Naar consistente voedselproductdata

Hoe kunstmatige intelligentie datakwaliteit kan bewaken en bevorderen

dr. R.J. Vlek, dr. D.J.M. Willems

WAGENINGEN
UNIVERSITY & RESEARCH

Naar consistente voedselproductdata

Hoe kunstmatige intelligentie datakwaliteit kan bewaken en bevorderen

Auteurs: dr. R.J. Vlek, dr. D.J.M. Willems

Dit onderzoek is uitgevoerd door Wageningen Food & Biobased Research in opdracht van Ministerie van Landbouw, Natuur en Voedselkwaliteit en gefinancierd door Ministerie Landbouw, Natuur en Voedselveiligheid, in het kader van Naar consistente voedselproductdata (projectnummer 6234165600).

Wageningen Food & Biobased Research
Wageningen, juni 2019

Openbaar

Rapport 1977
ISBN 978-94-6395-131-9

Versie: definitief

Reviewer: Jan Top

Goedgekeurd door: Henk Wensink

Opdrachtgever: Ministerie van Landbouw, Natuur en Voedselkwaliteit

Financier: Ministerie Landbouw, Natuur en Voedselveiligheid

Financier: Ministerie van Landbouw, Natuur en Voedselkwaliteit

Dit rapport is gratis te downloaden op <https://doi.org/10.18174/500883/> of op www.wur.nl/wfbr (onder publicaties).

© 2019 Wageningen Food & Biobased Research, instituut binnen de rechtspersoon Stichting Wageningen Research.

Het is de opdrachtgever toegestaan dit rapport integraal openbaar te maken en ter inzage te geven aan derden. Zonder voorafgaande schriftelijke toestemming van Wageningen Food & Biobased Research is het niet toegestaan:

- a. dit door Wageningen Food & Biobased Research uitgebrachte rapport gedeeltelijk te publiceren of op andere wijze gedeeltelijk openbaar te maken;
- b. dit door Wageningen Food & Biobased Research uitgebrachte rapport, c.q. de naam van het rapport of Wageningen Food & Biobased Research, geheel of gedeeltelijk te doen gebruiken ten behoeve van het instellen van claims, voor het voeren van gerechtelijke procedures, voor reclame of antireclame en ten behoeve van werving in meer algemene zin;
- c. de naam van Wageningen Food & Biobased Research te gebruiken in andere zin dan als auteur van dit rapport.

Postbus 17, 6700 AA Wageningen, T 0317 48 00 84, E info.wfbr@wur.nl, www.wur.nl/wfbr.
Wageningen Food & Biobased Research is onderdeel van Wageningen University & Research.

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand of openbaar gemaakt in enige vorm of op enige wijze, hetzij elektronisch, hetzij mechanisch, door fotokopieën, opnamen of enige andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever. De uitgever aanvaardt geen aansprakelijkheid voor eventuele fouten of onvolkomenheden.

Inhoud

	Woord vooraf	4
	Samenvatting	5
	Definities	6
1	Inleiding	7
1.1	Achtergrond	7
1.2	Afbakening	7
1.3	Vooruitblik	8
2	Productgegevens	9
2.1	Hoe voedselproductdata gebruikt worden	9
2.2	Definitie van datakwaliteit	10
2.3	Urgentie	10
2.4	Databronnen en datatypen	11
2.5	Kwaliteitsbeheer	11
3	Datakwaliteit	13
3.1	Aard van tekortkomingen in datakwaliteit	13
3.2	Oorzaak van tekortkomingen in datakwaliteit	14
3.2.1	Datamodel	14
3.2.2	Semantiek	15
3.2.3	Definitie van datakwaliteit	15
3.2.4	Datakwaliteitsmanagement	15
4	Technieken en prototypes	17
4.1	Ingrediëntdata	17
4.1.1	Taalherkenning	17
4.1.2	Gebruik van leestekens	18
4.1.3	Segmentatie en ontologische spellingscontrole	19
4.2	Nutriëntdata	21
4.2.1	Logische controles	21
4.2.2	Zelflerende regressiemodellen voor controle en correctie	22
4.3	Productomschrijving en rubricering	24
4.3.1	Zelflerende classificatiemodellen voor controle en correctie	24
5	Aanbevelingen	26
5.1	Datamanagement en beleid	26
5.2	Technieken voor bewaking van datakwaliteit	27
5.2.1	De mogelijkheden	27
5.2.2	De onmogelijkheden	29
6	Conclusie	31
	Bijlage 1 Variaties in schrijfwijze	32
	Bijlage 2 Ingrediënttermen op volgorde van frequentie	33

Woord vooraf

Voedselproductdata wordt op allerlei manieren verwerkt en benut. Dankzij de bereidwillige medewerking van verschillende partijen in het delen van hun kennis en ervaring, was het mogelijk een groot aantal facetten van datakwaliteit te bestuderen, en inspiratie te putten uit reeds individueel ontwikkelde deeloplossingen. Hiervoor gaat dank uit naar het Voedingscentrum, Rijksinstituut voor Volksgezondheid en Milieu (RIVM) en Questionmark. Naast het delen van hun kennis en ervaring werd door GS-1 voor dit project een dataset ter beschikking gesteld, waarmee het mogelijk werd om ontwikkelde concepten in de praktijk te toetsen.

Voor de totstandkoming van de resultaten van dit project gaat bovendien dank uit naar Klaas-Andries de Graaf en Anton Smeenk, die met hun kennis en kunde een substantiële bijdrage leverden aan implementatie van diverse onderzochte technieken.

Samenvatting

Dankzij snelle ontwikkelingen in digitalisering wordt steeds intensiever gebruik gemaakt van digitale voedselproductdata. Nieuwe digitale toepassingen, bijvoorbeeld voor advies over gezonde voeding of duurzame keuzes, stellen steeds zwaardere eisen aan de kwaliteit van deze data. Het bewaken van datakwaliteit volgens de huidige aanpak vraagt veel mankracht, en is mede door die beperking vaak onvolledig. Dit kan soms tot risico's leiden voor consumenten, bijvoorbeeld wanneer informatie omtrent allergenen incorrect is. Voor dit project werd een studie gemaakt van de aard en oorzaak van tekortkomingen in datakwaliteit. Vervolgens werden de mogelijkheden verkend om deze tekortkomingen met behulp van moderne informatietechnieken, zoals kunstmatige intelligentie, te signaleren en (soms) te herstellen.

Uit de resultaten van het project blijkt dat kunstmatige intelligentie (AI) voor verschillende aspecten van datakwaliteit meerwaarde heeft ten opzichte van de huidige manier van werken bij beheer van datakwaliteit. Met name het bewaken van complexere relaties, en dus consistentie, tussen verschillende soorten gegevens is bij uitstek het terrein waarop AI zich goed laat inzetten. Daarbij blijft het echter belangrijk te realiseren dat AI niet in alle gevallen in staat is om absolute zekerheid te geven over datakwaliteit, en soms dus beter kan worden ingezet ter ondersteuning van menselijke beoordeling van data. In dat laatste geval draagt AI nog steeds bij aan reductie van de benodigde mankracht in het beheer van datakwaliteit.

Op het gebied van beleid en datamanagement werden ook een aantal aanbevelingen gedaan. In plaats van het fysieke (papieren) etiket, wordt aangeraden om een goed gecontroleerde digitale tegenhanger hiervan leidend te maken bij de beoordeling van datakwaliteit. Dit zou tevens het deels geautomatiseerd handhaven van wet- en regelgeving omtrent etiketdata kunnen faciliteren. Beheerders van voedselproductdata wordt aangeraden meer aandacht te besteden aan ingangscontroles voor datakwaliteit, verbetering van de gebruikte datamodellen, en standaardisatie van met name meta-data, om zo datakwaliteit in al haar vormen, en het geautomatiseerde beheer daarvan, te bevorderen.

Definities

AI	Artificiële intelligentie, zie KI.
API	Application Programming Interface; een set definities waarmee computerprogramma's onderling kunnen communiceren (zonder dat ontwikkelaars de interne werking van beiden exact hoeven te kennen).
GPC	Global Product Classification. Het systeem waarnaar GS-1 haar producten in categorieën indeelt.
GTIN	Global Trade Identification Number. Een uniek, internationaal erkend nummer waarmee producten geïdentificeerd kunnen worden. Deze code is vaak terug te vinden op de verpakking van een product in de vorm van een streepjescode, maar wordt ook in digitale data bronnen gebruikt om producten te identificeren. De GTIN is door GS-1 ontwikkeld en wordt door hen verstrekt. De GTIN is uniek tot op het niveau van een merkproduct (bijv. huismerk cola) en haar verpakkingsvormen (flesjes, blikjes), maar niet uniek voor elk individueel product in een schap: meerdere flessen cola in een schap met verschillende houdbaarheidsdata bevatten dezelfde GTIN, en kunnen a.d.h.v. de GTIN dus niet van elkaar onderscheiden worden.
KI	Kunstmatige intelligentie. De wetenschap die zich bezighoudt met het creëren van een artefact (bijv. software of robot) dat een vorm van intelligentie vertoont.
ML	machine learning. Een vorm van AI waarbij een computeralgoritme zelfstandig, maar vaak gestuurd met voorbeelden, een patroon in bestaande data leert herkennen.
NEVO	Nederlands Voedingsstoffenbestand. Een database met generieke voedselproducten en hun voedingswaarden. Zie: https://nevo-online.rivm.nl/
POC	Proof Of Concept; een studie waarin het werkingsprincipe wordt aangetoond van een bepaalde techniek.
URI	Uniform resource identifier: een manier om middels een reeks karakters (tekst) de identiteit van een stukje informatie vast te leggen en ernaar te verwijzen. De URI is onderdeel van een internet-protocol waarmee relaties tussen verschillende stukken informatie in een graaf kunnen worden opgeslagen, resulterend in een semantisch web. Het semantisch web wordt door velen gezien als de belangrijkste volgende stap in de ontwikkeling van het internet.
USDA	United States Department of Agriculture

1 Inleiding

1.1 Achtergrond

Kiezen voor gezond voedsel dat op verantwoorde wijze wordt geproduceerd vraagt over het algemeen om bewustwording en gedragsverandering bij de consument. Moderne informatiesystemen kunnen de consument hierbij helpen door de juiste, specifieke kennis en informatie over het voedsel te verstrekken, of impliciet te helpen een gezonde en duurzame keuze te maken. Dit biedt de consument beter inzicht in de consequenties van hun keuze, voor henzelf (onder andere gezondheid) en hun omgeving (onder andere duurzaamheid). Voorwaarde hiervoor is de digitale beschikbaarheid van hoogwaardige productdata, bijvoorbeeld omtrent ingrediënten, voedingswaarden, allergenen en land van oorsprong. In de praktijk blijkt de kwaliteit van deze gegevensbronnen vaak tegen te vallen, waardoor ze niet of slechts beperkt inzetbaar zijn, en soms leiden tot gezondheidsrisico's, bijvoorbeeld in geval van incorrecte allergeneninformatie. Of de kwaliteit van gegevens toereikend is hangt sterk samen met hoe de gegevens in een specifieke toepassing gebruikt (zouden) worden. Met nieuwe ontwikkelingen in toepassingen van voedselproductdata kunnen dus ook nieuwe tekortkomingen geconstateerd worden.

Doel van het project 'naar consistente voedselproductdata' is om aanbevelingen te doen over hoe, met behulp van moderne technieken uit bijvoorbeeld de AI, een hogere datakwaliteit in voedselproductdata bereikt kan worden. Van enkele technieken, en voor de meest urgente tekortkomingen in datakwaliteit, werd een proof-of-concept (POC) uitgewerkt: dat wil zeggen een prototype van een specifieke techniek aan de hand waarvan kon worden beoordeeld of deze techniek geschikt is om datakwaliteit te bewaken of verbeteren. Tijdens de loop van het project hebben een aantal partijen de rol van 'stakeholder' vertegenwoordigd, te weten: GS-1, Voedingscentrum, Rijksinstituut voor Volksgezondheid en Milieu (RIVM) en Questionmark. GS-1 treedt namens een groot aantal producenten van levensmiddelen op als tussenpersoon voor verstrekking van productdata. De andere organisaties zijn maatschappelijk gebruikers van voedselproductdata, en uiteindelijk gericht op de consument. Ten behoeve van het project is door GS-1 een dataset ter beschikking gesteld. Deze werd aangevuld met openbare data van de USDA ten behoeve van het bestuderen van voedingswaardendata (die data waren in de GS-1 dataset niet beschikbaar). Hoewel voor dit project vooral werd gekeken naar data van bovengenoemde bronnen, worden aanbevelingen geformuleerd die zinvol zijn voor iedereen die met voedselproductdata werkt, als verstrekker, tussenpersoon, of afnemer. De inzichten uit dit rapport mogen dan ook door hen ingezet worden om datakwaliteit te bewaken en bevorderen. Tevens bevat dit rapport enkele beleidsmatige aanbevelingen.

