/retirement savings products, private wealth and institutional asset management products. Gross flows is a common industry metric for WAM businesses as it provides a measure of how successful the businesses are at attracting assets.

**Net flows** is presented for our Global WAM business and includes gross flows less redemptions for mutual funds, college savings 529 plans, group pension/retirement savings products, private wealth and institutional asset management products. Net flows is a common industry metric for WAM businesses as it provides a measure of how successful the businesses are at attracting and retaining assets.

Figuur A.15: Origineel document.

Global Wealth and Asset Management		2018	2017
For the years ended December 31,			
General			
<b>Core earnings</b>			
Amortization of deferred acquisition costs and other depreciation on			
Amortization of deferred sales commissions			
<b>Interest</b>			
Core earnings before income taxes		\$ 1,498	\$ 1,426
<b>Income taxes</b>		(341)	(344)
Core earnings after income taxes		(98)	(99)
<b>Core earnings</b>		\$ 1,099	\$ 983
Core earnings less expense recovered		(113)	(167)
<b>Core earnings</b>		\$ 986	\$ 816

**Expense efficiency ratio** is a non-GAAP measure which Manulife uses to measure progress towards our target to be more efficient. Efficiency ratio is defined as pre-tax general expenses included in core earnings divided by the sum of pre-tax core earnings and pre-tax general expenses included in core earnings.

**Embedded value ("EV")** is a measure of the present value of shareholders' interests in the expected future distributable earnings on in-force business reflected in the Consolidated Statements of Financial Position of Manulife, excluding any value associated with future new business. EV is calculated as the sum of the adjusted net worth and the value of in-force business. The adjusted net worth is the IFRS shareholders' equity adjusted for goodwill and intangibles, fair value of surplus assets, the carrying value of debt and preferred shares, and local statutory balance sheet, regulatory reserve, and capital for Manulife's Asian business. The value of in-force business in Canada and the U.S. is the present value of expected future IFRS earnings on in-force business less the present value of the cost of holding capital to support the in-force business under the MCCC framework. The MCCC framework was replaced by the LICAT framework on January 1, 2018 and LICAT was used to calculate EV as at December 31, 2018. It has been used to calculate quarterly NBV starting January 1, 2018. The value of in-force business in Asia reflects local statutory earnings and capital requirements. The value of in-force excludes our Global WAM, Manulife Bank and Property & Casualty Reinsurance businesses.

**New business value ("NBV")** is the change in embedded value as a result of sales in the reporting period. NBV is calculated as the present value of shareholders' interests in expected future distributable earnings, after the cost of capital, on actual new business sold in the period using assumptions that are consistent with the assumptions used in the calculation of embedded value. NBV excludes businesses with immaterial insurance risks, such as the Company's Global WAM, Manulife Bank and the short-term Property and Casualty Reinsurance businesses. NBV is a useful metric to evaluate the value created by the Company's new business franchise.

**New business value margin** is calculated as NBV divided by APE excluding non-controlling interests. APE is calculated as 100% of annualized first year premiums for recurring premium products, and as 10% of single premiums for single premium products. Both NBV and APE used in the NBV margin calculation are after non-controlling interests and exclude our Global WAM, Manulife Bank and Property & Casualty Reinsurance businesses. The NBV margin is a useful metric to help understand the profitability of our new business.

#### Sales are measured according to product type:

For individual insurance, sales include 100% of new annualized premiums and 10% of both excess and single premiums. For individual insurance, new annualized premiums reflect the annualized premium expected in the first year of a policy that requires premium payments for more than one year. Single premium is the lump sum premium from the sale of a single premium product, e.g. travel insurance. Sales are reported gross before the impact of reinsurance.

For group insurance, sales include new annualized premiums and administrative services only premium equivalents on new cases, as well as the addition of new coverages and amendments to contracts, excluding rate increases.

APE sales is comprised of 100% of regular premiums/deposits and 10% of single premiums/deposits for both insurance and insurance-based wealth accumulation products.

Insurance-based wealth accumulation product sales include all new deposits into variable and fixed annuity contracts. As we discontinued sales of new Variable Annuity contracts in the U.S. in 1Q13, subsequent deposits into existing U.S. Variable Annuity contracts are not reported as sales. Asia variable annuity deposits are included in APE sales.

Bank new lending volumes include bank loans and mortgages authorized in the period.

**Gross flows** is a new business measure presented for our Global WAM business and includes all deposits into mutual funds, college savings 529 plans, group pension/retirement savings products, private wealth and institutional asset management products. Gross flows is a common industry metric for WAM businesses as it provides a measure of how successful the businesses are at attracting assets.

**Net flows** is presented for our Global WAM business and includes gross flows less redemptions for mutual funds, college savings 529 plans, group pension/retirement savings products, private wealth and institutional asset management products. Net flows is a common industry metric for WAM businesses as it provides a measure of how successful the businesses are at attracting and retaining assets.

90 Manulife Financial Corporation | 2018 Annual Report | Management's Discussion and Analysis

Figuur A.16: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

For the years ended December 31,\n(\$ millions)	2018	2017
0 mortization of deferred acquisition costs and ...	\$ 1,498	\$ 1,426
1 Core earnings before income taxes\nCore income...	1,099\n(113)	983\n(167)
2 Core earnings	S\$ 986	\$ 816

Getransformeerde tabellen met algoritme B:

	For the years ended December 31,	NaN	NaN	NaN
0 (§ millions)	2018	None	2017	
1 Core EBITDA	\$1,498	None	\$ 1,426	
2 Amortization of deferred acquisition costs and...	(301)	depreciation	(344)	
3 Amortization of deferred sales commissions	(9a)	None	(99)	
4 Core earnings before income taxes	1,099	None	983	
5 Core tax (expense) recovery	(113)	None	(167)	
6 Core earnings	\$ 986	None	\$ 816	

## A.9 Document 9

in which suitable services for citizens (as well as businesses) are only available online. In countries like Denmark, the Netherlands and the United Kingdom, specific measures are taken for the 15 to 20 per cent of citizens who, in practice, will not be able to access online services at least in the short term. Overall, the government still makes very significant efforts by making online services compulsory through boosting usage, even though it needs to make special arrangements to assist this small group. (An example of the UK's Assisted Digital Team is provided in Chapter 5.)

Moreover, there are substantial usability challenges for certain groups of citizens (e.g. older persons interacting with Medicare).<sup>10</sup> Seen from an inclusive, sustainable development perspective, usability is therefore a particularly important usage issue; warranting governments' primary attention in their efforts to advance sustainable development through increased uptake of such groups.

Better usability means greater ability to use e-services, hence increasing the chance for e-government uptake. If a website is difficult to use, people leave. The design needs to make it easy for visitors to find what they need and do the tasks that they have come for. The recent usability flaws of Healthcare.gov (one of the centrepieces of the US Affordable Care Act) show the critical importance of usability for user uptake and e-government success. Likewise, in the EU, difficulties in usage were also identified as a key barrier to adopting e-government services. Search features are the most common usability features found on Member States' websites, with 168 countries including such features. Audio or video contents were present on 148 websites. Advanced search features are less common with only 101 countries (see Table 7.1).

Table 7.1. Number of countries with selected usability features

Usage-enhancing features—available on national portals	Number of countries
'Contact us' feature	185
Search feature	168
Audio or video contents	148
Site map or index	131
Advanced search options such as search filters	101
'Help' feature or 'Frequently Asked Questions (FAQs)' section	89
Information on how to make use of datasets	34

Ultimately, it is however not easy to answer the question: what "usability" means and what its components are. Our understanding of usability is also changing and it needs to be interpreted in a variety of new ways, which will also depend on the particular stage of development and the specific needs of each country or locality. In more advanced countries like the United Kingdom, the United States of America, Finland, Singapore and the Republic of Korea, usability is moving away from reliance on sophisticated one-stop-shop navigation portals to a number of other approaches, including the "findability" of a specific service through advanced search engines. The new UK portal ([www.gov.uk](http://www.gov.uk)) has adopted this approach after it was shown that only about 10 per cent of citizens found a service

Figuur A.17: Origineel document.

## CHAPTER 7 • E-GOVERNMENT FOR THE POST-2015 ERA: THE USAGE PERSPECTIVE

in which suitable services for citizens (as well as businesses) are only available online. In countries like Denmark, the Netherlands and the United Kingdom, specific measures are taken for the 15 to 20 per cent of citizens who, in practice, will not be able to access online services at least in the short term. Overall, the government still makes very significant efforts by making online services compulsory through boosting usage, even though it needs to make special arrangements to assist this small group. (An example of the UK's Assisted Digital Team is provided in Chapter 5.)

Moreover, there are substantial usability challenges for certain groups of citizens (e.g. older persons interacting with Medicare).<sup>13</sup> Seen from an inclusive, sustainable development perspective, usability is therefore a particularly important usage issue, warranting governments' primary attention in their efforts to advance sustainable development through increased uptake of such groups.

Better usability means greater ability to use e-services, hence increasing the chance for e-government uptake. If a website is difficult to use, people leave. The design needs to make it easy for visitors to find what they need and do the tasks that they have come for. The recent usability flaws of Healthcare.gov (one of the centrepieces of the US Affordable Care Act) show the critical importance of usability for user uptake and e-government success. Likewise, in the EU, difficulties in usage were also identified as a key barrier to adopting e-government services. Search features are the most common usability features found on Member States' websites, with 168 countries including such features. Audio or video contents were present on 148 websites. Advanced search features are less common with only 101 countries (see Table 7.1).

Table 7.1 Number of countries with selected usability features

Boxed feature	Number of countries
Usage-enhancing features—available on national portals	168
Contact us' feature	185
Search feature	168
Search feature	168
Audio or video contents	148
Site map or index	131
Advanced search options such as search filters	101
'Help' feature or 'Frequently Asked Questions (FAQs)' section	89
Information on how to make use of datasets	34

Ultimately, it is however not easy to answer the question: what "usability" means and what its components are. Our understanding of usability is also changing and it needs to be interpreted in a variety of new ways, which will also depend on the particular stage of development and the specific needs of each country or locality. In more advanced countries like the United Kingdom, the United States of America, Finland, Singapore and the Republic of Korea, usability is moving away from reliance on sophisticated one-stop-shop navigation portals to a number of other approaches, including the "findability" of a specific service through advanced search engines. The new UK portal ([www.gov.uk](http://www.gov.uk)) has adopted this approach after it was shown that only about 10 per cent of citizens found a service

146

Figuur A.18: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Usage-enhancing features—available on national portals		Number of countries
0	'Contact us' feature	185
1	Search feature	168
2	Audio or video contents	148
3	Site map or index	131
4	Advanced search options such as search filters	101
5	'Help' feature or 'Frequently Asked Questions ...	B89
6	Information on how to make use of datasets	34

Getransformeerde tabellen met algoritme B:

	Usage-enhancing features—available on national portals	Number of countries
0	‘Contact us’ feature	185
1	Search feature	168
2	Audio or video contents	148
3	Site map or index	131
4	Advanced search options such as search filters	101
5	‘Help’ feature or ‘Frequently Asked Questions ...	89
6	Information on how to make use of datasets	34

## A.10 Document 10

UNITED NATIONS E-GOVERNMENT SURVEY 2014

months. Ease of use was further supported by access to content in more than one language in 74 per cent of cases (142 countries), availability of a site map or index 68 per cent of the time (131 countries) and the online publication of a help or frequently asked questions document in 46 per cent of cases (89 countries).

The Survey shows continued effort of most countries in building and maintaining site-specific tools, notwithstanding the ubiquity and convenience of commercial search engines. There is also a growing recognition of the importance of providing content in different languages. In 2012, just over half of all countries had multilingual sites whereas according to the 2014 assessment almost three quarters had expanded language options in some form.

Enhanced (Stage 2) features are less common as Table 2.4 suggests. Roughly half of the United Nations Member States maintain an advanced search engine or publish a statement setting out a privacy policy in connection with the use of the government website. User opinion features such as tag clouds and 'hot topics' lists could be found on only 41 per cent of websites. Less than one third of national portals surveyed showed availability of a secure connection.

Table 2.3. Availability of selected basic features

	Number of countries	Per centage of countries
Find website using search tools	168	87%
Site updated within past three months	148	77%
Access in more than one language	142	74%
Availability of map/index	131	68%
Help/FAQ feature	89	46%

Table 2.4. Availability of selected enhanced features

	Number of countries	Per centage of countries
Advanced search option	101	52%
Privacy statement	97	50%
Tag cloud or 'hot topics'	80	41%
Secure website	53	27%

Figure 2.4 provides a breakdown of typical transactional services and the number of countries for which these services could be readily identified through the national website. Of the transactional services included in the Survey instrument, the most commonly found were setting up of personal online accounts (101 countries), income tax filing (73 countries) and business registration (60 countries). An open-ended 'other' category also scored well (76 countries) reflecting a diversity of priorities in building and expanding online services at national level.

51

Figuur A.19: Origineel document.

## UNITED NATIONS E-GOVERNMENT SURVEY 2014

months. Ease of use was further supported by access to content in more than one language in 74 per cent of cases (142 countries), availability of a site map or index 68 per cent of the time (131 countries) and the online publication of a help or frequently asked questions document in 46 per cent of cases (89 countries).

The Survey shows continued effort of most countries in building and maintaining site-specific tools, notwithstanding the ubiquity and convenience of commercial search engines. There is also a growing recognition of the importance of providing content in different languages. In 2012, just over half of all countries had multilingual sites whereas according to the 2014 assessment almost three quarters had expanded language options in some form.