Het project vindt plaats in opdracht van, en gefinancierd door het Ministerie van Landbouw, Natuur en Voedselkwaliteit. Het onderzoek werd onafhankelijk uitgevoerd door Wageningen Research. Dit project loopt parallel aan het project "Trusted Source", waarin wordt gekeken hoe informatie door de gehele voedselketen verbeterd en vooral beter gedeeld kan worden. Uiteindelijk is het doel om meer transparantie te bereiken in de keten, zodat consumenten gefundeerde keuzes kunnen maken omtrent gezondheid en duurzaamheid, en bedrijven hun processen verder kunnen optimaliseren.

1.2 Afbakening

Wat bedoelen we als het hebben over voedselproductdata? Uiteraard kan dit gaan over de identiteit van een product (GTIN), verpakkingseenheid, het gewicht, maten, maar ook over de inhoud (ingrediënten en voedingswaarde), bereidingswijze, wettelijk verplichte vermeldingen en waarschuwingen, oorsprong, enzovoorts. Ten behoeve van dit project is een afbakening gemaakt aan de hand van twee belangrijke maatschappelijke doelen: bevordering van gezondheid en

duurzaamheid. In overleg met verschillende maatschappelijk gebruikers¹ die voedselproductdata voor deze doelen inzetten, is vastgesteld welke data van belang zijn, en welke verbeteringen aan datakwaliteit het meest urgent zijn. Hieruit kwam naar voren dat de digitale tegenhanger van productetiket-informatie de meeste aandacht verdient, en vooral informatie omtrent:

- Ingrediënten
- Allergenen
- Voedingswaarden
- Productnaam en -omschrijving
- Productgroep-indeling (niet altijd op etiket aanwezig, wel in digitale bronnen)
- Herkomst (de plaats waar het uiteindelijke product werd gefabriceerd, geoogst of geslacht)

1.3 Vooruitblik

In hoofdstuk 2 en 3 wordt nader ingegaan op de verscheidenheid in gebruik van voedselproductdata en hoe dit leidt tot verschillende eisen aan datakwaliteit, en ernst van de gevolgen van tekortkomingen in data. Verschillende facetten van datakwaliteit worden aan de hand van voorbeelden geïllustreerd.

In hoofdstuk 4 wordt een aantal oplossingen voor het beheer van specifieke aspecten van datakwaliteit in de vorm van een proof-of-concept uitgewerkt. Uit de resultaten hiervan blijkt dat kunstmatige intelligentie (AI) voor verschillende aspecten van datakwaliteit meerwaarde heeft ten opzichte van de huidige manier van werken. Met name het bewaken van complexere relaties, en dus consistentie, tussen verschillende soorten gegevens is bij uitstek terrein waarop AI zich goed laat inzetten. Daarbij blijft het echter belangrijk te realiseren dat AI niet in alle gevallen in staat is om absolute zekerheid te geven over datakwaliteit (maar enkel over een mate van waarschijnlijkheid), en soms dus beter kan worden ingezet ter ondersteuning van menselijke beoordeling van data. In dat laatste geval draagt AI nog steeds bij aan reductie van de benodigde mankracht in het beheer van datakwaliteit.

In hoofdstuk 5 worden aanbevelingen gedaan voor bevordering van de kwaliteit van voedselproductdata. Naast de in hoofdstuk 4 uitgewerkte oplossingen, wordt ook voor diverse andere aspecten van datakwaliteit een technologische oplossing voorgesteld. Op het gebied van beleid en datamanagement werden tevens aanbevelingen gedaan. In plaats van het fysieke (papieren) etiket, wordt aangeraden om een goed gecontroleerde digitale tegenhanger hiervan leidend te maken bij de beoordeling van datakwaliteit. Dit zou tevens het deels geautomatiseerd handhaven van wet- en regelgeving omtrent etiketdata kunnen faciliteren. Beheerders van voedselproductdata wordt aangeraden meer aandacht te besteden aan ingangsc controles voor datakwaliteit, verbetering van de gebruikte datamodellen, en standaardisatie van met name meta-data, om zo datakwaliteit in al haar facetten, en het geautomatiseerde beheer daarvan, te bevorderen.

De resultaten in dit rapport zijn voor een deel tot stand gekomen op basis van specifieke datasets, zoals door GS-1 en USDA beschikbaar gesteld, maar de onderliggende principes zijn in vrijwel alle gevallen generiek, en dus ook geschikt voor toepassing op andere databronnen. Daar waar machine learning werd ingezet om een model te verkrijgen uit voedselproductdata is het belangrijk te realiseren dat omvang en inhoud van de gebruikte dataset mede bepalend zijn geweest voor de bereikte resultaten. Aanvullende machine learning expertise kan noodzakelijk zijn om vergelijkbare resultaten te bereiken met andere datasets.

¹ Dit betreft: Voedingscentrum, Rijksinstituut voor Volksgezondheid en Milieu (RIVM) en Questionmark

2 Productgegevens

2.1 Hoe voedselproductdata gebruikt worden

Het begrip 'datakwaliteit' is een relatief begrip, of data van voldoende kwaliteit zijn wordt hoofdzakelijk bepaald door hoe deze data worden gebruikt. Een voorbeeld scenario:

Wanneer iemand een app zou ontwikkelen die op basis van beschikbare voedselproductdata consumenten met een notenallergie adviseert welke producten zij veilig kunnen eten, heeft incorrecte allergeneninformatie in de gebruikte databron potentieel grote consequenties. Bovendien maakt het uit of de allergeneninformatie automatisch door de app wordt geïnterpreteerd en tot een advies leidt: "eet dit niet" dan wel "dit kun je veilig eten", of dat de allergeneninformatie in zijn oorspronkelijke vorm aan de consument wordt getoond. Waar de consument meestal nog in staat is te begrijpen dat wanneer er "nten" in een product zit, vermoedelijk "noten" bedoeld wordt, is deze interpretatie voor een app minder vanzelfsprekend, wat mogelijk onterecht tot het advies leidt dat het product veilig te eten is. In het laatste geval is de impact van de datakwaliteit een stuk groter.

Behalve in applicaties voor consumentenadvies, zoals in bovengenoemd voorbeeld, wordt voedselproductdata op diverse andere manieren gebruikt, die mede bepalend zijn voor de impact die datakwaliteit heeft. Tabel 1 toont de meest voorkomende vormen van gebruik van voedselproductdata, zoals door de belanghebbenden beschreven, per gebruikersgroep. Met al deze vormen van gebruik in het achterhoofd wordt binnen dit project naar datakwaliteit gekeken.

Tabel 1 Verschillende gebruikers en vormen van gebruik van voedselproductdata

Consumenten	Retailers / groothandels	(Semi-)overheid
<ul style="list-style-type: none">Apps/websites voor advies over, of inzicht in: duurzaamheid, gezondheid (bijvoorbeeld dieet of allergie)Apps/websites voor receptenApps/websites voor online aankopen	<ul style="list-style-type: none">Digitaal etaleren van producten op een website (meestal direct verbonden met verkoop middels webshop).Logistiek, voorraadbeheer, inkoopmanagementRecept suggesties	<ul style="list-style-type: none">Bewaken voedselveiligheid (adequaat handelen bij incidenten, bepalen scope van terugroepacties)Bewaken volksgezondheid (overschrijding toelaatbare consumptienormen, tekorten aan belangrijke nutriënten)Informerend consument omtrent gezondheid / duurzaamheid (bijvoorbeeld de app "Kies ik gezond?" van het Voedingscentrum²)Beleid naar duurzame productie ondersteunen (bijvoorbeeld Questionmark).

In grote lijnen kan worden geconcludeerd dat veel toepassingen zich momenteel nog richten op het verstrekken van ruwe informatie aan een gebruiker. Het interpreteren van de verstrekte informatie wordt daarbij overlaten aan de gebruiker, die daartoe over de benodigde kennis moet beschikken, maar meestal ondanks kleine tekortkomingen in datakwaliteit toch de juiste interpretatie kan maken (een spelfout leidt niet snel tot verkeerde conclusies). Met de snelle toename in hoeveelheid en complexiteit van beschikbare data, wordt deze benadering steeds minder goed houdbaar: de gebruiker ziet door de bomen het bos niet meer, of beschikt niet over voldoende kennis voor een juiste interpretatie van de gegevens. Dat betekent dat slimme software de gebruiker beter zal moeten gaan ondersteunen, door interpretatie van de gegevens (deels) voor haar rekening te nemen. Hiermee

² <https://www.voedingscentrum.nl/nl/thema-s/apps-en-tools-voedingscentrum/kies-ik-gezond-.aspx>

worden hogere eisen gesteld aan de kwaliteit en computerbegrijpelijkheid ('machine-readability') van de gegevens.

2.2 Definitie van datakwaliteit

In eerste instantie denkt men bij een definitie van datakwaliteit vooral aan (in)correctheid van data: spelfouten, verkeerde getallen, vergeten ingrediënten, maar er zijn meer facetten van datakwaliteit.

- *Data actualiteit*: wanneer een product van de markt verdwijnt, of verandert, moet de data actueel blijven. Volgens een databron komt een komkommer uit Spanje, maar geldt dat ook nog voor de komkommers die nu in de winkel liggen, of die ik volgende week ontvang als ik nu online vooruit bestel?
- *Betekenis van data (semantiek)*: data kan verstrekt worden in de vorm van bijvoorbeeld tekst en getallen, maar vaak is er aanvullende informatie (meta-data) nodig om te weten hoe die moeten worden geïnterpreteerd. Wordt een getal in gram of kilogram uitgedrukt? Worden voedingswaarden verstrekt van het product in bereide of juist in onbereide vorm? Worden de voedingswaarden per 100g of een andere maat (15g) uitgedrukt?
- *Ontbrekende data*: informatie die er wel had moeten zijn, maar ontbreekt is voor veel toepassingen disruptief, en helemaal wanneer ontbrekende getallen (onterecht) worden vervangen door een 0. Ook de betekenis van ontbrekende informatie is vaak onhelder. Er staat niets over de hoeveelheid suiker, betekent dit dat er *geen* suiker in het product zit, dat het *onbekend* is hoeveel suiker erin zit, of dat het beter uitkomt om het niet te vermelden?
- *Doublures en strijdigheid*: regelmatig komt een stroom van informatie van verschillende bronnen (producten, distributeurs) ergens samen, en resulteert dit in (bijna-)doublures. Hetzelfde product komt twee maal voor, soms met kleine verschillen. Welke informatie is dan correct?
- *Nauwkeurigheid*: hoe nauwkeurig is de methode waarmee informatie (getallen) verkregen zijn? Wat is de omvang van de willekeurige en systematische fout van de gebruikte methode? Hoeveel spreiding zit er tussen verschillende instanties van hetzelfde product (hoe variabel is het productieproces)?

Wetgeving stelt dat levensmiddelenproducenten zelf verantwoordelijk zijn voor naleven van regelgeving en voorschriften³ omtrent het (fysieke) productetiket, en vraagt van hen dat digitale data die zij beschikbaar maken over hetzelfde product in lijn is (en dus niet strijdig) met dit etiket. De consistentie tussen etiketinformatie en haar digitale tegenhanger is dus een vorm van datakwaliteit. Voor dataverstrekker GS-1, die namens producenten digitale data beschikbaar maakt, is dit de belangrijkste vorm van datakwaliteit en deze wordt actief bewaakt. Dit garandeert dat de digitale data correct is, aangenomen dat de informatie op het etiket ook correct is (de verantwoordelijkheid hiervoor ligt bij de producent). Deze aanname is een kwetsbaarheid in de huidige manier van werken, want blijkt niet altijd terecht.

2.3 Urgentie

De gevolgen van inconsistente of incorrecte voedselproductdata verschillen per toepassing, maar zijn in een paar grote lijnen samen te vatten. De gevolgen voor consumenttoepassingen bestaan met name uit desinformatie, verwarring en in enkele gevallen gezondheidsrisico's (bijvoorbeeld omtrent allergenen). Desinformatie maakt het voor consumenten vooral lastig om de juiste keuzes te maken wanneer ze gezondheid of duurzaamheid nastreven, of wanneer de overheid dit gedrag wil bevorderen. Voor retailers zullen de consequenties met name financieel merkbaar zijn, bijvoorbeeld vanwege producten die verkeerd weergegeven worden in hun webshop, of vanwege disruptieve problemen met logistiek, voorraadbeheer of inkoopmanagement, en soms via reputatieschade bij ernstige incidenten. Tekortkomingen in datakwaliteit maken het voor overheden lastig om snel de

³ EU verordening nr. 1169/2011 speelt hiertoe de meest prominente rol, maar ook andere wet- en regelgeving is van toepassing. Opmerkelijk hierin is de aandacht voor de vorm waarin informatie wordt verstrekt, en ondergeschikte rol voor de totstandkoming en correctheid van de op het etiket gepresenteerde gegevens.

juiste beslissingen te maken in geval van incidenten met voedselveiligheid, of om belangrijke trends in volksgezondheid op tijd te registreren. Daarnaast heeft de structuur van een keten invloed op de mate waarin tekortkomingen in datakwaliteit vroeg in de keten zich voortplanten tot verderop in de keten: een tomaat kan langs allerlei wegen, via een verscheidenheid aan samengestelde producten, zoals pizza of pastasaus, uiteindelijk bij de consument terecht komen.