Enhanced (Stage 2) features are less common as Table 2.4 suggests. Roughly half of the United Nations Member States maintain an advanced search engine or publish a statement setting out a privacy policy in connection with the use of the government website. User opinion features such as tag clouds and 'hot topics' lists could be found on only 41 per cent of websites. Less than one third of national portals surveyed showed availability of a secure connection.

Table 2.3. Availability of selected basic features

Indicator (2.0)	2012 (n=142)	2014 (n=148)
	Number of countries	Per centage of countries
Search tool	168	87% 88/100
Site updated within past three months	148	27% 27/100
Available in more than one language	142	74% 74/100
Availability of map/index	131	89% 89/100
HelpFAQ feature	89	60% 60/100

Table 2.4. Availability of selected enhanced features

Indicator (2.0)	2012 (n=142)	2014 (n=148)
	Number of countries	Per centage of countries
Advanced search option	101	52% 52/100
Privacy statement	97	50% 50/100
Tag cloud or 'hot topics'	80	41% 41/100
Secure website	53	27% 27/100

Figure 2.4 provides a breakdown of typical transactional services and the number of countries for which these services could be readily identified through the national website. Of the transactional services included in the Survey instrument, the most commonly found were setting up of personal online accounts (101 countries), income tax filing (73 countries) and business registration (60 countries). An open-ended 'other' category also scored well (76 countries) reflecting a diversity of priorities in building and expanding online services at national level.

51

Figuur A.20: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Number\nof countries	Per centage\nof countries
0 Advanced search optioor	52%	
1 Privacy statement	50%	
2 Taq cloud or 'hot topics'	AIM	
3 Secure website	27%	

	Number\nof countries	Per centage\nof countries
0	Find website using search tools	87%
1	Site updated within past three months	77%
2	Access in more than one lanquaas	TA%
3	Availability of map/index	68%
4	Help/FAQ feature	Ab%

Getransformeerde tabellen met algoritme B:

	Number	Per centage	NaN
0	of countries	of countries	None
1	101	52%	Advanced search option
2	97	50%	Privacy statement
3	80	41%	Tag cloud or ‘hot topics’
4	53	27%	Secure website

	Number	Per centage	NaN
0	of countries	of countries	None
1	168	87%	Find website using search tools
2	148	71%	Site updated within past three months
3	142	74%	Access in more than one language
4	131	68%	Availability of map/index
5	89	46%	Help/FAQ feature

## A.11 Document 11

Bordered Table Layout	
Country	Capital
India	Delhi
Bangladesh	Dhaka
Nepal	Kathmandu

Figuur A.21: Origineel document.

Bordered 0.97 Bordered Table Layout	
Country	Capital
India	Delhi
Bangladesh	Dhaka
Nepal	Kathmandu

Figuur A.22: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Country	Capital
0	India	Delhi
1	Bangladesh	Dhaka
2	Nepal	Kathmandu

Getransformeerde tabellen met algoritme B:

	Country	Capital
0	India	Delhi
1	Bangladesh	Dhaka
2	Nepal	Kathmandu

## A.12 Document 12

### Chapter 2

#### CHAPTER 2 • PROGRESS IN ONLINE SERVICE DELIVERY

Table 2.6. High online service performance relative to income

Country	Online Service Index	Income group
Rwanda	0.5118	Low
Colombia	0.7874	Upper Middle
Ethiopia	0.4567	Low
Kazakhstan	0.7480	Upper Middle
Morocco	0.6929	Lower Middle
Kenya	0.4252	Low
Sri Lanka	0.6535	Lower Middle
Malaysia	0.6772	Upper Middle
Tunisia	0.6378	Upper Middle
Mongolia	0.6142	Lower Middle

Table 2.7. Low online service performance relative to income

Country	Online Service Index	Income group
Equatorial Guinea	0.0315	High
Monaco	0.2205	High
Libya	0.0157	Upper Middle
Saint Kitts and Nevis	0.1339	High
San Marino	0.2756	High
Tuvalu	0.0394	Upper Middle
Barbados	0.2205	High
Algeria	0.0787	Upper Middle
Sao Tome and Principe	0.0079	Lower Middle

Ultimately, the measure of online service utility is its impact on development either directly in provision of services to citizens or indirectly, for example through investment linked to apparent ease of doing business. Low- and middle-income countries with relatively low levels of Internet use such as Ethiopia (1.48 per cent of the population are Internet users), Rwanda (8.02 per cent of the population are Internet users) and Sri Lanka (18.29 per cent of the population are Internet users), and relatively high online service scores may need to invest more in securing telecommunication infrastructure to fully optimize the benefit of e-services.

The converse also applies to high income countries with widespread telecommunication infrastructure and low online service scores such as Monaco (87.00 per cent of the population are Internet users), Saint Kitts and Nevis (79.35 per cent of the population are Internet users) and Barbados (73.33 per cent of the population are Internet users). These are all small countries, and it may be the case that a larger critical mass of Internet users, or potential users, makes it more worthwhile for a country to invest in resource intensive forms of online service delivery such as remote health care, smart energy grids and real-time environmental monitoring. The Survey does not, however, require such technological advancement for high scores reflecting the view that even relatively simple information sharing and interaction can produce important benefits when the primary needs and attributes of population segments are reflected in online service design.

Figuur A.23: Origineel document.

## CHAPTER 2 • PROGRESS IN ONLINE SERVICE DELIVERY

Table 2.6. High online service performance relative to income

Country	Online Service Index	Income group
Colombia	0.0419	Low, 1.00
Ethiopia	0.0403	Upper Middle
Kazakhstan	0.0402	Low, 1.00
Monaco	0.0393	Upper Middle
Rwanda	0.0392	Lower Middle
Sri Lanka	0.0383	Lower Middle
Mauritius	0.0373	Upper Middle
Tunisia	0.0372	Upper Middle
Mongolia	0.0362	Lower Middle

Table 2.7. Low online service performance relative to income

Country	Online Service Index	Income group
Equatorial Guinea	0.0318	High
Monaco	0.0318	High, 1.00
Libya	0.0157	Upper Middle
Saint Kitts and Nevis	0.0139	High
Sao Tome and Principe	0.0094	High, 1.00
Togo	0.0094	Upper Middle
Barbados	0.0085	High, 1.00
Albania	0.0082	Upper Middle
San Tome and Principe	0.0075	Lower Middle

Ultimately, the measure of online service utility is its impact on development either directly in provision of services to citizens or indirectly, for example through investment linked to apparent ease of doing business. Low- and middle-income countries with relatively low levels of Internet use such as Ethiopia (1.48 per cent of the population are Internet users), Rwanda (8.02 per cent of the population are Internet users) and Sri Lanka (18.29 per cent of the population are Internet users), and relatively high online service scores may need to invest more in securing telecommunication infrastructure to fully optimize the benefit of e-services.

The converse also applies to high income countries with widespread telecommunication infrastructure and low online service scores such as Monaco (87.00 per cent of the population are Internet users), Saint Kitts and Nevis (79.35 per cent of the population are Internet users) and Barbados (73.33 per cent of the population are Internet users). These are all small countries, and it may be the case that a larger critical mass of Internet users, or potential users, makes it more worthwhile for a country to invest in resource intensive forms of online service delivery such as remote health care, smart energy grids and real-time environmental monitoring. The Survey does not, however, require such technological advancement for high scores reflecting the view that even relatively simple information sharing and interaction can produce important benefits when the primary needs and attributes of population segments are reflected in online service design.

56

Figuur A.24: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Country	Online Service Index	Income group
0	Equatorial Guinez	0.0315	High
1	Monaco	0.2205	None
2	Libya	0.0157	Upper Middle
3	Saint Kitts and Nevis	0.1339	High
4	San Marino	0.2756	High
5	Tuvalu	0.0394	Upper Middle
6	Barbados	0.2205	None
7	Algeria	0.0787	Upper Middle
8	Sao Tome and Principe	0.0079	Lower Middle

	Country	Online Service Index	Income group
0	Rwanda	0.5118	Low
1	Colombia	0.7874	Upper Middle
2	Ethiopia	0.4567	Low
3	Kazakhstan	0.7480	Upper Middle
4	Morocco	0.6929	Lower Middle
5	Kenya	0.4252	Low
6	Sri Lanka	0.6535	Lower Middle
7	Malaysia	0.6772	Upper Middle
8	Tunisia	0.6378	Upver Middle
9	Mongolia	0.6142	Lower Middle

Getransformeerde tabellen met algoritme B:

	Country	Online Service Index	income group
0	Equatorial Guinea	0.0315	High
1	Monaco	0.2205	High
2	Libya	0.0157	Upper Middle
3	Saint Kitts and Nevis	0.1339	High
4	San Marino	0.2756	High
5	Tuvalu	0.0394	Upper Middle
6	Barbados	0.2205	High
7	Algeria	0.0787	Upper Middle
8	Sao Tome and Principe	0.0079	Lower Middle

	Country	Online Service index	Income group
0	Rwanda	0.5118	Low
1	Colombia	0.7874	Upper Middle
2	Ethiopia	0.4567	Low
3	Kazakhstan	0.7480	Upper Middle
4	Morocco.	0.6929	Lower Middle
5	Kenya	0.4252	Low
6	Sti Lanka	0.6535	Lower Middle
7	Malaysia	0.6772	Upper Middle
8	Tunisia	0.6378	Upper Middle
9	Mongolia	0.6142	Lower Middle

## A.13 Document 13

### Users

First Name	Last Name	Username	Email
John	Doe	john_doe1234	john_doe@email.com
Jane	Doe	jane_doe1234	jane_doe@email.com
John	Doe	john_doe1234	john_doe@email.com
Jane	Doe	jane_doe1234	jane_doe@email.com
John	Doe	john_doe1234	john_doe@email.com
Jane	Doe	jane_doe1234	jane_doe@email.com

Figuur A.25: Origineel document.

### Users

First Name	Last Name	Username	Email
John	Doe	john_doe1234	john_doe@email.com
Jane	Doe	jane_doe1234	jane_doe@email.com
John	Doe	john_doe1234	john_doe@email.com
Jane	Doe	jane_doe1234	jane_doe@email.com
John	Doe	john_doe1234	john_doe@email.com
Jane	Doe	jane_doe1234	jane_doe@email.com

Figuur A.26: Gedetecteerde tabellen.

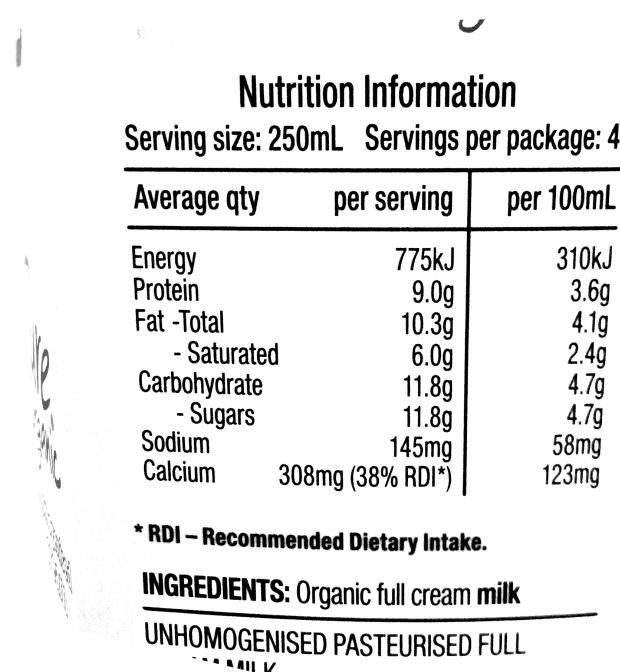
Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

	First Name	Last Name	Username	Email
0	John	Doe	john_doe1234	john_doe@email.com
1	Jane	Doe	jane_doe1234	jane_doe@email.com
2	John	Doe	john_doe1234	john_doe@email.com
3	Jane	Doe	jane_doe1234	jane_doe@email.com
4	John	Doe	john_doe1234	john_doe@email.com
5	Jane	Doe	\jane_doe1234	\jane_doe@email.com

## A.14 Document 14



Figuur A.27: Origineel document.

<b>Nutrition Information</b>		
Serving size: 250mL Servings per package: 4		
Average qty	per serving	per 100mL
Energy	775kJ	310kJ
Protein	9.0g	3.6g
Fat -Total	10.3g	4.1g
- Saturated	6.0g	2.4g
Carbohydrate	11.8g	4.7g
- Sugars	11.8g	4.7g
Sodium	145mg	58mg
Calcium	308mg (38% RDI*)	123mg

\* RDI – Recommended Dietary Intake.