Met de komst van slimme applicaties die data automatisch verwerken en interpreteren worden de gevolgen van tekortkomingen in datakwaliteit echter groter en soms disruptief. De urgentie varieert per belanghebbende (retailer, consument, (semi-)overheid), maar een belangrijke gemene deler werd gevonden in de informatie op het productetiket, en dan met name: allergenen, ingrediënten, nutriënten (voedingswaarden), productomschrijving. Ook werd aangemerkt dat de indeling van een product in de, voor specifieke toepassingen, juiste productcategorie van groot belang is. Op deze vormen van data werd voor het vervolg van dit project dan ook de nadruk gelegd.

2.4 Databronnen en datatypen

Op het gebied van voedselproductdata verstrekken een aantal grote partijen (namens levensmiddelenproducenten) digitale data. Dit zijn onder meer: GS-1, Brandbank, PSinFoodService. Daarnaast verstrekken sommige grotere retailers (o.a. Aldi, Lidl, Albert Heijn) ook zelf digitale voedselproductdata. Van consumenten komen ook steeds meer voedselproductdata beschikbaar, die door hen zelf worden verzameld, vaak met behulp van een app die de etiketten van een producten kan lezen, en soms handmatig ingevoerd⁴. Vooral wanneer informatie handmatig ingevoerd wordt door een groot aantal verschillende consumenten is het de vraag hoe betrouwbaar en volledig deze is.

Verskillende typen data kunnen worden gebruikt om voedselproducten te beschrijven. Naam, merk, omschrijving en ingrediënten worden over het algemeen als tekstveld opgeslagen. In sommige gevallen worden ook nutriënten als tekst vastgelegd, maar in de meeste gevallen betreft het numerieke data die per individueel nutriënt worden vast gelegd. De eenheden waarin deze getallen zijn uitgedrukt verschillen per databron. Voor het beschrijven van het land van oorsprong, de taal waarin data is uitgedrukt, of de categorie waartoe een product behoort worden vaak codes gebruikt. Helaas zijn dit soort codes meestal niet tussen verschillende databronnen geharmoniseerd, maar per databron gedefinieerd.

Voor dit project werd samengewerkt met GS-1, een van de grootste nationale (en internationale) spelers als het gaat om verstrekking van voedselproductdata. Ten behoeve van dit onderzoek werd door hen een dataset beschikbaar gesteld met gegevens van 118702 levensmiddelen, waaronder gegevens over ingrediënten, allergenen, productomschrijving, en rubricering⁵. Ter identificatie van producten in deze dataset werd een substituut-GTIN gebruikt, om anonimiteit van individuele producenten te garanderen. Om het werkingsprincipe van algoritmes voor nutriëntdata aan te tonen werd gebruik gemaakt van een openbare dataset van de USDA⁶ met data van 239095 merkartikelen.

2.5 Kwaliteitsbeheer

Omdat tekortkomingen in datakwaliteit meestal aan het licht komen bij specifieke toepassingen, met data die langs allerlei wegen beschikbaar zijn gekomen, is het niet verwonderlijk dat bij de ontwikkelaars van die toepassingen vaak de eerste maatregelen worden genomen om datakwaliteit te bewaken of verbeteren. Dit leidt ertoe dat veel partijen elk hun eigen methodes ontwikkelen om kwaliteit van deels dezelfde data te beheren, en niet profiteren van elkaars inspanningen. Wanneer datakwaliteit dicht bij de oorsprong wordt bewaakt hebben alle afnemers hiervan profijt.

⁴ Zie bijvoorbeeld het initiatief Open Food Facts: <https://world.openfoodfacts.org/>

⁵ GS-1 gebruikt het GPC-systeem (global product classification) om hun producten naar in te delen

⁶ De openbare merkproductdata van USDA is te vinden via: <https://ndb.nal.usda.gov/ndb/search/list>

De manier waarop datakwaliteit wordt beheerd hangt vooral samen met de beoogde toepassingen van die data, maar wordt ook bepaald door andere factoren, waaronder de beschikbare (domein-)kennis over data (kent men de relatie tussen nutriënten?) en data-analyse-technieken, de technologische infrastructuur, en beschikbare mankracht. Automatisering lijkt de ideale oplossing om tekort aan mankracht te overkomen, maar is alleen zinvol indien er voldoende vertrouwen is dat volledige automatisering in alle gevallen tot de juiste beslissingen omtrent kwaliteit leidt. In veel gevallen is uiteindelijk toch nog een menselijk eindoordeel gewenst, en in die gevallen is een semi-automatische benadering de meest effectieve: automatisering wordt gebruikt om te zorgen dat de beschikbare mankracht optimaal wordt benut. Bepaalde tekortkomingen in datakwaliteit kunnen zonder twijfel betrouwbaar gedetecteerd (en soms ook gecorrigeerd) worden door een algoritme, en zijn dus uitstekend automatiseerbaar. In andere situaties kan een algoritme geen eindoordeel vellen, maar behulpzaam zijn door de meest onwaarschijnlijke datapunten naar boven halen, zodat die als eerste aan een menselijk eindoordeel onderworpen kunnen worden (mankracht wordt zo effectiever benut). Soms is het zelfs mogelijk dat een algoritme voor verdachte datapunten al suggesties doet over wat de correcte waarde had moeten zijn. Elk menselijk oordeel over een verdacht datapunt leidt bovendien tot meer kennis en inzicht over datakwaliteit, en zou benut kunnen worden om de algoritmes in het geautomatiseerde gedeelte verder te versterken. Zo verschuift geleidelijk de balans in semi-geautomatiseerde systemen, met een groeiende rol voor de computer en een steeds kleinere voor de mens.

3 Datakwaliteit

3.1 Aard van tekortkomingen in datakwaliteit

Zoals in hoofdstuk 2 duidelijk werd is datakwaliteit een breed begrip met veel facetten. Om op een systematische manier te werk te kunnen gaan hebben we onderscheid gemaakt tussen een aantal manieren om naar datakwaliteit te kijken⁷. Deze werden afzonderlijk bestudeerd.

Intrinsieke kwaliteit: de kwaliteit van een individueel dataveld blijkt uit de inhoud van het veld zelf, eventueel tegen het licht van algemene kennis over bijvoorbeeld de Nederlandse taal, logica of basale context (grootheid, eenheid).

Voorbeelden:

- Spelfouten, zoals "aardoppel" in plaats van "aardappel". Zie Bijlage 1 voor een illustratie van de diversiteit in (foutieve) schrijfwijzen van enkele specifieke woorden.
- Andere syntactische fouten, zoals "Ingrediënten: Aardappel. Ui. Conserveermiddelen: (E-200,, E-221, zout. " (verkeerd gebruik punten en komma's, ontbrekend haakje sluiten).
- Ontbrekende informatie (leeg veld, waar wel verplicht of gewenst)
- Onlogische informatie, zoals: "Ingrediënten: Pinda, olie, verpakkingsmiddel, zout." (vermoedelijk wordt verdikkingsmiddel bedoeld)
- Onlogische getallen, zoals: 120 gram eiwit per 100 gram product
- Onwaarschijnlijke getallen, zoals: 100000 kCal energie.

Relationele kwaliteit: de kwaliteit van een individueel dataveld blijkt pas in relatie tot andere datavelden.

Voorbeelden:

- Ingrediënten- en allergenendeclaratie inconsistent: (noten staan als ingrediënt vermeld, maar ontbreken in de declaratie van allergenen).
- Inconsistentie tussen nutriënten: een product bevat 80 gram koolhydraten per 100 gram, maar slechts 4kJ energie per 100 gram (de hoeveelheid energie is niet consistent met die van de koolhydraten). Of een product bevat 10 gram dierlijke vetten, 20 gram plantaardige vetten, en slechts 15 gram voor het totaal aan vetten (de som van delen die samen het totaal aan vetten opmaken is te groot).
- Onlogische set nutriënten: een product bevat 80 gram vet, 17 gram koolhydraten, en 20 gram eiwit per 100 gram product (de som van nutriënten is groter dan wat logischerwijs mogelijk is).
- Uitbijter in de groep: in de categorie zoutjes, nootjes en chips staat een product dat vruchtensap en veel suiker bevat, en waarvan de verpakkingseenheid in liters wordt uitgedrukt. Dat is erg onrepresentatief voor de andere producten in deze groep, waarvan geen enkele vruchtensap bevat, slechts een enkeling suiker, en waarvan de verpakkingseenheid consequent in grammen wordt uitgedrukt.

Semantische kwaliteit: de kwaliteit van een individueel dataveld blijkt uit de informatie (of afwezigheid daarvan) over de betekenis van het veld.

Voorbeelden:

⁷ Op theoretische gronden is het mogelijk om een veel genuanceerder onderscheid te maken, met een striktere scheiding op basis van een aantal zorgvuldig gekozen definities. Met de beoogde doelgroep van dit rapport in het achterhoofd is echter gekozen voor een gesimplificeerde indeling.

- Eenheden ontbreken: het is niet duidelijk of een getal in gram, kilogram of milliliter is uitgedrukt
- Betekenis lege velden en 0 niet vastgelegd: betekent een ontbrekende waarde voor de hoeveelheid suiker dat het niet in het product zit? Of dat niet bekend is of het in het product aanwezig is?
- Onduidelijke synoniemen: ascorbinezuur, vitamine C en E300 verwijzen naar dezelfde stof⁸, maar weet de consument dit?
- Nutriënten en bereiding: worden nutriënten vermeld van het product in onbereid of bereide toestand? Indien voor het bereide product: volgens welk recept wordt het dan bereid?
- Informatie in verkeerde veld: "Ingrediënten: suiker, water" staat in het veld van allergenen.

Tot slot spelen ook de volgende onderwerpen een rol bij de beoordeling van datakwaliteit:

- *Doublures* of bijna-doublures (ontstaan met name als data langs verschillende routes in dezelfde bron komen). Vooral de bijna-doublures zijn een probleem, omdat daardoor soms niet meer duidelijk is welke van de twee varianten correct is.
- *Actualiteit*: met name voor seizoensgroente en -fruit problematisch, evenals voor wijzigingen in producten of nieuwe/vervallen producten⁹.
- *Provenance (herkomst)*: hoe is de data tot stand gekomen? Betreft het resultaten van lab-analyses, van modelmatige berekeningen of zijn gegevens overgenomen van soortgelijke producten (NEVO tabel).
- *Nauwkeurigheid*: hoeveel variatie zit er tussen potjes van hetzelfde product in de gedeclareerde hoeveelheid eiwit? Hoe variabel is het productieproces? Wat waren de willekeurige en systematische fout van het eventuele meetsysteem waarmee de waarden tot stand kwamen?

3.2 Oorzaak van tekortkomingen in datakwaliteit

3.2.1 Datamodel

Wie data wil opslaan in een database moet keuzes maken over een zogenaamd datamodel: de mal waarin de verschillende soorten informatie gegoten (mogen) worden. Voor de invoer van nutriënten in een database zou je bijvoorbeeld kunnen kiezen voor een tekstveld waarin iemand met tekst kan beschrijven welke nutriënten in welke mate aanwezig zijn in het product. Het zou ook anders kunnen. Het data model zou ook voor elk belangrijk nutriënt een kolom kunnen bevatten waar door de gebruiker een getal (de hoeveelheid) ingevoerd moet worden. Vergeleken met het tekstveld is deze optie minder vrij, en dankzij dit 'strakkere' datamodel is het bijvoorbeeld al niet meer mogelijk dat de gebruiker per ongeluk de ingrediëntendeclaratie in plaats van de nutriënten invoert. De user interface (UI), de plek waar gebruikers gegevens invoeren, helpt om de invoer van gegevens in goede banen te leiden en het datamodel af te dwingen. Een goed datamodel ondersteunt dus een goede datakwaliteit. Een groot aantal tekortkomingen in, met name intrinsieke en relationele datakwaliteit, zijn gevolg van de grote vrijheid in huidige datamodellen. De keuze voor dit soort datamodellen komt vooral voort uit behoefte naar flexibiliteit, nodig om met allerlei praktische variaties om te gaan.

Een te strak datamodel heeft ook nadelen. Voor marketingdoeleinden kan het bijvoorbeeld van belang zijn dat een fabrikant de vrijheid houdt om te kiezen uit de bewoording van ingrediënten. De een zal het liever over vitamine C hebben (wat beter te begrijpen is voor de consument), terwijl de andere al

⁸ Equivalentie van deze drie naamgevingen geldt alleen in de context van een ingrediëntendeclaratie, omdat het in die context altijd een (kunstmatige) toevoeging van de stof betreft. Vitamine C die van nature voorkomt in andere ingrediënten wordt niet apart vermeldt. Buiten de context van een ingrediëntendeclaratie biedt de naamgeving E300 dus wel extra informatie: het maakt expliciet dat het om een (kunstmatige) toevoeging gaat, terwijl het bij 'vitamine C' onduidelijk blijft of het om de van nature voorkomende stof gaat, of om een toevoeging.