**INGREDIENTS:** Organic full cream milk  
**UNHOMOGENISED PASTEURISED FULL**

Figuur A.28: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

	Average aty	per	serving	NaN	100mL
0	Energy	None	775kd	None	310kJ
1	Protein	None	9.09	None	3.69
2	Fat -Total	None	10.3g	None	41g
3	- Saturated	None	6.0g	None	2.49
4	Carbohydrate	None	11.89	None	4/9
5	- Sugars	None	11.89	None	479
6	sodium	None	None	None	None
7	None	None	145mg	None	58mg
8	Calcium	(38% RDI*)	308mg	123mg	

## A.15 Document 15

Opleidingsprogramma	WF	SP Examen	Score	Opmerking
2 Bedrijfsmgmt	3	3 EK1 S1	12	Creditbewijs
3 Analyse III	3	3 EK1 S1	15	Creditbewijs
3 Internationale Communicatie III	3	3 EK1 S1	12	Creditbewijs
3 Artificiële Intelligentie	3	3 EK1 S1	16	Creditbewijs
3 Native apps I: mobile apps voor Android	3	3 EK1 S1	--	Afwezig - Tweede examenkans
3 Native apps II: mobile apps voor Windows	4	4 EK1 S1	--	Afwezig - Tweede examenkans
3 Web apps	3	3 EK1 S1	--	Afwezig - Tweede examenkans
3 Project III: Mobile apps	5	5 EK1 S1	15	Creditbewijs
3 ITalent	3	3 EK1 S2	15	Creditbewijs
3 Internationale stage S2 20STP	20	20 EK1 S2	17	Verworven
3 Bachelorproef SJ 7STP	7	7 EK1 S2	--	Afwezig - Tweede examenkans

Figuur A.29: Origineel document.

UNITED NATIONS E-GOVERNMENT SURVEY 2014

Chapter 5

A channel can change a user's perception and confidence of a public service. For instance, in the context of value proposition of a public service, a channel can add value with a positive user experience, and in the same vein, it can subtract value through a negative user experience. Therefore, it is extremely important to select the right channel for the right service targeting the specific audience. At the same time, channel selection is a deciding factor to effectively reach out to specific groups of citizens, for example, rural population with limited ICT access.

Table 5.1 List of channels (non-exhaustive)

1. Counter (face-to-face service)
2. Telephone (voice) service and call centres
3. Web portal
4. Email
5. SMS and other messaging service
6. Mobile portal/mobile website
7. Mobile app
8. Social media
9. Public kiosk
10. Intermediaries through public-private partnership

Digital channels, with both their diversity and spread, are increasingly embraced by almost all countries, while counter (face-to-face service) and telephone (voice) services, have continued to serve as fundamental channels as preferred by some citizen groups. Through strategic and optimised mixed use of channels, governments will be able to provide ubiquitous 24 x 7 access to information and services to different user groups.

### 5.2.1. Web portal

The 2014 Survey concludes that all 193 United Nations Member States have some form of online presence, as compared to 18 countries with no online presence in 2003 and 3 countries in 2012 (see Figure 2.1 in Chapter 2).

Research shows that offering more online services and aiming for increased usage of these services improves efficiency and results in cost reduction. According to one research study,<sup>10</sup> the Government of the United Kingdom could save between GBP 3.30 and GBP 12 per transaction by moving public services online. Denmark has made mandatory the use of online public services in its e-government strategy and action plan covering the period 2011–2015.<sup>11</sup> When fully implemented, it will go "digital by default" for citizens and businesses with the aim of making public service delivery more cost-efficient. Similarly, the Government Digital Strategy<sup>12</sup> of the United Kingdom published in November 2012, stresses that all services should be "digital by default". The strategy states 11 principles and 14 actions to shape how central government departments and agencies will embrace digitalization of their services and improve usage by citizens and businesses. The Government of the United Kingdom has since committed to the

97

Figuur A.30: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Opleidingsprogramma	WE	12 Creditbewijs\n15 Creditbewijls\n1
0 2 Bedrijfsmanagement\n3. Analyse III	3 EK1/S1	12 Creditbewijs\n15 Creditbewijs
1 3 Internationale Communicatie III\n3. Artifici...	3 EK1/S1	12 Creditbewijs\n16 Creditbewijs
2 Native apps I: mobile apps voor Android	3 EK1/S1	Afwezig - Tweede
3 .~=Native apps II: mobile apps voor Windows	4 EK1/S1	Afwezig - Tweede
4 Web apps	3 FK1IS1	Afwezig - Tweede
5 3 ~~~Project III: Mobile apps	5 EK1/S1	15 Creditbewijs
6 Talent	3 EK1IS2	15 Creditbewijs
7 Internationale stage S2 20STP\nBachelorproef S...	20 EK1/S2	None
8 3 =Bachelorproef SJ 7STP	7 EK1IS2	Afwezig - Tweede\n‘examenkans

Getransformeerde tabellen met algoritme B:

Opleidingsprogramma	‘SP Examen	Score Opmerking
0 2 Bedrijfsmanagement	3 EK1IS1	Creditbewijs
1 3 Analyse III	3 EK1IS1	Creditbewijs
2 3 Internationale Communicatie III	3 EK1IS1	Creditbewijs
3 3 Artificiële Intelligentie	3 EK1IS1	Creditbewijs
4 3. Native apps I: mobile apps voor Android	3 EK1IS1	Afwezig- Tweede
5 3. Native apps II: mobile apps voor Windows	4 EK1IS1	Afwezig- Tweede
6 3. Web apps	3 EK1IS1	Afwezig- Tweede
7 3 Project III: Mobile apps	\$ EK1IS1	15 Creditbewijs
8 3 iTalent	3 EK1IS2	15 Creditbewijs
9 3 Internationale stage \$2 20STP	20 20 EK1IS2	7 Verworven
10 3 Bachelorproef SJ 7STP	7 7	Afwezig- Tweede
11 None	None	examenkans
12 None	None	examenkans
13 None	None	examenkans

## A.16 Document 16

The screenshot shows a web interface with a navigation bar at the top. The main content area has a sidebar on the left with categories: Gegevens, Contracten, Documenten, and Voorkeuren. The 'Contracten' section is selected. A box titled 'Download hier je contract' contains instructions. Below is a table with columns: Seizoen, Locatie, Rol leerkracht, Status, Status, Download, and Upload.

Seizoen	Locatie	Rol leerkracht	Status	Status	Download	Upload
2018s	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	Vrijwilliger	Contract is in orde	<a href="#">Download</a>	✓
2018w	GO! Middenschool MAD	Series Assistant	Jobstudent	Contract is in orde	<a href="#">Download</a>	✓
2019w	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	Jobstudent	Contract is in orde	<a href="#">Download</a>	✓
2019s	GO! Middenschool MAD	Series Assistant	Jobstudent	Contract is in orde	<a href="#">Download</a>	✓
2019s	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	Jobstudent	Contract is in orde	<a href="#">Download</a>	✓
2019w	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	Jobstudent	Contract is in orde	<a href="#">Download</a>	✓

Figuur A.31: Origineel document.

The screenshot shows the same web interface as Figure A.31, but with the entire page highlighted in green. The sidebar and table structure remain the same, indicating the detection of the entire document structure.

Figuur A.32: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Seizoen	Locatie	Rol leekrachi	Statuut	Status
0	2018s	Hogeschool Gent - Campus\\nSchoonmeersen	Series\\nAssistant	Vrijwilliger	Contract
1	2018w	GO! Middenschool MAD	Series\\nAssistant	Jobstudent	Contract
2	2018w	Hogeschool Gent - Campus\\nSchoonmeersen	Series\\nAssistant	Jobstudent	Contract
3	2019s	GO! Middenschool MAD	Series\\nAssistant	Jobstudent	Contract
4	2019s	Hogeschool Gent - Campus\\nSchoonmeersen	Series\\nAssistant	Jobstudent	Contract
5	2019W	Hogeschool Gent - Campus\\nSchoonmeersen	Series\\nAssistant	Jobstudent	Contract

Getransformeerde tabellen met algoritme B:

	Seizoen	Locatie	Rol leekracht	Statuut	Status
0	20188	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	Vrijwilliger	Contract is i
1	20180	GO! Middenschool MAD	Series Assistant	dobstudent	Contract is i
2	20180	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	dobstudent	Contract is i
3	20198	GO! Middenschool MAD	Series Assistant	dobstudent	Contract is i
4	20198	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	dobstudent	Contract is i
5	2019"	Hogeschool Gent - Campus Schoonmeersen	Series Assistant	dobstudent	Contract is i

## A.17 Document 17

### GeeksforGeeks

S. No.	Name	City	Age
1	Ajay	Patna	20
2	Rahul	Chandigarh	17
3	Parush	Kolkata	22

Figuur A.33: Origineel document.

### GeeksforGeeks

S. No.	Name	City	Age
1	Ajay	Patna	20
2	Rahul	Chandigarh	17
3	Parush	Kolkata	22

Figuur A.34: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	S. No.	Name	City	Age
0	1	Ajay	Patna	20
1	2	Rahul	Chandigarh	17
2	3	Parush	Kolkata	22

Getransformeerde tabellen met algoritme B:

	S.No.	Name	City	Age
0	1	Ajay	Patna	20
1	2	Rahul	Chandigarh	7
2	3	Parush	Kolkata	22

## A.18 Document 18

### 4. Profile of Responses

A summary of the types of organisation investigated and the responses received is set out in Table 1. The surveys identified in the Table cover a range of different types, which are discussed in the paragraphs below according to the type of organisation. Most do not specifically solicit the view of the public. Surveys of particular relevance to public opinion are discussed in Section 5 below.

**Table 1 Summary of Responses**

Type of Organisation	No response	No surveys known	Total of Type	Surveys identified
European Institutions		2	3	4
Government / Regulators	4	8	14	2
Air Navigation Service Providers	11	8	22	3
Airport Authorities	1	3	4	
Airlines	4		4	
Representative Organisations	2	6	8	2
Pressure Groups / Watchdogs		2	4	2
Commercial Market Researchers		2	5	6
Research Laboratories	2	5	7	
Academia	2	6	9	1
<b>TOTALS</b>	<b>26</b>	<b>42</b>	<b>80</b>	<b>20</b>

5

Figuur A.35: Origineel document.

#### 4. Profile of Responses

A summary of the types of organisation investigated and the responses received is set out in Table 1. The surveys identified in the Table cover a range of different types, which are discussed in the paragraphs below according to the type of organisation. Most do not specifically solicit the view of the public. Surveys of particular relevance to public opinion are discussed in Section 5 below.

**Table 1. Summary of Responses**

Type of Organisation	No response	No surveys known	Total of Type	Surveys identified
European Institutions		2	3	4
Government / Regulators	4	8	14	2
Air Navigation Service Providers	11	8	22	3
Airport Authorities	1	3	4	
Airlines	4		4	
Representative Organisations	2	6	8	2
Pressure Groups / Watchdogs		2	4	2
Commercial Market Researchers		2	5	6
Research Laboratories	2	5	7	
Academia	2	6	9	1
<b>TOTALS</b>	<b>26</b>	<b>42</b>	<b>80</b>	<b>20</b>

5

Figuur A.36: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	No\nresponse	No surveys\nknown	Total of\nType	Surveys:\nidentified
0	4	8	14	None
1	11	8	22	3
2	1	3	4	None
3	4	4	None	None
4	2	6	8	2
5	2	5	7	None
6	2	6	9	41
7	2	3	None	None
8	2	4	2	None
9	2	5	6	None

Getransformeerde tabellen met algoritme B:

	Type of Organisation	No response	Nosurveys   known	Totalof Type	Surveys id
0	European Institutions	None	2	3	4
1	Government / Regulators	"4	a	4	2
2	Air Navigation Service Providers	14	a	2	3
3	Airport Authorities	4	3	4	None
4	Airlines	4	None	4	None
5	Representative Organisations	2	6	a	2
6	Pressure Groups / Watchdogs	None	2	4	2
7	Commercial Market Researchers,	None	2	5	6
8	Research Laboratories	2	5	7	None
9	Academia	2	6	9	4
10	TOTALS	26	42	BD	20

## A.19 Document 19

Nährwerte pro 100 g	Flohsamen (Zirkulin)	Gemahlene Flohsamenschalen (dm)
<b>Brennwert</b>	234 kcal / 980 kJ	208 kcal / 842 kJ
<b>Kohlenhydrate</b>	2,6 g	0 g
<b>Eiweiß</b>	15,8 g	3,8 g
<b>Fett</b>	2,8 g	2,8 g, davon gesättigt: 1 g
<b>Ballaststoffe</b>	68 g	84 g

Figuur A.37: Origineel document.