⁹ Wijzigingen aan producten en dus verandering van data (bijvoorbeeld bij vernieuwde receptuur) vereisen volgens de definities van GS-1 in sommige gevallen een nieuwe GTIN, maar niet altijd. Juist in de gevallen dat geen nieuwe GTIN verstrekt wordt is het lastig om de identiteit, en dus ook actualiteit, van verschillende versies van hetzelfde product in het oog te houden.

veel informatie op het label heeft staan en liever het kortere E300 gebruikt. Een goed datamodel is dus een balans tussen vrijheid (ten koste van datakwaliteit) en inperking (ten gunste van datakwaliteit). Moderne technieken als 'Semantic Web' (RDF/OWL) zijn voor dit soort toepassingen zeer geschikt. Ze zijn flexibel, maar zonder grote concessies te hoeven doen aan onderliggend datamodel. Zo kan de identiteit van een stof op een eenduidige manier in een database worden vastgelegd (met een zogenaamde URI), terwijl tegelijkertijd de vrijheid wordt behouden om met verschillende benamingen (labels zoals E300, ascorbinezuur of vitamine C) naar deze stof te verwijzen.

3.2.2 Semantiek

Een database is vrijwel onbruikbaar zonder kennis over de betekenis van de gegevens erin (bijvoorbeeld over de eenheid waarin getallen worden uitgedrukt). Problemen met semantische datakwaliteit komen vaak voort uit het feit dat in de database zelf onvoldoende informatie (meta-data) is vastgelegd over de betekenis van de datavelden. Dit vloeit voort uit de beperkingen van de momenteel meeste populaire *relationele database* techniek, waarbij de betekenis van gegevens in tabelkolommen wordt vastgelegd in de kop van de tabel (header), maar niet kan worden vastgelegd hoe de informatie in betreffende kolom samenhangt met andere stukken informatie. Deze informatie is uiteraard meestal wel bekend bij de eigenaar van de database, en wordt dan meestal via andere kanalen gecommuniceerd (mondeling, via een website, etc.). Deze scheiding tussen inhoud en betekenis maakt de kans op fouten groter. Ook de vorm waarin meta-data wordt gecommuniceerd is zelden uniform, de een verwijst naar de gewichtseenheid gram als "g", de ander als "gr." of schrijft het voluit als 'gram'. Een gestandaardiseerde aanpak in het verstrekken van meta-data zou datakwaliteit en toegankelijkheid van data enorm verbeteren. Dan kan men denken aan eenduidige manieren om te verwijzen naar valuta, geografische locaties, eenheden, grootheden, etc. Ook dit principe wordt door Semantic Web ondersteund, en een mogelijke manier om te verwijzen naar eenheden en grootheden werd al eerder ontwikkeld¹⁰.

Wanneer invoer van gegevens plaatsvindt is het tevens van groot belang dat de betekenis van de velden naar de gebruiker wordt gecommuniceerd. Gegevensvalidatie kan aanvullend gebruikt worden om misinterpretaties van de betekenis van een veld tijdens invoer te ondervangen.

3.2.3 Definitie van datakwaliteit

De definitie van datakwaliteit kan verschillen per organisatie, en is typisch afhankelijk van een specifieke toepassing; het gaat om de geschiktheid van data voor die toepassing. Omdat deze definitie bepalend is voor het databasebeleid van de organisatie, heeft ze dus ook directe gevolgen voor wat door de gebruikers van deze data als 'datakwaliteit' ervaren. Met name wanneer gebruikers de data anders gebruiken dan de organisatie heeft voorzien worden tekorten in datakwaliteit ervaren.

3.2.4 Datakwaliteitsmanagement

De processen die door een organisatie worden gebruikt om datakwaliteit te bewaken leunen nog sterk op menselijke controles. Dit is niet per definitie slecht, maar onder druk van beperkte tijd en financiële middelen, moeten vaak keuzes worden gemaakt die direct gevolgen hebben voor datakwaliteit. Ook het moment waarop datakwaliteit bewaakt wordt is bepalend voor de effectiviteit van het proces. Momenteel wordt datakwaliteit vaak laat in de keten bewaakt, vlak voordat de data in een specifieke toepassing wordt benut. Wanneer er nu problemen optreden met de kwaliteit is het pad voor terugkoppeling van de bevindingen naar de oorspronkelijke verstrekker van de data erg lang, en de kans op herstel van datakwaliteit klein. Bewaking dicht bij de bron van data zou daarom verbeteringen moeten opleveren, en voorkomt tegelijkertijd dat foutieve versies van data al verspreid

¹⁰ Publicatie: H. Rijgersberg, M. v. Assem, and J. L. Top, "Ontology of Units of Measure and Related Concepts," Semantic Web, vol. 4, pp. 3-13, 2013. Broncode: <https://github.com/HajoRijgersberg/OM>.

worden voordat ze gecontroleerd zijn. De vraag is alleen of in dit stadium alle mogelijke relevante aspecten van datakwaliteit te overzien zijn (soms blijken die pas bij toepassing van de data).

4 Technieken en prototypes

In dit hoofdstuk laten we zien dat het in verschillende specifieke gevallen mogelijk is tekortkomingen in datakwaliteit automatisch te detecteren en (soms) te corrigeren. We doen dit voor drie typen data: ingrediënten, nutriënten, rubricering (categorie-indeling). Om de prestaties van een nieuw algoritme te kunnen beoordelen is het nodig om de daadwerkelijke kwaliteit van een dataset vooraf te kennen. Deze zogenaamde 'ground truth' dient als ijkpunt om de voorspellingen van een algoritme over dezelfde data mee te vergelijken. Op die manier kan men iets zeggen over het aantal terecht herkende problemen met datakwaliteit, maar ook het aantal vals positieve en vals negatieve resultaten van het algoritme. De rol van 'ground truth' verschilt echter per techniek:

- De definitie van 'goede data' die het algoritme voor detectie en (soms) correctie gebruikt is identiek aan de definitie van de 'ground truth'. Hierdoor weten we zeker dat wat het algoritme als tekortkoming in datakwaliteit aangeeft altijd terecht is. Dit geldt bijvoorbeeld voor detectie van het ontbreken van een haakje sluiten ')' wanneer reeds een haakje open "(" is geconstateerd.
- Het algoritme leert zelf een definitie van 'goede data' die tegen het licht van een 'ground truth' kan worden beoordeeld. Deze 'ground truth' moet eerst worden verkregen, want is in geen van beide datasets vooraf aanwezig. Via het steekproefsgewijs handmatig beoordelen van data kon in een aantal gevallen een 'ground truth' worden verkregen. De beperkte omvang van de steekproef geeft altijd enige onzekerheid omtrent de uiteindelijke resultaten.
- Het algoritme leert zelf een definitie van 'goede data', maar deze definitie is statistisch van aard, en brengt slechts datapunten aan het licht die relatief gezien afwijken t.o.v. de overige beschikbare datapunten. Het vaststellen van een 'ground truth' in deze situatie is niet altijd mogelijk, bijvoorbeeld omdat het om complexe relaties gaat waarover beperkte kennis beschikbaar is.

4.1 Ingrediëntdata

De meeste bronnen met voedselproductdata hanteren een vrij tekstveld (string) voor de ingrediënten in een product. Gevolg van de invoervrijheid van een dergelijk veld is dat de ingrediënten soms in een andere taal dan de gewenste worden ingevoerd. Ook komen, net als bij andere vormen van handmatig ingevoerde tekst, spelfouten en andere syntactische fouten regelmatig voor. Een menselijke lezer is vaak nog wel (maar soms ook niet) in staat om door de onvolkomenheden in een ingrediëntveld heen te kijken, en ondanks fouten te begrijpen wat er bedoeld wordt. Bij automatische verwerking is dit lastiger, en zijn het met name de syntactische fouten die de grote consequenties hebben. De belangrijkste informatie uit de ingrediëntendeclaratie komt namelijk pas beschikbaar wanneer segmentatie van individuele ingrediënten heeft plaatsgevonden. Syntactische problemen staan segmentatie soms in de weg, waardoor belangrijke informatie niet ontsloten kan worden. Voor de verkenning van technieken voor dit type data werd de dataset van GS-1 gebruikt.

4.1.1 Taalherkenning

Moderne taalherkenningsalgoritmen zijn in staat om de taal te voorspellen waarin een stuk tekst geschreven is. Ze kunnen daarom ingezet worden om te detecteren wanneer een ingrediëntenveld (of ander tekstueel veld) niet in de juiste taal is ingevoerd. Indien gewenst kunnen ze ook voorspellen welke taal er dan wel is ingevoerd. Voor de Nederlandse markt is het wenselijk tekstuele productinformatie in het Nederlands beschikbaar te hebben. We hebben ingrediëntvelden van alle

71942 beschikbare Nederlandse voedselproducten¹¹ uit de dataset van GS-1 door een taalherkenningsalgoritme¹² laten analyseren. In 9090 gevallen voorspelde het algoritme dat het een andere taal dan de Nederlandse betrof. Bij een steekproef van deze 9090 voorspellingen bleek het algoritme in sommige gevallen moeite te hebben de juiste taal vast te stellen, vooral bij korte ingrediëntenlijsten, en wanneer verschillende talen hetzelfde woord gebruiken voor een bepaald ingrediënt (ambigüiteit). Als oplossing werd een 'override' optie toegevoegd, die op basis van een lijst met taal-ambigue woorden¹³ de voorspellingen van het algoritme corrigeerde (aangenomen dat bij twijfel, de taal eerder Nederlands dan een andere zal zijn). Wanneer mag worden aangenomen dat een verkeerde taal consequent wordt gehanteerd over meerdere datavelden tegelijk zouden deze ook tezamen aan het algoritme verstrekt kunnen worden. Hiermee ontvangt het algoritme meer informatie, waardoor de betrouwbaarheid van de voorspellingen verder toeneemt.

Om te kunnen vaststellen in hoeverre het taalherkenningsalgoritme terecht aangeeft dat het om een andere taal ging is een willekeurige steekproef van 200 producten genomen. Voor deze 200 producten werd handmatig beoordeeld of de ingrediëntendeclaratie in het Nederlands was vermeldt (of niet). Bijvoorbeeld:

Ingrediëntendeclaratie	Taal NL?
Ingrediënten: Rundvlees 99%, zout, kruiden	Ja
Ingrédients: pommes, poires et mangues	Nee
Zutaten: Wasser, Gerstenmalz, Hefe, Hopfen	Nee

Deze informatie diende als 'ground truth' om de voorspellingen van het algoritme op dezelfde 200 producten te staven. Zonder de toevoeging van de 'override' optie voorspelde het algoritme voor 195 van de 200 producten (97,5%) correct of het wel/niet Nederlands betrof, en werden 5 producten ten onrechte aangemerkt als verkeerde taal (vals positief). Met de toevoeging van de 'override' steeg het percentage correcte voorspellingen naar 100% en kwamen vals positieven of negatieven niet meer voor.

4.1.2 Gebruik van leestekens

Syntactische fouten in ingrediëntvelden kunnen o.a. bestaan uit verkeerd gebruik van haakjes (), [] of {} waarvan er meer openen dan sluiten. Haakjes worden vaak gebruikt om meer specifieke informatie te verstrekken over een wat algemener ingrediënt, bijvoorbeeld: "olie (palm- en rapzaad-)" of "conserveermiddelen (Natriumbenzoaat, E221 & E210)". Vaak is de ingrediëntenlijst langer, en wanneer een haakje ontbreekt is niet meer duidelijk wanneer een specificatie begint en eindigt. Met een logische controle kan voor elk type haakjes geteld worden hoeveel er openen en sluiten, en waarschuwen wanneer deze aantallen niet overeenkomen. Het is meestal niet mogelijk om fouten die op deze manier aan het licht komen automatisch te corrigeren.

Voor sommige ingrediënten specificeert men het percentage dat ervan aanwezig is in een product, waardoor er getallen met decimalen kunnen voorkomen in een lijst met ingrediënten. In Nederland is het gebruikelijk om voor getallen het symbool ',' als decimaalteken te gebruiken, maar bij vertalingen uit andere talen komt het regelmatig voor dat het symbool '.' nog wordt gebruikt. Dit kan automatisch herkend en gecorrigeerd worden met behulp van een reguliere expressie.¹⁴ Consistent gebruik van een decimaalteken is met name van belang voor geautomatiseerde interpretatie van ingrediëntvelden.

¹¹ Het data-model van GS-1 staat toe dat voor hetzelfde product meerdere ingrediëntdeclaraties in verschillende talen worden ingevoerd. De taal van de declaratie wordt in een apart taalcode-veld gespecificeerd. In dit geval werden alleen declaraties geselecteerd waarvoor taalcode NL was gespecificeerd.

¹² Het gebruikte taalherkenningsalgoritme is beschikbaar in sommige spreadsheet-editors en gebaseerd op Google Translate. Voor grootschalig gebruik en software-integratie van een dergelijk algoritme bieden diverse partijen (waaronder Google) tegen betaling een service met API.

¹³ Een kleine lijst met de voor dit voorbeeld relevante taal-ambigue woorden werd handmatig opgesteld, maar voor grootschalige toepassingen zou het beter zijn deze lijst automatisch te genereren op basis van gegevens uit een digitaal woordenboek of vertalingsalgoritme. Als het woord niet veranderd bij vertaling naar een andere taal is het taal-ambigue.

¹⁴ Deze expressie is: (?<=\d)\.(?=\d). Waarbij het gematchte teken (de punt) wordt vervangen door een komma.

Het komt af en toe voor dat symbolen per ongeluk tweemaal worden ingevoerd, bijvoorbeeld `,,` of `..` of een dubbele spatie waar een spatie volstaat. Ook komt het voor dat men vergeet een spatie te plaatsen na een `.` wanneer er een nieuwe zin begint. Ook deze typografische fouten kunnen met reguliere expressies herkend en opgelost worden.

Een probleem dat aan het licht kwam, maar waarvoor nog geen passende oplossing ontwikkeld kon worden, betrof het gebruik van andere symbolen `;` of `.` als scheidingstekens tussen losse ingrediënten, in plaats van de gebruikelijke `.`. Deze scheidingstekens spelen een belangrijke rol bij het segmenteren van velden naar individuele ingrediënten.

Alle hierboven beschreven logische controles werden geïmplementeerd en uitgevoerd op ingrediëntvelden van 62852 voedselproducten (de resulterende Nederlandse velden na taalherkenning). Met deze methode werd in 9810 velden een probleem geconstateerd. In sommige velden speelden meerdere problemen tegelijkertijd. Indien deze apart geteld worden, werden 10362 syntactische problemen geïdentificeerd. 8736 hiervan konden automatisch opgelost worden, waarmee 8184 van de 9810 aangemerkte velden correct werden bevonden. De overblijvende 1626 ingrediëntvelden konden niet automatisch gecorrigeerd worden. Automatische correctie kon dus in dit geval ongeveer 81% van het werk verlichten.

4.1.3 Segmentatie en ontologische spellingscontrole

Een typische declaratie van ingrediënten is een lijst van meestal meerdere individuele ingrediënten, die van elkaar gescheiden worden door bijvoorbeeld een komma. Om de datakwaliteit van afzonderlijke ingrediënten te kunnen bewaken is het nodig om het ingrediëntenveld te segmenteren in haar afzonderlijke ingrediënten. Deze stap klinkt eenvoudiger dan ze is, omdat er ontzettend veel manieren zijn om een ingrediëntenlijst op te schrijven. In de meest gangbare vorm worden ingrediënten gescheiden door komma's, maar regelmatig worden haakjes gebruikt om meer detail te verstrekken (*nesting*), bijvoorbeeld:

Ingrediënten: groente (aardappel 12%, wortel 6,3%, ui 5%), specerijen,...

Nog complexer wordt het wanneer er ook gebruik gemaakt wordt van koppeltekens om herhaling van woorden te voorkomen:

Ingrediënten: bloem, plantaardige olie (raap-, en palm-), zout,...

Hoewel niet direct van belang voor bewaking van datakwaliteit, maar wel voor andere vormen van geautomatiseerde verwerking van ingrediënten, is het zinvol te realiseren dat verschillende termen soms hetzelfde kunnen betekenen: ascorbinezuur is hetzelfde als vitamine C, en beiden kunnen ook aangeduid worden met het E-nummer E-300¹⁵.

Voor het segmenteren van ingrediënten werd een zogenaamde tokenizer gebruikt die de informatie tussen specifieke scheidingstekens, in dit geval een van de tekens `;(){}[]` en het woord `en`, uitsplitst naar afzonderlijke termen. Een uitzondering werd gemaakt voor het teken `.` wanneer het voor komt tussen twee getallen en dus als decimaalteken fungeert. Deze tokenizer biedt nog geen oplossing voor bovengenoemde problemen omtrent nesting en koppeltekens, maar wel voor het hiervoor genoemde probleem van punten als decimaalteken in plaats van komma's.

De termen die door de tokenizer op deze manier van elkaar gescheiden werden betreffen meestal ingrediënten, maar lang niet altijd. Tekstvelden met ingrediënten bevatten regelmatig extra informatie omtrent allergenen, keurmerken (bijvoorbeeld biologisch of 'beter leven') en soms zelfs bereidingsinstructies. Om te kunnen herkennen welke 'termen' nu precies ingrediënten, en of die correct gespeld zijn, zijn achtergrond informatie nodig. Deze achtergrondinformatie werd verkregen uit een ingrediëntenontologie, een soort woordenboek dat beschrijft welke mogelijke ingrediënten er in een declaratie kunnen voorkomen.

¹⁵ Zoals reeds eerder vermeldt geldt dit enkel in de context van een ingrediëntendeclaratie.

Figuur 1 Absolute (links) en cumulatief-relatieve (rechts) frequenties van termen die voorkomen in ingrediëntdeclaraties. De plek waar de rode stippellijn doorkruist wordt geeft aan met hoeveel termen 90% van de informatie kan worden afgedekt.

Deze ontologie bestond nog niet, en moest eerst worden samengesteld. De eerste vraag die dit oproep was: welke ingrediënten moeten erin staan? Op basis van de resultaten uit de tokenizer werd een term-frequentie analyse gedaan op de dataset van GS-1 (zie Figuur 1), waaruit duidelijk werd dat met circa 1500 individuele termen ongeveer 90% van de hoeveelheid informatie in de ingrediëntvelden kon worden beschreven. Als uitgangspunt is dus gekozen voor een ontologie op basis van de 1500 meest voorkomende termen, aangenomen dat voor elke term de juiste spelling ook de meest voorkomende spelling is. In een proces dat deels handmatig en deels geautomatiseerd verliep werden deze 1500 termen vervolgens verder opgeschoond. Zo werden termen die erg dicht bij elkaar lagen in spelling, bijvoorbeeld 'antioxydant' en 'antioxydant' samengebracht tot dezelfde term in de juiste spelling (desgewenst zou 'antioxydant' als alternatieve schrijfwijze voor dezelfde term kunnen worden opgeslagen). Termen die veel voorkwamen, maar geen ingrediënt bleken te zijn, werden verwijderd. Termen die synoniemen van elkaar bleken te zijn werden dusdanig georganiseerd dat duidelijk werd dat beide benamingen voor hetzelfde concept gebruikt mogen worden. De ontologie (zie Bijlage 2 voor een klein deel eruit) werd verder verrijkt met informatie over E-nummers en de namen van stoffen die als E-nummer bekend staan¹⁶. Bij wijze van experiment werd deze ontologie ook verrijkt met automatische vertalingen¹⁷ van de Nederlandse ingrediënten naar verschillende andere talen. Dit zou in de toekomst benut kunnen worden voor het automatisch naar andere talen interpreteren van Nederlandse ingrediëntenvelden.

De ontwikkelde ontologie kon vervolgens als externe 'ground truth' dienen in een algoritme voor spellingscontrole van de termen die door de tokenizer uit de data werden gesplitst. Hiertoe werd een zogenaamde relatieve *Levenshtein edit-distance* (afstandsmaat) uitgerekend: het aantal typografische bewerkingen dat moet worden uitgevoerd om de ene term in de andere te veranderen, gewogen met de lengte van de betreffende term. Voor iedere individuele term werd steeds gezocht naar het ingrediënt in de ontologie met de kleinste afstandsmaat, waarbij verschillen tussen hoofdletters en kleine letters werden genegeerd. Afhankelijk van de hiervoor gevonden afstandsmaat werden de volgende conclusies getrokken:

- Afstand 0: de term komt in de ontologie voor en is reeds correct gespeld
- Afstand tussen 0 en 0,1: de term komt in de ontologie voor, maar is verkeerd gespeld en wordt automatisch gecorrigeerd.
- Afstand tussen 0,1 en 0,5: de term komt in de buurt van iets dat in de ontologie bekend is, maar onvoldoende om tot automatische correctie over te gaan. Wordt aangemerkt als 'verdacht'.
- Afstand > 0,5: de term komt (vermoedelijk) niet in de ontologie voor, of is dusdanig verkeerd gespeld dat niet meer te herleiden is wat bedoeld wordt. Geen verdere actie.

¹⁶ Bron: https://nl.wikipedia.org/wiki/Lijst_van_E-nummers

¹⁷ Met behulp van het Google Translate algoritme

Om bovenstaande methode te evalueren werd handmatig een *ground-truth* vastgesteld voor de GS-1 dataset. Dit werd gedaan op 1000 willekeurig gekozen unieke termen, zoals die uit de term-frequentie analyse naar boven kwamen. Dezelfde termen werden tevens aan het algoritme voor ontologische spellingscontrole voorgelegd. Aan de hand van de term-frequentie gegevens kon vervolgens worden berekend wat de prestatie van spellingscontrolealgoritme zou zijn in de context van de GS-1 dataset.

In het totaal werden 582681 individuele termen beoordeeld (waarin termen van de ground-truth vaker kunnen voorkomen). Volgens de vastgesteld ground-truth zijn 9939 hiervan verkeerd gespeld. Het algoritme herkende en corrigeerde 8235 spelfouten terecht, terwijl 1704 spelfouten ten onrechte niet werden herkend en (ze werden wel als 'verdacht' aangemerkt). Het algoritme voerde in geen van de gevallen ten onrechte correcties uit. Voor veel termen kon door het algoritme met zekerheid worden vastgesteld dat ze correct gespeld waren (ze kwamen met een afstand van 0 voor in de ontologie). Wanneer een term (los gezien van eventuele spelfouten) in de ontologie voorkwam, kon het algoritme in 99,71% van de gevallen terecht aangeven of de spelling correct was of niet. Het algoritme kwam 131072 termen tegen die het niet met zekerheid kon relateren aan de ontologie. Handmatige inspectie van deze termen bracht aan het licht dat het met name termen betrof omtrent voedingsclaims, biologische certificaten, allergenen en waarschuwingen, of zeldzame ingrediënten die (vanwege term-frequentie) niet in de ontologie waren opgenomen. Bij nadere inspectie viel ook op dat het algoritme in sommige gevallen erg strikt was omtrent schrijfwijze, waar die voor menselijke interpretatie meestal vrijblijvender is. Zo werd bijvoorbeeld 'palm olie' gecorrigeerd in 'palmolie' en 'glucosefructosestroop' in 'glucose-fructosestroop'. Dergelijke striktheid is met name voor machinematig interpreteren van ingrediënten nuttig, maar minder noodzakelijk bij menselijke interpretatie.

Behalve dat deze aanpak helpt om spelfouten te detecteren en (meestal) te corrigeren heeft het verbinden van individuele ingrediënten aan een ontologie nog veel meer potentie. Het faciliteert de opstap naar een rijker datamodel voor ingrediënten. Met een uitgebreidere ontologie is het bijvoorbeeld mogelijk allergenen automatisch te herkennen, of vertalingen naar andere talen automatisch te genereren. Ook kunnen ontologische relaties gelegd worden tussen verschillende termen die hetzelfde betekenen (zoals in het voorbeeld van ascorbinezuur = vitamine C = E300). Dit vergroot de transparantie voor consumenten, terwijl de vrijheid voor producenten blijft bestaan. Andere aan ingrediënten gekoppelde informatie, zoals bijvoorbeeld gemiddelde waarden omtrent duurzaamheid van een ingrediënt, of vermijdingsadvies van een ingrediënt ten behoeve van een specifiek dieet, zouden ook in een dergelijke ontologie kunnen worden opgenomen.

4.2 Nutriëntdata

Nutriënteninformatie omvat de voedingswaarde die een product voor het menselijk lichaam heeft, of in ieder geval een benadering daarvan (er kunnen verschillen zijn in de effectiviteit waarmee verschillende mensen stoffen opnemen). Voor veel consumenten hebben nutriënten onderling weinig met elkaar te maken, maar met de juiste achtergrondkennis zijn er diverse onderlinge relaties te vinden tussen de getallen, en dat biedt houvast voor het bewaken van datakwaliteit. Voor de verkenning van technieken voor nutriënt data werd een openbare USDA dataset gebruikt.

4.2.1 Logische controles

Nutriënten worden over het algemeen vermeld in een hoeveelheid die aanwezig is per 100 gram van een product. Nutriënten waarvoor meer dan 100 gram wordt gedeclareerd zijn dus per definitie incorrect, evenals nutriënten waarvoor negatieve waarden worden vermeldt (dat laatste komt in de praktijk zelden voor). Ook wanneer de som van alle aanwezige nutriënten boven de 100 gram uitstijgt moet er sprake zijn van een fout, maar men moet daarbij wel rekening houden met nutriënten waarvan tevens sub-nutriënten worden vermeldt (bijvoorbeeld 'vetten' en 'verzadigde vetten'); deze mogen niet dubbel geteld worden. Voor sommige producten is de referentie-eenheid echter geen 100 gram, en in die gevallen zullen de logische controles daar dus op moeten worden aangepast. Het zou verstandig zijn deze kennis over een product in het datamodel op te slaan, omdat het nu vaak

onduidelijk is welke referentie-eenheid er voor een product wordt gehanteerd en daarmee wordt bewaking van datakwaliteit bemoeilijkt.