Bordered[1.00]		
Nährwerte pro 100 g	Flohsamen (Zirkulin)	Gemahlene Flohsamenschalen (dm)
Brennwert	234 kcal / 980 kJ	208 kcal / 842 kJ
Kohlenhydrate	2,6 g	0 g
Eiweiß	15,8 g	3,8 g
Fett	2,8 g	2,8 g, davon gesättigt: 1 g
Ballaststoffe	68 g	84 g

Figuur A.38: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Brennwert	234 kcal / 980 kJ
0 Kohlenhydrate	2,69
1 Eiweif&BQ	15,89
2 Fett	28 9

Getransformeerde tabellen met algoritme B:

	Nahrwerte pro 100 g	Flohsamen (Zirkulin)	Gemahlene Flohsamenschalen (dm)
0 Brennwert	234 keal / 980 kd	208 keal / 842 kJ	
1 Kohlenhydrate	264	og	
2 Eiweiß	15,84	389	
3 Fett	289	2.8 g. davon gesattigt: 1.9	
4 Ballaststoffe	68g	Bag	

## A.20 Document 20

AP Van Schoubroeck  
Oosterwijkseweg 2  
2250 Olen



Patient:  
Laatst gewijzigd door:  
Afdruk door:  
INSZ:  
Tijdstip:

Tijdstip opvraging:

Chronisch medicatie	Frequente	Begin	Einde	Inname/ Eenheid	Ontbijt			Middagmaal			Avondmaal			Opmerkingen				
					% morgens	Voor Tijdens	Na	Voor Tijdens	Na	Voor Tijdens	Na	Slaap	08 u	12 u	16 u	20 u		
ASAFLOW 80MG MAAGSAPRES COMP BLI 168X 80MG	Dagelijks	04/09/2015		Tablet		1												
AZILECT 1 MG TABL 28 X 1 MG	Dagelijks	04/09/2015		Tablet		1												
FORLAX PI PHARMA 20 ZAK 10 G	Dagelijks	04/09/2015		Zakje			1											
INDERAL COMP 50X10MG	Dagelijks	04/09/2015		Tablet		1		1		1								
MIRAPEXIN 3.15 MG COMP VERLENGDE AFGEFTE 100	Dagelijks	04/09/2015		Tablet		1												
STALEVO 100/25/200 100 TABL	Dagelijks	04/09/2015		Tablet									1	1	1	1		

Indien Nodig	Frequente	Begin	Einde	Inname/ Eenheid	Ontbijt			Middagmaal			Avondmaal			Opmerkingen				
					% morgens	Voor Tijdens	Na	Voor Tijdens	Na	Voor Tijdens	Na	Slaap	08 u	12 u	16 u	20 u		
TEMESTA EXPIDET COMP. 50 X 1 MG	Dagelijks	04/09/2015		Tablet										1				
PARACETAMOL TEVA 30 TABL 1 G		04/09/2015																

tot 3 x per dag 1 tablet, 6 uur tussen 2 innames

Indicatie: Migraine  
Instructie: 6 uur wachten tussen 2 innames

Huisarts: tel.  
Opgepast: Dit schema bevat alle informatie zoals gekend in de apotheek op datum

1/1

Figuur A.39: Origineel document.

AP Van Schoubroeck  
Oosterwijkseweg 2  
2250 Olen



Patient:  
Laatst gewijzigd door:  
Afdruk door:  
INSZ:  
Tijdstip:

Afdruk door:  
Tijdstip opvraging:

Chronische medicatie	Frequente	Begin	Einde	Inname/ Eenheid	Ontbijt			Middagmaal			Avondmaal			Opmerkingen				
					% morgens	Voor Tijdens	Na	Voor Tijdens	Na	Voor Tijdens	Na	Slaap	08 u	12 u	16 u	20 u		
ASAFLOW 80MG MAAGSAPRES COMP BLI 168X 80MG	Dagelijks	04/09/2015		Tablet		1												
AZILECT 1 MG TABL 28 X 1 MG	Dagelijks	04/09/2015		Tablet		1												
FORLAX PI PHARMA 20 ZAK 10 G	Dagelijks	04/09/2015		Zakje			1											
INDERAL COMP 50X10MG	Dagelijks	04/09/2015		Tablet		1		1		1								
MIRAPEXIN 3.15 MG COMP VERLENGDE AFGEFTE 100	Dagelijks	04/09/2015		Tablet		1												
STALEVO 100/25/200 100 TABL	Dagelijks	04/09/2015		Tablet									1	1	1	1		

Medicatieplan

Indien Nodig	Frequente	Begin	Einde	Inname/ Eenheid	Ontbijt			Middagmaal			Avondmaal			Opmerkingen				
					% morgens	Voor Tijdens	Na	Voor Tijdens	Na	Voor Tijdens	Na	Slaap	08 u	12 u	16 u	20 u		
TEMESTA EXPIDET COMP. 50 X 1 MG	Dagelijks	04/09/2015		Tablet									1					
PARACETAMOL TEVA 30 TABL 1 G		04/09/2015																

tot 3 x per dag 1 tablet, 6 uur tussen 2 innames

Indicatie: Migraine  
Instructie: 6 uur wachten tussen 2 innames

Huisarts: tel.  
Opgepast: Dit schema bevat alle informatie zoals gekend in de apotheek op datum

1/1

Figuur A.40: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

Chronische medicatie	Frequentie	NaN	Opmerking
2 ASAFLOW 80MG MAAGSAPRES COMP BLI 168X 80MG	Dagelijks	04/09/2015	Indicatie
3 AZILECT 1 MG TABL 28 X 1 MG	Dagelijks	04/09/2015	Indicatie
4 FORLAX PI PHARMA 20 ZAK 10G	Dagelijks	04/09/2015	Indicatie
5 INDERAL COMP 50X10MG	None	04/09/2015	Indicatie
6 MIRAPEXIN 3,15 MG COMP VERLENGDE AFGIFTE 100	Dageliks	04/09/2015	Indicatie
7 STALEVO 100/25/200 100 TABL	Dageliks	04/09/2015	Indicatie

## A.21 Document 21

DEBUT	INJECTIONS	8 h	12 h	17 h	20 h	FIN
	<b>MÉDICAMENTS Per Os</b>					
	<b>PAROXETINE 20 mg</b>	1				
	<b>FOLAVIT 4mg</b>	1				
	<b>ALDACTONE cp 25mg</b>	1				
	<b>ALLOPURINOL SANDOZ 100mg</b>			1		
	<b>BURINEX LEO cp 5mg</b>	1				
	<b>CEREVIA caps</b>	1				
	<b>GLURENORM cp 30mg</b>	1	1	1		
	<b>L THYROXINE CHRISTIAENS cp0,025 mg</b>	1				
	<b>L THYROXINE CHRISTIAENS cp0,1mg</b>	1				
	<b>LERCANIDIPINE EG 10mg</b>	1		1		
	<b>MOXONIDINE EG 0,4mg</b>				1	
	<b>PARACETAMOL EG 1gr</b>	1	1	1		
	<b>LORMETAZEPAM 2mg</b>				1/2	
	<b>PROLOPA 250</b>	1/4			1/4	
	<b>SELECTOL 400mg</b>	1/2			1/2	
	<b>SPASMOMEN</b>	1			1	
	<b>TEMESTA EXPIDET 2,5mg</b>				1	
	<b>VIPIDIA 25mg</b>	1/2				
	<b>URI-CRAN FORTE</b>	1				
	<b>CALCIUM en Gel</b>			1		
	<b>D CURE 1x/sem le jeudi</b>					
	MOVICOL SACHET	1				
	<b>TRADONAL ODIS</b>	1				
	<b>MOTILIUM SN</b>					
	<b>PLACEBO (pour dormir)</b>				1	
	<b>DIVERS (sirop, collyre...)</b>					
	<b>GAVISCON ANIS</b>	10ml	10ml	10ml		

Figuur A.41: Origineel document.

Bordered(1.00)

DEBUT	INJECTIONS	8 h	12 h	17 h	20 h	FIN
	MEDICAMENTS Per Os					
	PAROXETINE 20 mg	1				
	FOLAVIT 4mg	1				
	ALDACTONE cp 25mg	1				
	ALLOPURINOL SANDOZ 100mg			1		
	BURINEX LEO cp 5mg	1				
	CEREVISIA caps	1				
	GLURENORM cp 30mg	1	1	1		
	L THYROXINE CHRISTIAENS cp0,025 mg	1				
	L THYROXINE CHRISTIAENS cp0,1mg	1				
	LERCANIDIPINE EG 10mg	1		1		
	MOXONIDINE EG 0,4mg				1	
	PARACETAMOL EG 1gr	1	1	1		
	LORMETAZEPAM 2mg				1/2	
	PROLOPA 250	1/4			1/4	
	SELECTOL 400mg	1/2			1/2	
	SPASMOMEN	1			1	
	TEMESTA EXPIDET 2,5mg				1	
	VIPIDIA 25mg	1/2				
	URI-CRAN FORTE	1				
	CALCIUM en Gel			1		
	D CURE 1x/sem le jeudi					
	MOVICOL SACHET	1				
	TRADONAL ODIS	1				
	MOTILIUM SN					
	PLACEBO (pour dormir)				1	
	DIVERS (sirop, colligre...)					
	GAVISCON ANIS	10ml	10ml	10ml		

Figuur A.42: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

	injections	8h	ith		loo
1	MEDICAMENTS Per Os	None	None	None	None
2	PAROXETINE 20 mg	1	None	None	None
3	FOLAVIT 4mg	1	None	None	None
4	ALDACTONE cp 25mg	=	None	None	None
5	ALLOPURINOL SANDOZ 100mg	None	None	4	None
6	BURINEX LEO cp 5mg	None	None	None	None
7	CEREVISA caps	None	None	None	None
8	GLURENORM cp 30mg	None	None	None	None
9	L THYROXINE CHRISTIAENS cp0,025 mg	None	None	None	None
10	L THYROXINE CHRISTIAENS cp0,1mg	None	None	None	None
11	LERCANIDIPINE EG 10mg	3]	None	None	None
12	MOXONIDINE EG 0,4mg	None	None	None	41
13	PARACETAMOL EG 1gr	1	41	41	None
14	LORMETAZEPAM 2mg	None	None	None	112
15	PROLOPA 250	4a	None	None	14
16	SELECTOL 400mg	12	None	None	12
17	SPASMOMEN	41	None	None	41
18	TEMESTA EXPIDET 2,5mg	None	None	None	41
19	VIPIDIA 25mg	412	None	None	None
20	URI-CRAN FORTE	41	None	None	None
21	CALCIUM en Gel	None	None	41	None
22	D CURE 1x/sem le jeudi	None	None	None	None
23	MOVICOL SACHET	1	None	None	None
25	TRADONAL ODIS	41	None	None	None
26	MOTILIUM SN	None	None	None	None
27	PLACEBO (pour dormir)	None	None	None	4
28	DIVERS (sirop, collyre...)	None	None	None	None
29	GAVISCON ANIS	410ml		10ml	None

## A.22 Document 22

Naam	Inw. 1990	Inw. 2000	Inw. 2006	Inw. 2013	Oppervlakte in km²	Inw./km² in 2013	jaarlijkse groei in % (2012-2013)	Grootstedt/ eerste keer	Deelstaat
Berlijn	3.433.695	3.382.169	3.404.037	3.421.829	891,70	3.837	1,38	1740	Berlijn
Hamburg	1.652.363	1.715.392	1.754.182	1.746.342	755,30	2.312	0,70	1787	Hamburg
München	1.229.026	1.210.223	1.294.008	1.407.836	310,70	4.531	1,41	1854	Beieren
Koelen	953.551	962.884	989.766	1.034.175	405,16	2.553	0,96	1855	Noordrijn-Westfalen
Frankfurt am Main	644.805	648.550	652.016	701.350	248,31	2.824	1,97	1875	Hessen
Stuttgart	579.988	583.674	593.923	604.297	207,35	2.914	1,06	1874	Baden-Württemberg
Düsseldorf	575.794	569.364	577.505	598.686	217,41	2.754	0,84	1882	Noordrijn-Westfalen
Dortmund	599.055	588.994	587.624	575.944	280,71	2.052	0,67	1893	Noordrijn-Westfalen
Essen	626.973	595.243	583.198	569.884	210,30	2.710	0,53	1893	Noordrijn-Westfalen
Bremen	501.219	539.403	547.938	548.547	325,42	1.688	0,38	1875	Bremen
Leipzig	511.079	493.208	506.578	531.562	297,37	1.788	2,06	1871	Saksen
Dresden	490.571	477.807	504.795	530.754	328,31	1.617	1,08	1852	Saksen
Hannover	513.010	515.001	516.343	518.386	204,14	2.539	0,83	1873	Nedersaksen
Neurenberg	493.692	488.400	500.855	498.876	186,37	2.677	0,76	1881	Beieren
Duisburg	535.447	514.915	499.111	486.855	232,80	2.091	0,01	1904	Noordrijn-Westfalen

## Geschiedenis

→ De geschiedenis van Duitsland voor het hoofd artikel over dit onderwerp.

## Vóór de natiestaat

Terwijl in andere streken van West-Europa enkele gecentraliseerde staten ontstonden, zoals Engeland, Frankrijk of Spanje, bleef het Heilige Roomse Rijk, van de Duitse ridders in de middeleeuwen en daarna een soort confederatie (staatenbond), die Duitse keizer had betrekkelijk weinig macht, en de modernisering in de eerste moderne tijd gebaseerd op het niveau van de individuele landen. Vele steden, vooral in het noorden en oosten, waren een conservatieve en zelfstandige deel van het 'Roomse Rijk' tegenover de modernisering achter ten opzichte van de meer verstedelijkt westelijk gelegen streken. De keizer werd door de vorsten van sommige landen gekozen, de keurvorsten. Vanaf de 15e eeuw kwamen de meeste keizers uit het huis Habsburg, dat toen in Oostenrijk regerde. Naast Oostenrijk en Beieren in het zuiden werd later ook Pruisen een van de belangrijkste Duitse staten.

In het tijdperk van de Franse Revolutie en Napoleon (1799-1815) veranderde Duitsland ingrijpend, onder meer door het verkrijgen van het aantal Duitse staten van ruim 350 naar enkele tientallen. Vooral in het zuiden werden enkele middelgrote staten gevormd. Een groot deel van West- en Noord-Duitsland werd door Frankrijk geannexeerd, een ander maaktel deel uit van de Rijnbond, feitelijk een Franse satellietstaat.

Na 1815 verdelen de meeste Duitstalige (maar ook niet-Duitstalige) gebieden de Duitse Bond, die niet als het oude Rijk geen bondstaat maar een staatbond was. Het voorname orgaan van de Bond was de Bondsdagvergadering (ook Bondsdag genoemd), een congres van afgevaardigden van de individuele vorsten. Onder hen was trouws ook de koning van Nederland in zijn functie als groothertog van Luxemburg (dáár ook Limburg). Bij de opkomende liberale en nationale beweging was de Bond weinig geliefd; hij diende in hun oogst voornamelijk als instrument van de vorsten om de persoonlijkheid en de oprichting van politieke partijen te onderdrukken.

In maart 1848, na een nieuwe revolutie in Frankrijk, ontstonden ook onrusten in Duitsland. Bang voor een gewelddadige revolutie beloofden vele Duitse vorsten grondwetten en gekozen parlementen (voor



→ De geschiedenis van Duitsland voor het hoofd artikel over dit onderwerp.

Figuur A.43: Origineel document.

Naam	Inw. 1990	Inw. 2000	Inw. 2006	Inw. 2013	Oppervlakte in km²	Inw./km² in 2013	jaarlijkse groei in % (2012-2013)	Grootstedt/ eerste keer	Deelstaat
Berlijn	3.433.695	3.382.169	3.404.037	3.421.829	891,70	3.837	1,38	1740	Berlijn
Hamburg	1.652.363	1.715.392	1.754.182	1.746.342	755,30	2.312	0,70	1787	Hamburg
München	1.229.026	1.210.223	1.294.008	1.407.836	310,70	4.531	1,41	1854	Beieren
Koelen	953.551	962.884	989.766	1.034.175	405,16	2.553	0,96	1855	Noordrijn-Westfalen
Frankfurt am Main	644.805	648.550	652.016	701.350	248,31	2.824	1,97	1875	Hessen
Stuttgart	579.988	583.674	593.923	604.297	207,35	2.914	1,06	1874	Baden-Württemberg
Düsseldorf	575.794	569.364	577.505	598.686	217,41	2.754	0,84	1882	Noordrijn-Westfalen
Dortmund	599.055	588.994	587.624	575.944	280,71	2.052	0,67	1893	Noordrijn-Westfalen
Essen	626.973	595.243	583.198	569.884	210,30	2.710	0,53	1893	Noordrijn-Westfalen
Bremen	501.219	539.403	547.938	548.547	325,42	1.688	0,38	1875	Bremen
Leipzig	511.079	493.208	506.578	531.562	297,37	1.788	2,06	1871	Saksen
Dresden	490.571	477.807	504.795	530.754	328,31	1.617	1,08	1852	Saksen
Hannover	513.010	515.001	516.343	518.386	204,14	2.539	0,83	1873	Nedersaksen
Neurenberg	493.692	488.400	500.855	498.876	186,37	2.677	0,76	1881	Beieren
Duisburg	535.447	514.915	499.111	486.855	232,80	2.091	0,01	1904	Noordrijn-Westfalen

## Geschiedenis

→ De geschiedenis van Duitsland voor het hoofd artikel over dit onderwerp.

## Vóór de natiestaat

Terwijl in andere streken van West-Europa enkele gecentraliseerde staten ontstonden, zoals Engeland, Frankrijk of Spanje, bleef het Heilige Roomse Rijk, van de Duitse ridders in de middeleeuwen en daarna een soort confederatie (staatenbond), die Duitse keizer had betrekkelijk weinig macht, en de modernisering in de eerste moderne tijd gebaseerd op het niveau van de individuele landen. Vele steden, vooral in het agrarische en conservatieve oostelijke deel van het 'Roomse Rijk' tegenover ontwikkeling achter ten opzichte van de meer verstedelijkt westelijk gelegen streken. De keizer werd door de vorsten van sommige landen gekozen, de keurvorsten. Vanaf de 15e eeuw kwamen de meeste keizers uit het huis Habsburg, dat toen in Oostenrijk regerde. Naast Oostenrijk en Beieren in het zuiden werd later ook

Pruisen een van de belangrijkste Duitse staten.

In het tijdperk van de Franse Revolutie en Napoleon (1799-1815) veranderde Duitsland ingrijpend, onder meer door het verkrijgen van het aantal Duitse staten van ruim 350 naar enkele tientallen. Vooral in het zuiden werden enkele middelgrote staten gevormd. Het groot deel van West- en Noord-Duitsland werd door Frankrijk geannexeerd, een ander maaktel deel uit van de Rijnbond, feitelijk een Franse satellietstaat.

Na 1815 vormden de meeste Duitstalige (maar ook niet-Duitstalige) gebieden de Duitse Bond, die niet als het oude Rijk geen bondstaat maar een staatbond was. Het voorname orgaan van de Bond was de Bondsdagvergadering (ook Bondsdag genoemd), een congres van afgevaardigden van de individuele vorsten. Onder hen was trouws ook de koning van Nederland in zijn functie als groothertog van Luxemburg (dáár ook Limburg). Bij de opkomende liberale en nationale beweging was de Bond weinig geliefd; hij diende in hun oogst voornamelijk als instrument van de vorsten om de persoonlijkheid en de oprichting van politieke partijen te onderdrukken.

In maart 1848, na een nieuwe revolutie in Frankrijk, ontstonden ook onrusten in Duitsland. Bang voor een gewelddadige revolutie beloofden vele Duitse vorsten grondwetten en gekozen parlementen (voor



→ De geschiedenis van Duitsland voor het hoofd artikel over dit onderwerp.

Figuur A.44: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

Jaarlijkse gro) «	Naam =e	2   1990	2000	Inw. 2006	Inw.   2013	Opper
0 1,38	Berlijn	3.433.695	13.382.169	13.404.037	13.421.829	891,7
1 0,70	Hamburg	1.652.363	11.715.392	11.754.182	11.746.342	755,3
2 1,41	Miinchén	1.229.026	11.210.223	11.294.608	11.407.836	310,7
3 0,96	Keulen	953.551]	962.884]	989.766]	1.034.175	405,1
4 1,97	Frankfurt am Mainl	644.865]	648.550	652.610]	701.350	248,3
5 1,06	Stuttgart	None	583.874]	593.923]	604.297	207,3
6 0,84	Dusseldorf	None	569.364]	577.505)	598.686	217,4
7 0,67	Dortmund	599.055]	588.994]	587.624)	575.944	280,7
8 0,53	Essen	626.973]	595.243]	583.198]	569.884	210,3
9 0,38	Bremen	551.219]	539.403]	547.934)	548.547	325,4
10 2,06	Leipzig	511.079]	493.208]	506.578)	531.562	297,3
11 1,08	Dresden	490.571]	477.807]	504.795]	530.754	328,3
12 0,83	Hannover	513.010]	515.001]	516.343)	518.386	204,1
13 0,76	Neurenberg	493.692	488.400]	500.855)	498.876	186,3
14 0,01	Duisburg	None	514.915	499.111]	486.855.	232,8

## A.23 Document 23

**60 FPS Ultra Build**

Last refreshed: June 26th 2020, Version 1.15, latest revision by AuE-Man1864. Provide feedback here.

**About**

A quality mid-range computer in the \$750 range. This build is suitable for playing at "Ultra" graphics settings for most games, on 60 Hz refresh rate. The base build offers 1440p (QHD) resolution, while the augmentation offers an option for 1080p. With the components in this build, it will be able to support VR.

**PCPartPicker Part List**

Type	Item	Price
CPU	AMD Ryzen 5 1600 (12nm) 3.2 GHz 6-Core Processor	\$104.99 @ Amazon
Motherboard	*Gigabyte B450M DS3H Micro ATX AM4 Motherboard	\$77.99 @ B&H
Memory	*Corsair Vengeance LPX 16 GB (2 x 8 GB) DDR4-3200 CL16 Memory	\$30.55 @ Amazon
Storage	*Team T-Force VULCAN 500 GB 2.5" Solid State Drive	\$55.99 @ Amazon
Video Card	*EVGA GeForce RTX 2060 6 GB KO GAMING Video Card	\$303.98 @ Newegg
Case	*Cougar MX330 ATX Mid Tower Case	\$49.99 @ B&H
Power Supply	*Antec High Current Gamer Gold 650 W 80+ Gold Certified Fully Modular ATX Power Supply	\$53.06 @ Amazon
Prices include shipping, taxes, rebates, and discounts		
Total		\$676.55

**Do not forget to overclock the memory, as every memory always runs at JEDEC 2133 MHz by default. Especially on AMD Ryzen computers as they would take a massive hit in performance without the overclock.** Depending on the motherboard, the memory overclock may be called DOCP (Direct Over Clock Profile), EOCOP (Extended Over Clock Profile), or XMP (Extreme Memory Profile).

**Augmentations**

- Don't take our words for the case! Cases are subjective, you can look at any case you want, just make sure the video card fits!
- Not sure if you need an SSD? Head to the FAQ down below
- 2TB HDD for more storage (Add ~\$50)
- Case fans to keep the system cooler (Add ~\$10 per fan, ~\$30 for a 5-pack)

**Benchmarks**

- GTX 1660 Super Benchmark
- RTX 2060 Benchmark

**Rules**

- 6.2 Pictures of purchase or shipping related notifications and information.
- 6.3 Pictures of shipping.
- 6.4 Reception images (unless very high-effort especially original).
- 6.5 Concern trolling.
- 6.6 Feigning idiocy when posting or commenting, or linking to such content.
- Rule #1 Age, nationality, race, gender, ability, sexuality, religion, political affiliation and economic status are all irrelevant. All are welcome to participate in the Master Race.
- Rule #2 No racism, sexism, homophobia, or any other type of discrimination. Any kind of analogies in this regard, is not acceptable in the slightest.
- Rule #3 No begging, No sniffing, No trading, explicit trading, No buying, No selling, No asking for valuation. No asking how or where to do these either.
- Rule #4 Please message the mods first, and wait for permission. If your post/comment is NSFW, Do not mark your post as NSFW or as Spoiler unless it has been approved by a moderator.
- Rule #10 Screenshots ("mashshots") including but not limited to conversations on Reddit, other websites or smishchat screenshots with general text being spewed and/or indicated, must be in text only.

**Giveaway Rules**

For an in-depth explanation of our giveaway rules don't forget to check our [rules wiki page](#).

- The Master Race is a closed group. When giving away please keep it within the bounds of our community. You can't submit members, Feel free to give prizes as you see fit, but don't make it a requirement to follow the rules. You can post about your channel or project, but don't make it a requirement to visit, comment or subscribe to an external place to enter a

Figuur A.45: Origineel document.

The screenshot shows a table of computer components and their prices from a PC build configuration tool. The columns are Type, Item, and Price.

Type	Item	Price
0 CPU	AMD Ryzen 5 1600 (12nm) 3.2 GHz 6-Core Processor	\$104.99 @ Amazon
1 Motherboard	"Gigabyte B450M DS3H Micro ATX AM4 Motherboard	\$77.99 @ B&H
2 Memory	"Corsair Vengeance LPX 16 GB (2 x 8 GB) DDR4-3...	\$30.55 @ Amazon
3 Storage	*Team T-Force VULCAN 500 GB 2.5"Solid State D...	\$55.99 @ Amazon
4 Video Card	*EVGA GeForce RTX 2060 6 GB KO GAMING Video Card	\$303.98 @ Newegg
5 Case	"Cougar MX330 ATX Mid Tower Case	\$49.99 @ B&H
6 Power Supply	\$53 06 @ Amazo!	None
7 Total	\$676.55	None

**Augmentations**

- Don't take our words for the case! Cases are subjective, you can look at any case you want, just make sure the video card fits!
- Not sure if you need an SSD? Head to the FAQ down below
- 2TB HDD for more storage (Add ~\$50)
- Case fans to keep the system cooler (Add ~\$10 per fan, ~\$30 for a 5-pack)

**Components**

- GTX 1660 Super Benchmark
- RTX 3060 Benchmark

**cell0.91**

For an in-depth explanation of our giveaway rules don't forget to check our rules wiki page.

**cell0.99**

The Master Race is generous. When doing giveaways please keep it within the bounds of your channel or project. If you're giving away a prize, feel free to give prizes as you see fit, but don't make it a requirement to follow your channel or project. If you're giving away a prize, it must be related to your channel or project, but don't make it a requirement to visit, comment, or subscribe to an external page to enter.

Figuur A.46: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Type	Item	Price
0 CPU	AMD Ryzen 5 1600 (12nm) 3.2 GHz 6-Core Processor	\$104.99 @ Amazon
1 Motherboard	"Gigabyte B450M DS3H Micro ATX AM4 Motherboard	\$77.99 @ B&H
2 Memory	"Corsair Vengeance LPX 16 GB (2 x 8 GB) DDR4-3...	\$30.55 @ Amazon
3 Storage	*Team T-Force VULCAN 500 GB 2.5"Solid State D...	\$55.99 @ Amazon
4 Video Card	*EVGA GeForce RTX 2060 6 GB KO GAMING Video Card	\$303.98 @ Newegg
5 Case	"Cougar MX330 ATX Mid Tower Case	\$49.99 @ B&H
6 Power Supply	\$53 06 @ Amazo!	None
7 Total	\$676.55	None

Getransformeerde tabellen met algoritme B:

Type	Item	Price
0 cPU	AMD Ryzen 5 1600 (12nm) 3.2 GHz 6-Core Processor	\$104.99 @ Amazon
1 Motherboard	*Gigabyte B450M DS3H Micro ATX AM4 Motherboard	\$77.99 @ B&H
2 Memory	"Corsair Vengeance LPX 16 GB (2 x 8 GB) DDR4-3...	\$30.55 @ Amazon
3 Storage	"Team T-Force VULCAN 500 GB 2.5" Solid State D...	\$55.99 @ Amazon
4 Video Card	"EVGA GeForce RTX 2060 6 GB KO GAMING Video Card	\$303.98 @ Newegg
5 Case	*Cougar MX330 ATX Mid Tower Case	\$49.99 @ B&H
6 Power Supply	*Antec High Current Gamer Gold 650 W 80+ Gold ...	\$53.06 @ Amazon
7 None	Prices include shipping, taxes, rebates, and d...	None
8 None	Total	\$676.55

## A.24 Document 24

Typical Values	100g	1 serving
	Contains	(27g)
<b>Energy</b>	1550kJ 370kcal	418kJ 100kcal
<b>Fat</b>	0g	0g
<b>of which saturates</b>	0g	0g
<b>Carbohydrate</b>	93g	25g
<b>of which sugars</b>	93g	25g
<b>Fibre</b>	0g	0g
<b>Protein</b>	0g	0g
<b>Salt</b>	1.1g	0.3g

Figuur A.47: Origineel document.