In sommige gevallen worden nutriënten uitgesplitst naar meer specifieke stoffen. Zo kan het totaal bijvoorbeeld aan vetten worden uitgesplitst naar verzadigde en onverzadigde vetten. Indien dit gebeurt is het zaak te controleren of de som van de onderdelen overeenstemt met het totaal aan vetten. Hiervoor is echter wel specifieke kennis nodig over definities van nutriëntnamen en hun onderlinge relaties. Standaardisatie en beschikbaar maken van dit type kennis in de vorm van meta-data zou een goede stap vooruit zijn.

Het omzetten van bepaalde nutriënten (met name vetten, koolhydraten, eiwitten) levert het lichaam energie op, maar er zijn boven- en ondergrenzen aan de hoeveelheid energie die uit deze nutriënten kan voortkomen. Deze kennis kan gebruikt worden om inconsistenties tussen dit soort nutriënten en de gedeclareerde hoeveelheid energie (in kJ of in kCal) te bewaken. Vuistregels voor omrekening van dergelijke nutriënten naar energie kunnen worden gevonden in bijlage XIV van de EU verordening 1169/2011 over etikettering op levensmiddelen. Om deze vuistregels toe te passen is echter aanvullende kennis nodig, onder andere over definities van nutriënten en hun onderlinge relaties; bijvoorbeeld kennis over welke stoffen *polyolen* worden genoemd. In de praktijk blijkt bovendien dat toepassen van deze vuistregels wordt gehinderd door de aanwezigheid van lege nutriëntvelden.

Vanuit de EU verordening 1169/2011 valt nog meer inspiratie te halen over bewaking van datakwaliteit, en daarmee tegelijkertijd handhaving van deze verordening. Met uitzondering van bepaalde soorten producten, is het voor producenten verplicht om informatie over ingrediënten te verstrekken en minimaal een aantal belangrijke nutriënten te kwantificeren (energie in zowel kCal en kJ, eiwitten, koolhydraten, mono-disachariden, vet, verzadigde vetzuren en zout). Wanneer een databron voldoende informatie bevat over de aard van producten kan automatisch herkend worden wanneer deze regels wel of niet van toepassing zijn. Zo kunnen de benodigde uitzondering gemaakt worden voor bijvoorbeeld fruit en groente (inclusief diepvries), water, thee, koffie en alcoholische dranken, en in alle andere gevallen de standaard richtlijnen toegepast worden.

Op veel producten staat energie in zowel kJ als kCal vermeld¹⁸. De relatie tussen deze twee eenheden is wiskundig bepaald: $1 \text{ kCal} = 4,1876 \text{ kJ}$. Wonderbaarlijk genoeg komt het vaak voor dat energiewaarden in de verschillende eenheden voor hetzelfde producten meer of minder dan deze factor van elkaar verschillen. Een kleine afrondingsfout is te verwachten, maar grotere verschillen dienen aangemerkt te worden als foutief.

Aan bovenstaande logische controles kleven behoorlijk wat randvoorwaarden. Omdat meerdere daarvan niet door de beschikbare dataset werden vervuld zijn ze niet geïmplementeerd, maar wel voor de volledigheid wel gedocumenteerd.

4.2.2 Zelflerende regressiemodellen voor controle en correctie

In de vorige sectie werd besproken hoe, op basis van kennis over energie-gevende nutriënten, logische grenzen kunnen worden gesteld. Hierbij wordt uitgegaan van enkele vuistregels. Wat daarbij buiten beschouwing blijft is dat de vermelde energie kan voortkomen uit een combinatie van verschillende nutriënten (en eventueel zelfs wordt beïnvloed door de aanwezigheid van andere nutriënten), die elk op een andere manier een bijdrage aan de energie leveren. Deze complexere relaties kunnen gemodelleerd worden met behulp van machine learning, wanneer data van een grotere hoeveelheid producten aanwezig is, en aangenomen mag worden dat het merendeel hiervan betrouwbaar is. Een andere voorwaarde is dat gegevens per nutriënt in dezelfde eenheid worden uitgedrukt. Het geleerde model kan vervolgens op twee manieren gebruikt worden. (1) Het model kan op basis van gedeclareerde nutriënten voorspellingen doen over de verwachte hoeveelheid energie in een product. Wanneer deze voorspelling teveel afwijkt van de werkelijk gedeclareerde hoeveelheid energie is er iets mis, dan zijn de nutriënten en energie-waarden onvoldoende consistent met elkaar.

¹⁸ Producten van Amerikaanse oorsprong hebben een vergrote kans op fouten, omdat in Amerika de afkorting "cal" wordt gebruikt voor "calories", terwijl daarmee feitelijk kilocalorieën (kCal) en niet calorieën (cal) bedoeld worden.

(2) Wanneer energie-informatie voor een product ontbreekt, maar de overige voedingswaarden wel gedeclareerd zijn, kan op basis hiervan het ontbrekende getal geschat worden.

Dit principe werd geïmplementeerd en uitgetest op nutriëntdata van 239095 producten uit de openbare USDA dataset (nutriëntgegevens waren in de GS-1 dataset niet beschikbaar). Als voorbewerkingsstap werden de nutriënten gestandaardiseerd. Dit zorgt ervoor dat gegevenskolommen met verschillende eenheden en spreiding als gelijkwaardig worden behandeld. Ook werden nutriënten die zelden of nooit gedeclareerd worden (onder andere sommige vitamines en mineralen) uitgesloten van gebruik. Machine learning op basis van lineaire regressie werd gebruikt om van alle beschikbare data een model te leren dat de relatie tussen nutriënten en energie kon beschrijven¹⁹. Met de keuze voor deze specifieke techniek wordt een extra aanname gedaan: dat de relaties tussen energie en nutriënten lineair zijn (hogere orde relaties kunnen dus niet goed gemodelleerd worden). Met het geleerde model werd vervolgens de hoeveelheid energie in een product voorspeld aan de hand van de beschikbare nutriënten, en het resultaat (per product) langs de daadwerkelijk gedeclareerde energiewaarden gelegd. Hiermee werd een correlatie van 0,9797 bereikt, met andere woorden: het model beschrijft het overgrote deel van de energierelaties in de dataset zeer goed (een correlatie van 1 is het hoogst haalbare, en zou zich moeten voordoen als alle data correct, volledig en nauwkeurig waren). Het resterende verschil (error) tussen voorspelde en daadwerkelijke energiewaarde kan nu worden gebruikt als maat voor datakwaliteit van individuele producten ten aanzien van de consistentie tussen energie en overige nutriënten. Een grote error duidt op een grote kans dat nutriënten en gedeclareerde energie niet consistent met elkaar zijn, en een van beiden mogelijk incorrect is. Proefondervindelijk zal moeten worden vastgesteld welke drempelwaarde voor deze error gehanteerd moet worden. De beslissing hiertoe zal afhangen van hoeveel zekerheid gewenst is over de datakwaliteit, en wat de kosten zijn voor menselijke controles van de verdacht aangemerkte data.

Ter illustratie van het werkingsprincipe werd een distributie van de error samengesteld (Figuur 2), waarin zichtbaar wordt hoe vaak errors tussen voorspelde en feitelijke energie van een bepaalde omvang voorkomen. De staarten van deze distributie vallen buiten het bereik van de grafiek. Aan de hand van deze distributie werd gekozen voor drempelwaardes van -100 en +100 kCal. Met deze drempels werden 3218 aangemerkt als verdacht. Deze producten werden gesorteerd op omvang van de error, om ter menselijke interventie te kunnen voorleggen, samen met de door het algoritme voorspelde energiewaarde. Voor deze menselijke beoordeling is echter een aanzienlijke hoeveelheid expert-kennis nodig. Het is daarom het overwegen waard om dezelfde methode anders toe te passen: het regressiemodel leren op basis van een andere dataset waarvan men de onderlinge relaties volledig vertrouwt (een soort 'ground truth' dus), en vervolgens toe te passen op de huidige dataset ter beoordeling van datakwaliteit.

¹⁹ Dit werd gedaan met Weka, versie: 3.8.0, zie <https://www.cs.waikato.ac.nz/ml/weka/>. Voor alle parameters van het lineaire regressie algoritme werden de default waardes gebruikt.

Figuur 2 *Distributie van de error die resulteert uit het verschil tussen gedeclareerde en voorspelde energiewaarde. De staarten van de verdeling liggen buiten bereik van de grafiek.*

Dit principe werkt onder de aanname dat het merendeel van de relaties tussen nutriënten en energie correct is, en dat er voldoende data beschikbaar is om een stabiel model te leren. Mede dankzij het gebruik van regularisatie (*ridge*) in het lineaire regressie-algoritme wordt voorkomen dat uitzonderlijk afwijkende datapunten (uitbijters) teveel invloed hebben op het leerproces. Hierdoor *generaliseert* het geleerde model goed over alle beschikbare data, en springen de uitbijters er bij het vergelijken van voorspellingen met daadwerkelijke energiewaardes sterker uit.

Het leerproces van deze techniek is vrij robuust tegen ontbrekende nutriëntwaardes, maar bij het maken van voorspellingen hebben ontbrekende waardes een grote invloed. Ontbreekt een waarde, dan wordt een product al snel als verdacht aangemerkt. Dit kan deels voorkomen worden door in een eerder stadium lege velden middels logische controle al af te vangen. Het zou ook mogelijk zijn de techniek dusdanig aan te passen dat ontbrekende getallen een minder grote invloed hebben op het aanmerken als 'verdacht', maar het is de vraag hoe wenselijk dat is.

Indien er weinig productdata beschikbaar zijn, zullen uitbijters het leerproces sterker beïnvloeden, en wordt het lastiger om een goed model te generaliseren. Als dit het geval is, is het verstandiger om middels een 'leave-one-out' principe de kwaliteit van de data te evalueren, omdat bij deze methode potentiële uitbijters buiten het leerproces worden gehouden. Deze aanpak vraagt echter wel om aanzienlijk meer rekenkracht, en is dus minder efficiënt.

Het principe dat hier is toegepast om uitbijters te vinden in numerieke relaties tussen verschillende kolommen met getallen is generiek en kan dus ook op andere numerieke data, waartussen relaties worden verondersteld, worden toegepast.

4.3 Productomschrijving en rubricering

4.3.1 Zelflerende classificatiemodellen voor controle en correctie

In sectie 4.2.2 werd beschreven hoe machine learning gebruikt kan worden om, middels een uit data geleerd model, de waarde van een continue variabele (zoals energie) te voorspellen en bewaken. Machine learning laat zich echter ook goed benutten voor het voorspellen van discrete variabelen, zoals bijvoorbeeld de indeling van een product in een productgroep. Het gebruik van een taxonomie voor het indelen van producten in productgroepen is onder andere van groot belang voor de vindbaarheid van producten in systemen en op websites. Dezelfde producten kunnen op allerlei manieren worden ingedeeld in groepen: op hun toepassing (ontbijt, lunch, avondeten), op waar ze van gemaakt zijn (fruit, groente, vlees), of op eigenschappen die relevant zijn voor transport en logistiek (verpakkingseenheid, gewicht, afmetingen). Wanneer een product bij een verkeerde

categorie is ingedeeld kunnen processen die hiervan afhankelijk zijn spaak lopen. Voor dit project werd gekozen om de GS-1 GPC categorie codes te gebruiken in een POC studie naar de toepassing van machine learning voor de bewaking van een juiste categorie-indeling. Dezelfde principes gelden echter evengoed voor andere taxonomieën.

Tekstvelden in voedselproductdata, zoals de omschrijving van een product, lijst van ingrediënten, functionele en wettelijke naamgeving en bereidingsinstructies geven samen veel informatie over een product. Machine learning algoritmes kunnen tekstuele data echter niet zonder tussenstap benutten, omdat ze alleen overweg kunnen met numerieke data (getallen). *Featurisation* is daarom eerst nodig; de informatie in tekst moet worden vertaald naar een reeks getallen. Voor deze studie werd hiervoor een TF-IDF (term-frequency inverse document frequency) algoritme gebruikt. Hierna werd een *multinomial naive bayes* classificatie algoritme gebruikt om een model uit de data te leren, en vervolgens ingezet om productgroepen te voorspellen op basis van hun tekstuele omschrijving. Deze methodiek kan wederom op twee manieren ingezet worden: (1) om verdachte productdata of een verkeerde indeling te signaleren, of (2) om te voorspellen tot welke categorie een product (dan wel) moet worden ingedeeld.