Borderless 0.99	cell 0.99	cell 0.99
<b>Typical Values</b>	<b>100g</b>	<b>1 serving</b>
	<b>Contains</b>	<b>(27g)</b>
<b>Energy</b>	1550kJ 370kcal	418kJ 100kcal
<b>Fat</b>	0g	0g
<b>of which saturates</b>	0g	0g
<b>Carbohydrate</b>	93g	25g
<b>of which sugars</b>	93g	25g
<b>Fibre</b>	0g	0g
<b>Protein</b>	0g	0g
<b>Salt</b>	1.1g	0.3g

Figuur A.48: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

	Typical Values	100g	1 serving
0	None	Contains (27g)	
1	Energy	1550kJ	418k)
2	None	370kcal	100kcal
3	Fat	Og	Og
4	of which saturates	Og	Og
5	Carbohydrate	93g	25g
6	of which sugars	93g	25g
7	Fibre	Og	Og
8	Protein	Og	Og
9	Salt	1.1g	0.3g

## A.25 Document 25

### Je bedrijf

Je straat 123  
1234 AB Je plaats  
(+31) 00 000 00 00

## Factuur

Verzonden op 04-09-2020

Factuur voor	Verschuldigd aan	Factuurnr.
Naam	Naam	123456
Bedrijfsnaam		
Bezoekadres	Project	Uiterste betaaldatum
Postcode en plaatsnaam	Projectnaam	16-09-00

Beschrijving	Aantal	Stukprijs	Totaalbedrag
Item 1	1	€ 200,00	€ 200,00
Item 2	2	€ 200,00	€ 400,00
			€ 0,00
			€ 0,00
Opmerkingen:		Subtotaal	€ 600,00
		Aanpassingen	- € 100,00
			€ 500,00

Figuur A.49: Origineel document.

---

<b>Je bedrijf</b>			
Je straat 123			
1234 AB Je plaats			
(+31) 00 000 00 00			
<b>Factuur</b>			
<b>Verzonden op 04-09-2020</b>			
<b>Factuur voor</b>	<b>Verschuldigd aan</b>		
Naam Bedrijfsnaam Bezoekadres Postcode en plaatsnaam	Naam 123456		
	<b>Project</b>		
	Projectnaam 16-09-00		
<b>Uiterste betaaldatum</b>			
<b>Productenlijst</b>			
Beschrijving	Aantal	Stukprijs	Totaalbedrag
Item 3	1	€ 200,00	€ 200,00
Item 2	2	€ 200,00	€ 400,00
			€ 600,00
			€ 600,00
			- € 100,00
		Subtotaal	€ 500,00
		Aanpassingen	

Figuur A.50: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Beschrijving	Aantal	T
0	Item 1	€ 200,00	€ 200,00
1	Item 2	€ 20000	€ 400,00
2	€ 0,00	None	None
3	€0,00	None	None
4	Opmerkingen:	Subtotaal\nAanpassingen	€ 600,00\n€ 100,00
5	€ 500 00	None	None

Getransformeerde tabellen met algoritme B:

	Beschrijving	Aantal	Stukprijs	Totaalbedrag	NaN
0	Item 1	1	€ 200,00	€ 200,00	None
1	Item 2	2	€ 200,00	€ 400,00	None
2	Opmerkingen:	None	Subtotaal	€ 600,00	None
3	None	None	None	€ 0,00	None
4	None	None	None	€ 0,00	None
5	None	None	Aanpassingen	- € 100,00	None
6	None	None	None	500 00	€

## A.26 Document 26

### Ice Cream Preferences

Field	Dislike	Neutral	Like
Pistachio	9	13	4
Vanilla	13	6	7
Strawberry	10	10	6

Figuur A.51: Origineel document.

Field		Dislike	Neutral	Like
	Pistachio	cell 0.99 9	cell 0.99 13	cell 1.00 4
	Vanilla	cell 0.99 13	cell 0.99 6	cell 0.96 7
	Strawberry	cell 1.00 10	cell 0.99 10	cell 0.96 6

Figuur A.52: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

	Field	Dislike	Neutral	Like
0	Pistachio	Qo	13	4
1	Vanilla	13	6	7
2	'Strawberry	10	10	6

Getransformeerde tabellen met algoritme B:

	Field	Dislike	Neutral	Like
0	Pistachio	9	33	4
1	Vanilla	13	6	7
2	Strawberry	10	10	6

**A.27 Document 27**

Compound	Concentration
Protein (g/100g)	1.3
Fat (g/100g)	0.2
Carbohydrate (g/100g)	7.1
Fibre (g/100g)	2.1
Vitamin	
Vitamin C (mg/100g)	7.0
Vitamin E (mg/100g)	0.14
Mineral	
Potassium (mg/100g)	170
Calcium (mg/100g)	25
Magnesium (mg/100g)	10
Phosphorus (mg/100g)	33
Sulfur	50
Iron (mg/100g)	0.3

Figuur A.53: Origineel document.

Compound	Concentration
Protein (g/100g)	1.3
Fat (g/100g)	0.2
Carbohydrate (g/100g)	7.1
Fibre (g/100g)	2.1
Vitamin	
Vitamin C (mg/100g)	7.0
Vitamin E (mg/100g)	0.14
Mineral	
Potassium (mg/100g)	170
Calcium (mg/100g)	25
Magnesium (mg/100g)	10
Phosphorus (mg/100g)	33
Sulfur	50
Iron (mg/100g)	0.3

Figuur A.54: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

	NaN	Concentration
0	Compound Protein (g/100g)	13
1	Fat (g/100g)	0.2
2	Carbohydrate (g/100g)	71
3	Fibre	2.1
4	(g/100g) Vitamin	None
5	Vitamin C (mg/100g)	70
6	Vitamin (mg/100g)	0.14
7	E Mineral	None
8	Potassium (mg/100g)	170
9	Calcium (mg/100g)	25
10	Magnesium (mg/100g)	10
11	Phosphorus (mg/100g)	33
12	Sulfur	50
13	Tron (mg/100g)	03

## A.28 Document 28

### Soorten websites

Tabel 3.5. Inhoud van ".be"-websites (in %) en verandering (in procentpunt)

	2014	2015	Verandering
Bedrijfswebsite	48,0	54,7	6,7
Niet-commerciële website	18,6	18,3	-0,3
Foutmelding	15,4	16,1	0,7
Pay per click	2,1	0,7	-1,4
Personlijke/gezinsblog	4,8	3,2	-1,6
Webshop	3,8	2,1	-1,7
Portaal/media	1,7	0,7	-1,0
Website om te verkopen	2,3	2,5	0,2
Andere	3,3	1,8	-1,5

Bron: DNS Belgium.

**54,7 %** van de domeinnamen ".be" heeft betrekking op websites voor ondernemingen. Er kan een lichte stijging van 6,7 procentpunt waargenomen worden voor deze categorie tussen 2014 en 2015.

64

Tabel 3.6. Inhoud van ".brussels"-websites (in %)

	Inhoud van ".brussels"-websites	2015
Parkeerpagina websites	36,3	
Foutmelding	29,7	
Bedrijfswebsite	13,6	
Parkeerpagina niet-commerciële websites	11,0	
Niet-commerciële website	6,6	
Website om te verkopen	1,0	
Pay-per-click	0,8	
Personlijke/gezinsblog	0,8	
Portaal/media	0,2	

Bron: DNS Belgium.  
**36,3 %** van de domeinnamen met ".brussels" zijn parkeerpagina's van ondernemingen. Actieve bedrijfs-websites vertegenwoordigen 13,6 % van het geheel.

Tabel 3.7. Inhoud van ".vlaanderen"-websites (in %)

	Inhoud van ".vlaanderen"-websites	2015
Foutmelding	33,8	
Parkeerpagina websites	27,4	
Bedrijfswebsite	14,6	
Parkeerpagina niet-commerciële websites	13,0	
Niet-commerciële website	6,6	
Website om te verkopen	2,8	
Personlijke/gezinsblog	1,4	
Pay-per-click	0,2	
Portaal/media	0,2	

Bron: DNS Belgium.

Figuur A.55: Origineel document.

**Soorten websites**

Tabel 3.5. Inhoud van ".be"-websites (in %) en verandering (in procentpunt)  
Bron: DNS Belgium

	2014	2015	Verandering
<b>Bedrijfswebsite</b>	<b>48,0</b>	<b>54,7</b>	<b>6,7</b>
Niet-commerciële website	18,6	18,3	-0,3
Foutmelding	15,4	16,1	0,7
Pay per click	2,1	0,7	-1,4
Persoonlijke/gezinsblog	4,8	3,2	-1,6
Webshop	3,8	2,1	-1,7
Portaal/media	1,7	0,7	-1,0
Website om te verkopen	2,3	2,5	0,2
Andere	3,3	1,8	-1,5

Bron: DNS Belgium.

**54,7 %** van de domeinnamen ".be" heeft betrekking op websites voor ondernemingen. Er kan een lichte stijging van 6,7 procentpunt waargenomen worden voor deze categorie tussen 2014 en 2015.

**36,3 %** van de domeinnamen met "brussels" zijn parkeerpagina's van ondernemingen. Actieve bedrijfs-websites vertegenwoordigen 13,6 % van het geheel.

Tabel 3.6. Inhoud van ".brussels"-websites (in %)  
Bron: DNS Belgium

	2015
<b>Parkeerpagina websites</b>	<b>36,3</b>
Foutmelding	29,7
Bedrijfswebsite	13,6
Parkeerpagina niet-commerciële websites	11,0
Niet-commerciële website	6,4
Website om te verkopen	1,0
Pay-per-click	0,8
Persoonlijke/gezinsblog	0,8
Portaal/media	0,2

Bron: DNS Belgium.

Tabel 3.7. Inhoud van "vlaanderen"-websites (in %)  
Bron: DNS Belgium

	2015
<b>Foutmelding</b>	<b>33,8</b>
Parkeerpagina websites	27,4
Bedrijfswebsite	14,6
Parkeerpagina niet-commerciële websites	13,0
Niet-commerciële website	6,6
Website om te verkopen	2,8
Persoonlijke/gezinsblog	1,4
Pay-per-click	0,2
Portaal/media	0,2

Bron: DNS Belgium.

Figuur A.56: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Foutmelding		33,8
0	Parkeerpagina websites	27,4
1	Bedrijfswebsite	14,6
2	Parkeerpagina niet-commerciële websites	13,0
3	Niet-commerciële website	6,6
4	Website om te verkopen	2,8
5	Persoonlijke/gezinsblog	1,4
6	Pay-per-click	0,2
7	Portal/media	0,2
8	Bron: DNS Belgium	None

Bedrijfswebsite	48.0	54.7	Verandering
0	Niet-commerciële web-	18,6	-0,3
1	Foutmelding	15,4	0,7
2	Pay per click	2,1	1,4
3	Persoonlijke/gezinsblog	48	1,6
4	Webshop	3,8	1,7
5	Portaal/media	1,7	1,0
6	Website om te verkopen	2,3	0,2
7	Andere	3,3	1,5

	Parkeerpagina websites	26,4
0	Foutmelding	29,7
1	Bedrijfswebsite	13,6
2	Parkeerpagina niet-commerciële websites	11,0
3	Niet-commerciële website	6,6
4	Website om te verkopen	1,0
5	Pay-per-click	0,8
6	Persoonlijke/gezinsblog	0,8
7	Portal/media	0,2

Getransformeerde tabellen met algoritme B:

HMM Vall	oN	NaN	NaN
0	None	33,8	Foutmelding
1	websites	27,4	Parkeerpagina
2	None	14,6	Bedrijfswebsite
3	niet-commerciële	13,0	Parkeerpagina
4	website	6,6	Niet-commerciële
5	verkopen	2,8	Website om te
6	None	1,4	Persoonlijke/gezinsblog
7	None	0,2	Pay-per-click
8	None	0,2	Portal/media

verandering	NaN	NaN	NaN	NaN
0	6,7	Bedrijfswebsite	48,0	54,7
1	-0,3	Niet-commercie	18,6	18,3
2	0,7	Foutmelding	15,4	16,1
3	-1,4	Pay per click	2,1	0,7
4	-1,6	Persoonlijke/gezinsblog	4,8	3,2
5	-1,7	Webshop	3,8	2,1
6	-1,0	Portaal/media	1,7	0,7
7	0,2	Website om te	2,3	2,5
8	-1,5	Andere	3,3	1,8

inhoud van	.Drussels	culo	NaN
0 websites	None	36,3	Parkeerpagina
1 None	None	29,7	Foutmelding
2 None	None	13,6	Bedrijfswebsite
3 niet-commerciële	websites	11,0	Parkeerpagina
4 website	None	6,6	Niet-commerciële
5 verkopen	None	1,0	Website om te
6 None	None	0,8	Pay-per-click
7 None	None	0,8	Persoonlijke/gezinsblog
8 None	None	0,2	Portal/media

## A.29 Document 29

Figuur A.57: Origineel document.

MIJN GENEESMIDDELEN							ONTBLUT			MIDDAGMAAL			AVONDMAAL			OPMERKINGEN				
Geneesmiddel	Dosis	Yorm	Frequentie	Inname	Eenheid	Nuchter	Voor	Tijdens	Na	10u	Voor	Tijdens	Na	16u	Voor	Tijdens	Na	20u	Slaap	
vb. Dafagran	500mg	invalstablet	2x dag	orale	tableet															

Figuur A.58: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Tabeltransformatie gefaald.

Getransformeerde tabellen met algoritme B:

MIJN	GENEESMIDDELEN	Nan	Nan	Nan	Nan	Nan
1 'Geneesmiddet	Basis	Yorm	Frequentie	Inname	Eenheid	Nuchter
2 vb. Dafa.gan	sooma —	truistable:	2xttag	cae	table:	None

## A.30 Document 30

Corona Virus as of 29.3.2020 at 16:00h GMT

Column1	Population	Number of Infected	Infected per 100 000	Number of Deaths	Deaths per 100 000	↑↓
Japan	126 000 000	1 693	1,34	52	0,04	
Finland	5 500 000	1 240	22,55	11	0,20	
South Korea	52 000 000	9 583	18,43	152	0,29	
Norway	5 400 000	4 239	78,50	25	0,46	
Germany	84 000 000	58 000	69,05	455	0,54	
USA	327 000 000	124 000	37,92	2 231	0,68	
Austria	8 800 000	8 536	97,00	86	0,98	
Sweden	10 000 000	3 700	37,00	110	1,10	
Denmark	5 800 000	2 400	41,38	72	1,24	
UK	66 000 000	20 000	30,30	1 200	1,82	
Switzerland	8 500 000	15 000	176,47	290	3,41	
France	67 000 000	38 000	56,72	2 300	3,43	
Belgium	11 000 000	10 800	98,18	431	3,92	
Holland	17 000 000	11 000	64,71	770	4,53	
Spain	47 000 000	79 000	168,09	6 500	13,83	
Italy	60 000 000	92 000	153,33	10 000	16,67	

Corona Virus as of 5.4.2020 at 12:00h GMT

Column1	Population	Number of Infected	Infected per 100 000	Number of Deaths	Deaths per 100 000	Change from 29/3/20
Japan	126 000 000	1 693	1,34	57	0,05	10 %
South Korea	52 000 000	10 237	19,69	183	0,35	76 %
Finland	5 500 000	1 927	35,04	25	0,45	56 %
Norway	5 400 000	5 645	104,54	62	1,15	148 %
Germany	84 000 000	97 074	115,56	1 478	1,76	225 %
Austria	8 800 000	11 821	134,33	204	2,32	240 %
USA	327 000 000	311 637	95,30	8 454	2,59	165 %
Denmark	5 800 000	4 077	70,29	161	2,78	152 %
Sweden	10 000 000	6 443	64,43	373	3,73	200 %
UK	66 000 000	41 903	63,49	4 313	6,53	259 %
Switzerland	8 500 000	20 505	241,24	666	7,84	130 %
Holland	17 000 000	16 627	97,81	1 651	9,71	183 %
France	67 000 000	89 953	134,26	7 560	11,28	188 %
Belgium	11 000 000	18 431	167,55	1 283	11,66	158 %
Spain	47 000 000	126 168	268,44	11 947	25,42	84 %
Italy	60 000 000	124 632	207,72	15 362	25,60	54 %
USA Prediction 31-3-2020	327 000 000			240 000	73,39	

Figuur A.59: Origineel document.

Corona Virus as of 29.3.2020 at 16:00h GMT						
Column1	Population	Number of Infected	Infected per 100 000	Number of Deaths	Deaths per 100 000	Change from 29/3/20
Japan	126 000 000	1 693	1,34	52	0,04	
Finland	5 500 000	1 240	22,55	11	0,20	
South Korea	52 000 000	9 583	18,43	152	0,29	
Norway	5 400 000	4 239	78,50	25	0,46	
Germany	84 000 000	58 000	69,05	455	0,54	
USA	327 000 000	124 000	37,92	2 231	0,68	
Austria	8 800 000	8 536	97,00	86	0,98	
Sweden	10 000 000	3 700	37,00	110	1,10	
Denmark	5 800 000	2 400	41,38	72	1,24	
UK	66 000 000	20 000	30,30	1 200	1,82	
Switzerland	8 500 000	15 000	176,47	290	3,41	
France	67 000 000	38 000	56,72	2 300	3,43	
Belgium	11 000 000	10 800	98,18	431	3,92	
Holland	17 000 000	11 000	64,71	770	4,53	
Spain	47 000 000	79 000	168,09	6 500	13,83	
Italy	60 000 000	92 000	153,33	10 000	16,67	

Corona Virus as of 5.4.2020 at 12:00h GMT						
Column1	Population	Number of Infected	Infected per 100 000	Number of Deaths	Deaths per 100 000	Change from 29/3/20
Japan	126 000 000	1 693	1,34	57	0,05	10 %
South Korea	52 000 000	10 237	19,69	183	0,35	76 %
Finland	5 500 000	1 927	35,04	25	0,45	56 %
Norway	5 400 000	5 645	104,54	62	1,15	148 %
Germany	84 000 000	97 074	115,56	1 478	1,76	225 %
Austria	8 800 000	11 821	134,33	204	2,32	240 %
USA	327 000 000	311 637	95,30	8 454	2,59	165 %
Denmark	5 800 000	4 077	70,29	161	2,78	152 %
Sweden	10 000 000	6 443	64,43	373	3,73	200 %
UK	66 000 000	41 903	63,49	4 313	6,53	259 %
Switzerland	8 500 000	20 505	241,24	666	7,84	130 %
Holland	17 000 000	16 627	97,81	1 651	9,71	183 %
France	67 000 000	89 953	134,26	7 560	11,28	188 %
Belgium	11 000 000	18 431	167,55	1 283	11,66	158 %
Spain	47 000 000	126 168	268,44	11 947	25,42	84 %
Italy	60 000 000	124 632	207,72	15 362	25,60	54 %
USA Prediction 31-3-2020	327 000 000		240 000		73,39	

Figuur A.60: Gedetecteerde tabellen.

Getransformeerde tabellen met algoritme A:

Column <sup>c</sup>	Population	DT yao\ndiiiiste cst: a	infected per\nST a	DTT a\nDeaths	[oye
0	126 000 00C	1693	1,34	52	0,04
1	Finland	5 500 000	1240	22,55	11
2	South Korea	52 000 00¢	9583	18,43	152
3	Norway	5S AN0 00c	4239	78,50	25
4	Germany	84 000 ONC	58 000	69,05	ASS
5	USA	377 000 000	124 000	37,92	2231
6	Austria	2 A200 ONC	8536	97,00	86
7	Sweden	10 000 00C	3700	37,00	110
8	Denmark	5S 200 00c	2400	41,38	72
9	UK	66 000 O00C	20000	30,30	1200
10	Switzerland	8 500 OOC	15 000	176,47	290
11	France	67 000 000	38 000	56,72	2300
12	Belgium	11 000 000	10 800	98,18	A31
13	Holland	17000 000	11000	64,71	770
14	47 000 000	79 000	168,09	6500	13,83
15	Italy	60 000 000	92? 000	153,33	10 000

	Population	1693	1,34	57	0,05	10%
0	South Korea	10 237	19,69	183	0,35	76%
1	Finland	1927	35,04	56%	None	None
2	Norway	5 645	104,54	62	1,15	148%
3	Germany	97074	115,56	1478	1,76	225%
4	Austria	11 821	134,33	204	2,32	240 %
5	377 000 000	311 637	95,30	8454	2,59	165%
6	Denmark	4077	70,29	161	2,78	152%
7	Sweden	6 443	64,43	373	3,73	200 %
8	66 000 00C	41903	63,49	4313	6,53	259 %
9	Switzerland	20505	241,24	666	7,84	130%
10	Holland	16 627	97,81	1651	9,71	183%
11	France	89953	134,26	7560	11,28	188 %
12	Belgium	18 431	167,55	1283	11,66	158%
13	Spain	126 168	268,44	11 947	25,42	84%
14	60 000 000	124 632	207,72	15 362	25,60	54%
15	USA Prediction 31-3-2020	240 000	73,39	None	None	None

Getransformeerde tabellen met algoritme B:

	Number of	iInfectedper	Numberof	Deaths per 100	NaN	NaN
0	None	None	None	None	Population	None
1	alae l	el	~l	None	rw	None
2	None	100	a ii	ut	None	None
3	1693	1,34	52	0,04	126 000 000	Japan
4	1240	22,55	11	0,20	5 500 000	Finland
5	9 583	18,43	152	0,29	52 000 000	South Korea
6	4239	78,50	25	0,46	5 400 000	Norway
7	58 000	69,05	455	0,54	84 000 000	Germany
8	124 000	37,92	2 231	0,68	327 000 000	USA
9	8 536	97,00	86	0,98	8 800 000	Austria
10	3 700	37,00	110	1,10	10 000 000	Sweden
11	2 400	41,38	72	1,24	5 800 000	Denmark
12	20 000	30,30	1200	1,82	66 000 000	UK
13	15 000	176,47	290	3,41	8 500 000	Switzerland
14	38 000	56,72	2 300	3,43	67 000 000	France
15	10 800	98,18	431	3,92	11 000 000	Belgium
16	11 000	64,71	770	4,53	17 000 000	Holland
17	79 000	168,09	6 500	13,83	47 000 000	Spain
18	92 000	153,33	10 000	16,67	60 000 000	Italy l

	fet {4	NaN	NaN	NaN	NaN	NaN	NaN
0	Pans	Number of	iInfectedper	Numberof Deaths per 100	None	Population	Columni
1	None	Infected	va	ET) 000	None	None	None
2	None	None	None	[l	Ir	Ir	None
3	10%	1693	1,34	57 0,05	None	126 000 000	Japan
4	76 %	10 237	19,69	183 0,35	None	52.000 000	South Korea
5	56%	1927	35,04	25 0,45	None	5 500 000	Finland
6	None	5 645	104,54	62 [ 115 1	None	5 400 000	Norway
7	225 %	97 074	115,56	1478 1,76	None	84 000 000	Germany
8	240 %	11.821	134,33	204 2,32	None	8 800 000	Austria
9	165 %	311 637	95,30	8454 2,59	None	327 000 000	USA
10	152 %	4077	70,29	161 2,78	None	5 800 000	Denmark
11	200 %	6 443	64,43	373 3,73	None	10 000 000	Sweden
12	259 %	41903	63,49	4313 6,53	None	66 000 000	UK
13	130%	20 505	241,24	666 7,84	None	8 500 000	Switzerland
14	183 %	16 627	97,81	1651 9,71	None	17.000 000	Holland
15	188 %	89 953	134,26	7.560 11,28	None	67 000 000	France
16	158 %	18.431	167,55	1283 11,66	None	11.000 000	Belgium
17	84%	126 168	268,44	11.947 25,42	None	47 000 000	Spain
18	54%	124 632	207,72	15 362 25,60	None	60 000 000	Italy
19	None	None	None	340 000 73.39	None	327000000	USA Predict



## B. Onderzoeksvoorstel

Het onderwerp van deze bachelorproef is gebaseerd op een onderzoeksvoorstel dat vooraf werd beoordeeld door de promotor. Dat voorstel is opgenomen in deze bijlage.

### B.1 Introductie

Het medicatieschema is een geheel van gestandaardiseerde informatie over de actieve medicatie van een patiënt, met inbegrip van de identiteit van de geneesmiddelen, hun dosering, indicatie, relevante gebruiksaanwijzingen en bijkomende informatie waar nodig. Het omvat zowel voorgeschreven als niet-voorgeschreven geneesmiddelen en voedingssupplementen (Apothekersnetwerk, 2013).

Deze oplijsting van de actieve medicatie van de patiënt is niet enkel een essentieel hulpmiddel voor de patiënt bij de correct inname van medicatie maar ook voor medische professionelen om bv. over- of onderdosering, dubbelmedicatie, en andere geneesmiddelgebonden problemen te voorkomen. Ook wordt het gebruikt bij de communicatie tussen zorgverstrekkers. Het medicatieschema wordt eveneens door verpleegsters geraadpleegd voor het klaarzetten van de medicatie.

Dit schema wordt grafisch steeds in tabulaire vorm gepresenteerd. Echter is de lay-out hiervan niet gestandaardiseerd; afhankelijk van de apotheker of andere zorgverstrekker worden andere kolomnamen, kolomverdeling, rand- en verdelingstijl, celgrootte en andere tabelelementen aangewend. Dit bemoeilijkt ernstig het ontwikkelen van een transformatiesysteem die ingescande medicatieschema's omzet in instanties van een uniform digitale

datastructuur in bv. XML- of JSON-formaat voor digitale verwerking van de medicatiedata in gezondheidszorgplatformen.