Voor deze studie werd een classificatiealgoritme gekozen dat in staat is om niet alleen een voorspelling af te leveren over de productgroep waartoe een product behoort, maar hier ook een waarschijnlijkheidswaarde aan te verbinden. Deze waarschijnlijkheidswaarde is van groot belang om de techniek te kunnen inzetten voor bewaking van datakwaliteit. Hoe lager de waarschijnlijkheidswaarde, hoe meer moeite het algoritme heeft om een product tot een categorie onder te brengen, en hoe zinvoller een menselijke interventie is. Daarnaast zijn natuurlijk discrepanties tussen voorspelde categorie en daadwerkelijke categorie indicatief voor een mogelijk probleem met datakwaliteit. Bij interpretatie is het belangrijk om te realiseren dat deze methodiek een product om twee redenen kan aanmerken als verdacht (in de zin van verkeerd gerubriceerd). Wanneer de tekstuele omschrijving van een product zeer afwijkend is in de context van de groep waarin het is ingedeeld kan dit betekenen dat het product in de verkeerde groep is ingedeeld, maar het kan evengoed betekenen dat de tekstuele informatie incorrect is. In de praktijk, en afhankelijk van de gewenste doelen en beschikbare menselijke bronnen, zal moeten worden bepaald wanneer de voorspellingen van een algoritme dusdanig zijn dat er een menselijk review van het product nodig is.

Voor deze POC werden data van 309056 producten uit de GS-1 dataset gebruikt (dit aantal omvat ook verschillende versies van hetzelfde product). Het algoritme werd getraind om aan de hand van bovengenoemde tekstvelden de categorie van het product (GPC code) te voorspellen. In de gebruikt data kwamen 47 verschillende GPC codes voor. Het getraind algoritme voorspelde in 99,89% van de gevallen dezelfde code als door GS-1 reeds was toegekend, maar in 341 gevallen een andere code. Deze werden aangemerkt als 'verdacht' ter nadere inspectie. Omdat kennis omtrent de exacte criteria voor de GPC-indeling ontbrak kon geen ground truth worden vastgesteld, maar op basis van gezond verstand leek de toegekende GPC code van een aantal van de 'verdacht' aangemerkte producten inderdaad erg onwaarschijnlijk. Op basis hiervan kon in ieder geval worden vastgesteld dat de techniek zich goed leent om onwaarschijnlijke GPC's aan het licht te brengen. Toch heeft de kwaliteit van de voorspellingen, net als bij andere technieken, te lijden onder (bijna) lege velden. Hoe meer lege velden, hoe minder informatie om betrouwbaar een categorie te voorspellen.

In sommige gevallen kunnen ook andere lessen getrokken worden uit de resultaten van een dergelijk algoritme. Zo werd bijvoorbeeld duidelijk dat de definitie van bepaalde GPC-groepen soms weinig onderscheidend zijn, of zelfs overlap hebben. Zo werden de volgende twee groepen vaak met elkaar verward als het noten betrof:

- 10000207 - Fruit/Noten/Zaadjes Gemengd – Bereid/Bewerkt (Houdbaar)
- 10000236 - Noten/Zaadjes – Bereid/Bewerkt (Houdbaar)

Dit soort inzichten kan helpen de GPC taxonomie verder te verbeteren. Overigens zou machine learning ook ingezet kunnen worden om op basis van data zelf suggesties te doen voor een optimale taxonomie (bijvoorbeeld op basis van een clustering algoritme).

5 Aanbevelingen

Aanbevelingen worden gedaan op twee verschillende terreinen. Het eerste is gericht op datamanagement en beleid omtrent data (sectie 5.1), die mede bepalend zijn voor resultaten die met technologie bereikt kunnen worden. Het tweede is gericht op specifieke problemen met datakwaliteit en de daarvoor relevante technieken (sectie 5.2), maar ook op de (huidige) onmogelijkheden en wensen voor de toekomst.

5.1 Datamanagement en beleid

Maak digitale data leidend

Om historische redenen is het begrijpelijk dat het fysieke etiket van een product momenteel leidend is bij beoordeling van de kwaliteit van productinformatie en handhaving van wetgeving omtrent voedselproducten. Echter staat deze beslissing optimaal gebruik van moderne databeheer- en automatiseringsmogelijkheden in de weg. Hierdoor blijft controle en handhaving een zeer arbeidsintensief en daardoor vaak onvolledig proces. Vanuit het perspectief van datamanagement zou het logischer zijn de digitale tegenhanger van de etiketinformatie leidend te maken, te onderwerpen aan (deels) geautomatiseerde controles, en daarna te gebruiken als gecontroleerde bron van waaruit andere informatiedragers, zoals het etiket, gegenereerd kunnen worden. Eenmaal correct bevonden is een digitale databron een veel gemakkelijker ijkpunt om andere vormen van productdata langs te vergelijken.

Handhaving wet- en regelgeving

Voortvloeiend uit de aanbeveling om digitale etiketdata leidend te maken, ontstaan ook mogelijkheden om het voor producenten van voedsel makkelijker te maken om goede data te verstrekken. Handhaving van wet- en regelgeving omtrent etiketdata kan voor een deel geautomatiseerd worden, wat het voor producenten eenvoudiger maakt om snel te weten of eraan voldaan wordt.

Controle aan de ingang

Dankzij automatisering van datakwaliteitsbewaking kunnen producenten onmiddellijk feedback krijgen over data die ze verstrekken. Dit vergroot de kans dat fouten rechtgezet kunnen worden, en zorgt er bovendien voor dat data die gedeeld worden vanaf het begin gecontroleerd zijn (in plaats van achteraf).

Een rijk en flexibel datamodel

De keuze voor een geschikt datamodel, de mal waarin data is opgeslagen, is mede bepalend voor datakwaliteit. Hoe specifiek en rijker het data model, hoe minder kans op fouten. Tegelijkertijd is om praktische redenen flexibiliteit gewenst. Moderne technologieën voor dataopslag, zoals Semantic Web, kunnen helpen om met behoud van de nodige flexibiliteit data rijker op te slaan.

Meta-data en standaardisatie

Direct verbonden aan het datamodel is de opslag van meta-data. Hierin is nog veel ruimte voor verbetering, waardoor de interpretatie van de data zelf eenvoudiger zal worden, en de kans op fouten daarin afneemt. Standaardisatie van meta-data tussen verschillende databronnen, en binnen diezelfde bronnen, is hierin ook een belangrijk element. Dit faciliteert namelijk het eenvoudiger combineren van data uit diverse bronnen, ook wanneer die buiten het voedingsdomein liggen (zoals bronnen met financiële, logistieke of geografische data of weerdata). Elk van deze bronnen kan context bieden die helpt bij bewaking van de kwaliteit van voedselproductdata. Als eerste stap is het aan te bevelen om in ieder geval metadata omtrent nutriëntnamen, ingrediëntnamen, allergenen, geografische locaties, eenheden en grootheden (bijvoorbeeld gewicht, valuta, volume, afmetingen) te standaardiseren.

Meegroeien met data

Voor een aanzienlijk aantal belangrijke aspecten van datakwaliteit is geautomatiseerde bewaking momenteel niet mogelijk, omdat context ontbreekt. Een voorbeeld hiervan is de herkomst van een product. Zonder aanvullende data is het niet mogelijk om vast te stellen of een product daadwerkelijk uit Spanje komt, maar wanneer meer logistieke data beschikbaar komt (bijvoorbeeld vanuit GPS) is dit wel mogelijk. Hoe meer databronnen ontsloten worden hoe meer relaties er gelegd kunnen worden. Dit soort context biedt meer mogelijkheden om datakwaliteit te beoordelen. Standaardisatie en goede data-organisatie faciliteren het leggen van dit soort relaties en zijn dus van groot belang om in de toekomst meer aspecten van datakwaliteit te kunnen gaan bewaken.

5.2 Technieken voor bewaking van datakwaliteit

5.2.1 De mogelijkheden

De eerste, en meest voor de hand liggende stap richting het automatisch bewaken van datakwaliteit, is het gebruik van logische regels voor het controleren van data. Dit kan uitstekend werken voor bewaking voor een aantal specifieke aspecten van datakwaliteit, maar is meestal onvoldoende wanneer er hogere eisen aan datakwaliteit gesteld worden. Voordeel is dat het mechanisme van logische controles relatief eenvoudig te begrijpen is, en daarmee vertrouwen wekt. In aanvulling op logische controles kan kunstmatige intelligentie uitstekend ingezet worden voor bewaking van complexere aspecten van datakwaliteit. In sommige gevallen is het echter niet eenvoudig om te begrijpen welk mechanisme schuil gaat achter de resultaten van dergelijke AI. Zorgvuldige validatie van de gekozen technieken is in dat geval gewenst. Wanneer hiermee voldoende vertrouwen ontstaat in de techniek kan deze volautomatisch toegepast worden. Is dit niet het geval, dan blijft menselijke interventie gewenst. Met behulp van AI is het dan bijvoorbeeld mogelijk om data met een verhoogde kans op fouten voor te leggen voor menselijke beoordeling, waarmee mankracht bespaard wordt die anders nodig zou zijn om alle data te doorlopen. De rolverdeling tussen mens en machine in dit soort semi-automatische scenario's kan ook geleidelijk opschuiven. In principe leidt elke menselijke interventie namelijk tot nieuwe kennis omtrent de herkenning van datakwaliteit, waarmee zelflerende algoritmes zichzelf weer kunnen verbeteren. Uiteindelijk kan zo steeds minder menselijke inmenging nodig zijn.

Voor een aantal van de belangrijkste tekortkomingen in productdata werden mogelijke oplossingen bedacht, en hieronder in een tabel weergegeven. Enkele van deze oplossingen zijn als proof-of-concept verder uitgewerkt in hoofdstuk 4.

Tekortkoming in data	Relevante techniek
Lege velden	Middels logische controle kan een leeg veld (ook als het enkel spaties of tabs bevat) opgespoord worden. Soms is extra context nodig (zoals productcategorie), omdat een veld voor bepaalde producten wel leeg mag zijn en andere niet.
Ontbrekende informatie	Indien een veld is gevuld met informatie, maar niet alle informatie bevat die men verwacht is dit zeer lastig te detecteren. Ingrediënten die ontbreken zijn zeer lastig op te sporen zonder aanvullende informatie (bijv. uit het productieproces of de logistieke keten van de producent). Het ontbreken van wettelijk verplichte nutriëntgegevens is middels logische controles te bewaken.
Inconsistente allergenen	In een ingrediëntendeclaratie worden allergene (delen van) ingrediënten met hoofdletters geschreven. In een aparte allergenendeclaratie worden deze allergenen nog eens expliciet vermeldt, aangevuld met informatie over allergenen waarvan sporen in het product kunnen voorkomen (maar die geen ingrediënt zijn). Als het om allergenen gaat is er dus een verschil tussen "bevat" en "kan sporen bevatten van". De gecombineerde informatie uit allergenen- en ingrediëntendeclaratie kan op verschillende manieren inconsistent zijn. Indien deze informatie aan een ontologie (of

	eenvoudige lijst) gelinked kan worden is dit de makkelijkste manier om consistentie te bewaken. In de ontologie kan de mate van allergeniciteit vastgelegd worden in relatie tot specifieke stoffen (een ingrediënt of als spoor in product aanwezig). Uiteraard moet men rekening houden met alternatieve schrijfwijzen en eventuele spelfouten. In sommige gevallen is het belangrijk de context van potentieel allergeen te overwegen, omdat variaties in verwerking van een ingrediënt van invloed kunnen zijn op allergeniciteit. In een samenwerking hebben GS-1 en Voedingscentrum samen reeds een oplossing geïmplementeerd voor het bewaken van allergeneninconsistenties.
Spelfouten in ingrediënten	Spellingscontrole, desgewenst met auto-correctie, op basis van <i>edit-distance</i> ²⁰ aan de hand van een generiek digitaal woordenboek in de juiste taal of (domein-specifieke) ontologie van ingrediënten. Zie sectie 4.1.3.
Spelfouten in andere tekstvelden	Spellingscontrole, desgewenst met auto-correctie, op basis van <i>edit-distance</i> ²⁰ aan de hand van een generiek digitaal woordenboek in de juiste taal.
Invoer van tekst in verkeerde taal	Taalherkenning op basis van een reeds getraind machine-learning algoritme (e.g. Google Translate). Het algoritme kan voorspellen wat de waarschijnlijke taal van een stuk tekst is, of een waarschijnlijkheidswaarde afgeven over een specifieke taal. Zie sectie 4.1.1. Voor ingrediënten zou men tevens een ontologie van ingrediënten kunnen gebruiken die meerdere talen ondersteunt, en die als ground truth dient voor herkenning op basis van edit-distance.
Verkeerd gebruik van leestekens	Logische controles kunnen een aantal fouten in het gebruik van leestekens opsporen, en ze soms ook herstellen. Zie sectie 4.1.2.
Invoer van verkeerde getallen, invoer in verkeerde eenheid of verkeerd geplaatst decimaalteken	Met enige voorkennis over de betekenis van het getal kunnen vaak logische grenzen gesteld worden. Daarnaast is soms kennis beschikbaar over relaties met andere getallen (of tekst), aan de hand waarvan fouten opgespoord kunnen worden (zie sectie 4.2.1) of kan machine learning ingezet worden om inconsistente relaties met andere getallen te identificeren (zie sectie 4.2.2). Tot slot kan ook statistiek gebruikt worden om te herkennen of een individueel getal erg sterk afwijkt van gangbare waarden van vergelijkbare getallen (eventueel binnen een productgroep).
Verkeerd decimaalteken gebruikt	In Nederland gebruikt men de komma als decimaalteken in getallen, in andere landen vaak een punt. Middels het datamodel kan correct gebruik van een decimaalteken meestal afgedwongen worden, maar komt het voor in een veld dat ook tekst bevat dan kan met een <i>reguliere expressie</i> het verkeerde decimaalteken opgespoord en vervangen worden door het juiste. De volgende <i>reguliere expressie</i> identificeert iedere verkeerd gebruikte punt als decimaalteken: $(?<=d),(?=d)$.
Verkeerde herkomst	Als een product feitelijk in Spanje wordt gemaakt, terwijl productdata Italië als herkomst noemen is dit zonder aanvullende gegevens vrijwel niet te herkennen. Met toegang tot logistieke gegevens van de producent of distributeur zou meer houvast ontstaan voor databewaking.
Invoer in verkeerd veld	Wanneer men informatie in een verkeerd veld invoert, bijvoorbeeld de productnaam in het veld voor ingrediënten, is dit soms goed te herkennen. Op basis van statistiek of machine learning, en eerder ingevoerde data, is het mogelijk een model te maken van de 'gangbare' eigenschappen van informatie in een specifiek veld, waarmee afwijkende invoer meestal herkend kan worden. Indien de invoer tekst betreft kan

²⁰ Er zijn diverse implementaties voor het berekenen van edit-distance. De populairsten zijn o.a. de Levenshtein-distance en de Jaro-Winkler-distance.