Hierdoor is er een nood aan een digitalisatiesysteem die medicatieschema's van verschillende vormen en met verschillende lay-outs nauwkeurig omzet in corresponderende instanties van een uniforme datastructuurschema. Voor deze bachelorproef wordt gebruik gemaakt van het datastructuurschema van Into Care by Pridictiv NV. De doelstelling van dit onderzoek is het bestuderen van de mogelijkheden om een dergelijk systeem tot stand te brengen en het implementeren van een proof-of-concept van een optimale oplossing. De volgende onderzoeksvragen kunnen gesteld worden bij dit onderzoek:

- Wat zijn de structuren en de relaties tussen de entiteiten in tabulaire data?
- Wat zijn de uitdagingen en complicaties bij tabelherkenning en -analyse? Kan er meer complexiteit ondervonden worden bij medicatieschematabellen?
- Hoe kan de correctheid en nauwkeurigheid van de transformatie van een tabel geëvalueerd worden?
- Welke oplossingen bestaan er reeds voor tabelherkenning en/of tabelanalyse?
- Wat is de optimale oplossing voor medicatieschema's? Hoe kan deze bepaald worden?
- Hoe kan domeinkennis gebruikt worden om de oplossing te optimaliseren?

## B.2 State-of-the-art

Verschillende oplossingen voor tabeldetectie zijn reeds beschikbaar:

- Vervormbare convolutionele neurale netwerken (Siddiqui e.a., 2018)
- Verticale en horizontale lijndetectie (Gatos e.a., 2005)
- Naïve Bayes en documentstructuur (Li e.a., 2006)

Ook voor tabelanalyse zijn enkele oplossingen voorgesteld:

- Cellsegmentatie (Nazemi e.a., 2016)
- Fast CNN (Oliveira & Viana, 2017)
- Faster R-CNN (Schreiber e.a., 2017)
- Graafgebaseerde neurale netwerken (GNN's) (Qasim e.a., 2019)

## B.3 Methodologie

Het uitvoeren van het onderzoek zal beginnen met het ontwerpen van een scoresysteem, ook wel een benchmarksysteem genoemd, waarbij de nauwkeurigheid, precisie, performantie en andere factoren van de tabelherkenningsoplossingen in rekening gebracht zullen worden.

Hiervoor zullen reeds bestaande geannoteerde, geanonimiseerde medicatieschemadatasets gebruikt worden.

Hierna zullen de verschillende oplossingen geïmplementeerd en tevens geëvalueerd worden a.d.h.v. de benchmarksysteem. De optimale oplossing zal op deze manier bepaald worden.

Verder zullen potentiële optimalisatieopportuniteten bestudeerd worden, zowel algemene optimalisaties als optimatisatimogelijkheden binnen een medisch-farmaceutisch context zoals anomaliedetectie van tijdstippen van medicatieinnamen.

## B.4 Verwachte resultaten

Enerzijds bestaan er in tabellen relaties tussen kolommen en cellen, en relaties tussen cellen onderling die voorgesteld kunnen worden door grafen en anderzijds vertonen de verschillende lay-outs van tabellen een patroon die door het menselijke brein maar dus ook door diepe neurale netwerken zeer snel herkend kan worden. Er wordt daarom verwacht dat een graafgebaseerde Deep Learning-oplossing de best resultaten zal opleveren.

## B.5 Verwachte conclusies

Aangezien zowel state-of-the-art algoritmen als reeds bestaande softwareimplementatieoplossingen beschikbaar zijn, wordt er verwacht dat een performante proof-of-concept van een digitalisatiesysteem voor medicatiesystemen succesvol gecreëerd zal worden. Eveneens wordt er verwacht dat domeinkennis de nauwkeurigheid van het systeem zal verhogen.



# Bibliografie

- Apothekersnetwerk, V. (2013, juli 27). *Standpunt medicatieschema*. <https://vlaamsapothekersnetwerk.be/index.php/informatie/nieuws/8-berichten-van/54-van-standpunt-medicatieschema>
- Cesarini, F., Marinai, S., Sarti, L. & Soda, G. (2002, augustus 11). Trainable table location in document images, In *Object recognition supported by user interaction for service robots*, IEEE Comput. Soc. <https://doi.org/10.1109/icpr.2002.1047838>
- Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., ... Lin, D. (2019). MMDetection: Open MMLab Detection Toolbox and Benchmark.
- Coüasnon, B. & Lemaitre, A. (2014). Recognition of Tables and Forms.
- Embley, D. W., Hurst, M., Lopresti, D. & Nagy, G. (2006). Table-processing paradigms: a research survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(2-3), 66–86. <https://doi.org/10.1007/s10032-006-0017-x>
- Fan, M. & Kim, D. S. (2015). Detecting Table Region in PDF Documents Using Distant Supervision.
- Federale Overheidsdienst Economie, M. e. E., K.M.O. (2019). *Barometer van de informatiemaatschappij (2019)* (onderzoeksrap.). Federale Overheidsdienst Economie, K.M.O., Middenstand en Energie.
- Fedosejev, A. (2016, januari). *React.js Essentials*.
- Financial, A. (2020). *Ant Design*. <https://ant.design/>
- Gatos, B., Danatsas, D., Pratikakis, I. & Perantonis, S. (2005). Automatic Table Detection in Document Images. [https://doi.org/10.1007/11551188\\_67](https://doi.org/10.1007/11551188_67)
- GeeksforGeeks. (2020). *Python | Pandas DataFrame*. <https://www.geeksforgeeks.org/python-pandas-dataframe/>

- Gilani, A., Qasim, S. R., Malik, I. & Shafait, F. (2017, november 9). Table Detection Using Deep Learning, In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE. <https://doi.org/10.1109/icdar.2017.131>
- Gobel, M., Hassan, T., Oro, E. & Orsi, G. (2013, oktober 15). ICDAR 2013 Table Competition, In *12th International Conference on Document Analysis and Recognition*, IEEE. <https://doi.org/10.1109/icdar.2013.292>
- Green, E. & Krishnamoorthy, M. (1996, augustus 14). Model-based analysis of printed tables, In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, IEEE Comput. Soc. Press. <https://doi.org/10.1109/icdar.1995.598979>
- Grinberg, M. (2018). *Flask web development: developing web applications with python*. O'Reilly Media, Inc.
- Hao, L., Gao, L., Yi, X. & Tang, Z. (2016, april 11). A Table Detection Method for PDF Documents Based on Convolutional Neural Networks, In *12th IAPR Workshop on Document Analysis Systems (DAS)*, IEEE. <https://doi.org/10.1109/das.2016.23>
- Jain, N., Bhansali, A. & Mehta, D. (2014). AngularJS: A modern MVC framework in JavaScript. *Journal of Global Research in Computer Science*, 5(12), 17–23.
- Kasar, T., Barlas, P., Adam, S., Chatelain, C. & Paquet, T. (2013, augustus 25). Learning to Detect Tables in Scanned Document Images Using Line Information, In *12th International Conference on Document Analysis and Recognition*, IEEE. <https://doi.org/10.1109/icdar.2013.240>
- Kasar, T., Barlas, P., Sébastien, A., Chatelain, C. & Paquet, T. (2014). Learning to Detect Tables in Scanned Document Images using Line Information. *12th International Conference on Document Analysis and Recognition (ICDAR)*.
- Kay, A. (2007). Tesseract: An Open-Source Optical Character Recognition Engine. *Linux J.*, 2007(159), 2. <https://doi.org/10.5555/1288165.1288167>
- Kieninger, T. & Dengel, A. (2001, september 13). Applying the T-Recs table recognition system to the business letter domain, In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, IEEE Comput. Soc. <https://doi.org/10.1109/icdar.2001.953843>
- Laurentini, A. & Viada, P. (1992, augustus 30). Identifying and understanding tabular material in compound documents, In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, IEEE Comput. Soc. Press. <https://doi.org/10.1109/icpr.1992.201803>
- Lee, M. (2009). *Python-tesseract*. <https://pypi.org/project/pytesseract/>
- Li, J., Tang, J., Song, Q. & Xu, P. (2006). Table Detection from Plain Text Using Machine Learning and Document Structure (X. Zhou, J. Li, H. T. Shen, M. Kitsuregawa & Y. Zhang, Red.). In X. Zhou, J. Li, H. T. Shen, M. Kitsuregawa & Y. Zhang (Red.), *Frontiers of WWW Research and Development - APWeb 2006*, Berlin, Heidelberg, Springer Berlin Heidelberg.
- Long, V. (2010, mei 23). *An Agent-Based Approach to Table Recognition and Interpretation* (proefschrift). Department of Computing, Macquarie University.
- Mandal, S., Chowdhury, S. P., Das, A. K. & Chanda, B. (2006). A simple and effective table detection system from document images. *International Journal of Document Analysis and Recognition (IJDAR)*, 8(2-3), 172–182. <https://doi.org/10.1007/s10032-005-0006-5>

- McKinney, W. (2010). Data Structures for Statistical Computing in Python, In *Proceedings of the 9th Python in Science Conference*, SciPy. <https://doi.org/10.25080/majora-92bf1922-00a>
- Microsoft. (2020). *De opmaak van een Excel-tabel wijzigen*. <https://support.microsoft.com/nl-nl/office/de-opmaak-van-een-excel-tabel-wijzigen-6789619f-c889-495c-99c2-2f971c0e2370>
- Nazari, M. (2020). *Open intelligence*. <https://github.com/nazarimilad/open-intelligence-backend>
- Nazemi, A., Murray, I., Fernaando, C. & McMeekin, D. A. (2016). Converting Optically Scanned Regular or Irregular Tables to a Standardised Markup Format to be Accessible to Vision-Impaired. *World Journal of Education*, 6(5), p9–19. Verkregen 2019, van <https://eric.ed.gov/?id=EJ1158245>
- Oliphant, T. (2006). *A guide to NumPy* (Deel 1). Trelgol Publishing USA.
- Oliveira, D. A. B. & Viana, M. P. (2017, oktober 22). Fast CNN-Based Document Layout Analysis, In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. <https://doi.org/10.1109/ICCVW.2017.142>
- Oro, E. & Ruffolo, M. (2009, juli 26). PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents, In *10th International Conference on Document Analysis and Recognition*, IEEE. <https://doi.org/10.1109/icdar.2009.12>
- Prasad, D., Gadpal, A., Kapadni, K., Visave, M. & Sultanpure, K. (2020, juni 14). Cascade-TabNet: An approach for end to end table detection and structure recognition from image-based documents, In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE. <https://doi.org/10.1109/cvprw50498.2020.00294>
- Pyreddy, P. & Croft, W. B. (1997, juli). TINTIN: A System for Retrieval in Text Tables, In *Proceedings of the second ACM international conference on Digital libraries - DL '97*, New York, NY, USA, ACM Press. <https://doi.org/10.1145/263690.263816>
- Qasim, S. R., Mahmood, H. & Shafait, F. (2019). *Rethinking Table Recognition using Graph Neural Networks*.
- Rashid, S. F., Akmal, A., Adnan, M., Aslam, A. A. & Dengel, A. (2017, november 9). Table Recognition in Heterogeneous Documents Using Machine Learning, In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE. <https://doi.org/10.1109/icdar.2017.132>
- Schreiber, S., Agne, S., Wolf, I., Dengel, A. & Ahmed, S. (2017, november 9). DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images, In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. <https://doi.org/10.1109/ICDAR.2017.192>
- Shahzad, M. A., Noor, R., Ahmad, S., Mian, A. & Shafait, F. (2019). Feature Engineering Meets Deep Learning: A Case Study on Table Detection in Documents, In *Digital Image Computing: Techniques and Applications (DICTA)*.
- Shelhamer, E., Long, J. & Darrell, T. (2016). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 640–651. <https://doi.org/10.1109/tpami.2016.2572683>
- Siddiqui, S. A., Malik, M. I., Agne, S., Dengel, A. & Ahmed, S. (2018). DeCNT: Deep Deformable CNN for Table Detection. *IEEE Access*, 6, 74151–74161. <https://doi.org/10.1109/access.2018.2880211>

- Silva, A. C. (2009, juli 26). Learning Rich Hidden Markov Models in Document Analysis: Table Location, In *10th International Conference on Document Analysis and Recognition*, IEEE. <https://doi.org/10.1109/icdar.2009.185>
- Simonyan, K. & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 1409.1556.
- Tran, D., Aly, T., Oh, A., Kim, S. & Na, I. (2015). Table Detection from Document Image using Vertical Arrangement of Text Blocks. *International Journal of Contents*, 11(4), 77–85.
- Tsjoen, A. (2020). *Medicatieschema*. <https://apotheektsjoen.be/service/medicatieschema/>
- Van Rossum, G. & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA, CreateSpace.
- Wang, X. (1996). *Tabular Abstraction, Editing, and Formatting* (proefschrift). University of Waterloo.
- Wangt, Y., Phillipst, I. T. & Haralick, R. (2001). Automatic table ground truth generation and a background-analysis-based table structure extraction method, In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, IEEE Comput. Soc. <https://doi.org/10.1109/icdar.2001.953845>
- Watanabe, T., Naruse, H., Luo, Q. & Sugie, N. (1991). Structure analysis of table-form documents on the basis of the recognition of vertical and horizontal line segments, In *Intl. Conf. Document Analysis and Recognition*.
- Zanibbi, R., Blostein, D. & Cordy, J. (2003). A Survey of Table Recognition: Models, Observations, Transformations, and Inferences.
- Zeiler, M. D. & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014* (pp. 818–833). Springer International Publishing. [https://doi.org/10.1007/978-3-319-10590-1\\_53](https://doi.org/10.1007/978-3-319-10590-1_53)