	<p>als tussenstap tekst-featurisatie gebruikt worden (zie sectie 4.3.1) alvorens machine learning toe te passen.</p> <p>Lastiger wordt het wanneer slechts een deel van het veld verkeerd is ingevoerd, bijvoorbeeld wanneer voedingsclaims samen met ingrediënten bij het ingrediëntveld worden ingevoerd. In die gevallen kan men soms gebruik maken van reguliere expressies, een ontologie met in het veld verwachte termen, of een andere vorm van <i>named-entity recognition</i> om gegevens van verschillende aard te scheiden.</p> <p>Indien invoer in het verkeerde veld plaatsvindt duidt dit vaak ook op teveel vrijheid in het gebruikte datamodel, of gebrekkige semantiek in de user-interface.</p>
Onduidelijkheid over betekenis van gegevens	<p>Indien een databron onvoldoende gegevens verstrekt over de betekenis van informatie (bijvoorbeeld de grootte en eenheid of betekenis van waarde 0) in die bron is dit met technologie niet meer te corrigeren: gegevens blijven moeilijk te interpreteren. Wel zou het mogelijk zijn dit probleem (semi-)automatisch te detecteren, maar dat lijkt niet erg zinvol. Inrichting van een goed datamodel en daarbij behorende rijke, gestandaardiseerde, meta-data (bijvoorbeeld in RDF/OWL) is typisch een eenmalig en grotendeels handmatig proces.</p>
Doublures	<p>(Bijna-)doublures waarbij (bijna) alle informatie identiek is tussen twee producten zijn makkelijker te herkennen wanneer alle informatie in getallen kan worden gevat. Voor tekstvelden kan gebruik gemaakt worden van tekst-featurisatie technieken, zoals bag-of-words of TF-IDF (zie sectie 4.3.1) waarmee tekst in getallen uitgedrukt wordt. Vervolgens kan men aan de hand van de ontstane getallenruimte en een afstandsmaat (bijvoorbeeld de Mahalanobis afstandsmaat) de gelijkenis bepalen tussen producten onderling. Producten die onderling weinig verschillen, en dus vermoedelijk (bijna-)doublures van elkaar zijn, zijn herkenbaar aan een erg lage afstandsmaat. Lege velden zullen een uitdaging zijn bij dergelijke toepassingen.</p>
Product in verkeerde groep geplaatst	<p>Met behulp van machine learning kan een model geleerd worden dat aan de hand van producteigenschappen voorspelt tot welke categorie een product hoort. Afwijkingen tussen voorspelde en daadwerkelijke categorie duiden op een mogelijk probleem (zie sectie 4.3.1). Lege velden zijn een uitdaging. Evenals kleine datasets, of datasets met zeer grote hoeveelheden fouten. Als alternatieve benadering kan men het machine learning model leren op basis van handmatig gecontroleerde data, zodat men meer zekerheid heeft over het geleerde model en resulterende voorspellingen.</p>

5.2.2 De onmogelijkheden

Naast het ontdekken van de mogelijkheden van AI is het even zo belangrijk te weten wat de onmogelijkheden zijn. Wie logische controles wil gebruiken om datakwaliteit (semi-)automatisch te bewaken is afhankelijk beschikbare kennis over de ingevoerde gegevens en de expertise om de deze te vertalen naar een algoritme (expert systeem). Wie machine learning wil gebruiken om datakwaliteit (semi-)automatisch te bewaken moet zich realiseren dat dit soort methoden sterk data-gedreven zijn. Wanneer men machine learning gebruikt als geavanceerde vorm van uitbijter-identificatie is dit geen probleem, omdat het daarbij gaat om relatieve eigenschappen van data in verhouding tot de hele dataset. Wanneer men machine learning modellen meer absolute waarde wil geven, is het van belang modellen te leren uit een gecontroleerde en betrouwbare dataset, die te valideren, en vervolgens pas toe te passen op bewaking van datakwaliteit in een andere bron.

Men hoopt dat digitale gegevens van een product een afspiegeling zijn van de werkelijkheid. Dat ingrediëntendeclaraties beschrijven welke grondstoffen gebruikt zijn voor de bereiding van het product, dat in geval van een biologisch keurmerk alle richtlijnen omtrent biologische teelt zijn

nageleefd, en dat bij een gedeclareerde energie van 250 kCal dit overeenkomt met de feitelijke energie in het product. De consistentie tussen digitale en fysieke werkelijkheid is echter niet te bewaken zonder fysieke meetmiddelen. De in dit rapport beschreven bewaking van datakwaliteit richt zich hoofdzakelijk op de onderlinge consistentie van verstrekte gegevens, en biedt dus geen garanties voor de fysieke werkelijkheid.

Data-actualiteit is een facet van datakwaliteit dat op dit moment erg lastig te bewaken is. Dit heeft niets zozeer te maken met een passende technologie maar met afwezigheid van de benodigde informatie. Om te kunnen beoordelen of data nog actueel is, is het nodig om de identiteit van verschillende versies van een product te kunnen bepalen. Afhankelijk van het seizoen kan een komkommer uit verschillende landen komen, de herkomstinformatie kan dus veranderen, maar de komkommer blijft dezelfde GTIN houden. Dankzij deze keuze is het niet meer mogelijk de identiteit (middels GTIN) te relateren aan de herkomst. Producenten beschikken vaak over aanvullende gegevens (batch nummer, productiedatum) over hun producten, maar die worden zelden meegegeven met overige productdata, terwijl ze wel zouden kunnen helpen de identiteit van een product beter te bepalen. Daarnaast hanteren verschillende databronnen verschillende regels omtrent versiebeheer van producten. Dezelfde GTIN mag vaak in gebruik blijven, ondanks kleine wijzigingen in een product, wat problemen omtrent data-actualiteit in stand houdt.

De herkomst van productgegevens, ook wel *provenance* genoemd, is zelden beschreven als het gaat om voedselproductdata. Het is vaak niet bekend of productgegevens gemeten zijn (en zo ja, hoe?), modelmatig berekend zijn, of zelfs geschat. Net als *provenance* is ook de nauwkeurigheid van (vooral numerieke) gegevens zelden bekend. Dit is typisch een probleem dat opgelost kan worden met goede (bij voorkeur gestandaardiseerde) meta-data, maar in het geval van nauwkeurigheid is het ook de vraag welke bronnen van informatie beschikbaar en relevant zijn. Wie een meetinstrument gebruikt voor het verkrijgen van voedingswaarden zou gegevens over de meetnauwkeurigheid van dit apparaat kunnen vastleggen en, als meta-data, meegeven met elke meting die door dit apparaat verricht wordt. Op deze manier worden databronnen met voedselproductinformatie rijker en veelzijdiger, maar omdat gegevens van interne processen van een producent nu gemakkelijker de buitenwereld bereiken is het zaak zorgvuldig na te denken over toegangsbeperking tot bepaalde (concurrentiegevoelige) gegevens.

Tot slot kan men overwegen om gegevens over voedselproductdata te verrijken met een maat voor datakwaliteit waarin verschillende aspecten, volgens een gestandaardiseerde methode, verwerkt zitten. Sommige bronnen met duurzaamheidsdata implementeren een zogenaamde data quality rating (DQR) op deze manier. Of dit wenselijk is hangt vooral af van hoe een eventuele DQR zou kunnen worden toegepast, en tot welke meerwaarde dit voor bedrijven, consumenten of overheid leidt. Indien men een DQR toepast is het voor het vertrouwen van data-gebruikers essentieel om transparant te zijn over de manier waarop de DQR wordt berekend, en welke aspecten van datakwaliteit hierbij wel en niet worden meegenomen.

6 Conclusie

(Semi-)automatische bewaking en (soms) herstel van datakwaliteit is mogelijk met verschillende technieken. De eerste, en meest voor de hand liggende techniek is het gebruik van logische controles op basis van specifieke kennis over de data en onderlinge relaties tussen gegevens. In de voorgaande hoofdstukken werd beschreven hoe daarnaast machine learning meerwaarde kan hebben bij bewaking van meer complexe relaties tussen gegevens, of de statistische eigenschappen van complexe gegevens. In sommige gevallen kan machine learning geheel zelfstandig problemen met datakwaliteit identificeren en (soms) corrigeren. In andere gevallen kan het behulpzaam zijn bij het ondersteunen van menselijke beoordeling van datakwaliteit, door bijvoorbeeld gegevens met de grootste kans op fouten naar boven te halen. Beide benaderingen reduceren de benodigde menselijke arbeid voor bewaking van datakwaliteit.

Op het gebied van datamanagement en beleid werden ook een aantal aanbevelingen gedaan. In plaats van het fysieke (papieren) etiket, wordt aangeraden om een goed gecontroleerde digitale tegenhanger hiervan leidend te maken bij de beoordeling van datakwaliteit. Dit zou tevens het deels geautomatiseerd handhaven van wet- en regelgeving omtrent etiketdata kunnen faciliteren. Beheerders van voedselproductdata wordt aangeraden meer aandacht te besteden aan ingangscontroles voor datakwaliteit, verbetering van de gebruikte datamodellen, en standaardisatie van met name meta-data, om zo datakwaliteit in al haar vormen, en het geautomatiseerde beheer daarvan, te bevorderen.

Bijlage 1 Variaties in schrijfwijze

Ter illustratie van de diversiteit aan spelfouten in ingrediëntendeclaraties werd voor een klein aantal veel gebruikte woorden gezocht naar alle mogelijke variaties in schrijfwijzen, zoals deze voorkwamen in de door GS-1 beschikbaar gestelde dataset. Variaties in schrijfwijze werden toegestaan voor enkelvoud en meervoudsvormen, en voor gangbare alternatieve schrijfwijzen (bijvoorbeeld 'emulgators' en 'emulgatoren'). De frequentieverhouding tussen correcte en foutieve varianten van hetzelfde woord werd tevens berekend en als percentage in de tabel uitgedrukt. In een paar gevallen vloeien foutieve schrijfwijzen voort uit een probleem met segmentatie van woorden.

Correcte schrijfwijze	antiklontermiddel antiklontermiddelen	antioxidant antioxidanten	conserveermiddel conserveermiddelen conserveringsmiddel conserveringsmiddelen	ingrediënt ingrediënten	kleurstof kleurstoffen	stabilisator stabilisatoren stabilisators	zuurteregelaar zuurteregelaars
Percentage fout gespeld	9.3%	3.3%	2.9%	6.3%	1.3%	3.2%	2.7%
Foutieve schrijfwijzen	aniklontermiddel	anioxidant	concerveermiddel	1 ngrediënten	gleurstof	s stabilisator	aazuurteregelaar
	anti klontermiddel	anitoxidant	concserveermiddel	b ingrediënten	kelurstoffen	sabilisatoren	s zuurteregelaar
	antiklntermiddel	anti oxidant	conderveermiddel	grediënten	keurstof	satbilisator	ssuurteregelaar
	antiklomtermiddel	anti oxidanten	conser veermiddel	igrediënten	kleur stoffen	satbilisatoren	suurteregelaar
	antiklonetrmiddel	anti oxydant	conserveeermiddel	iingredienten	kleurenstoffen	ssstabilisator	szuurteregelaar
	antiklontermiddel	antiaoxidant	conserveelmiddel	iingrediënten	kleursof	ssstabilisator	szuurteregelaars
	antiklontermiddelen	antiaxidant	conserveemiddel	in grediënten	kleurstif	ssstabilisatoren	zuurgeregelaar
	antiklontenmiddel	antiodixant	conserveemiddelen	ingedienten	kleurstoen	stabailisator	zuurregelaar
	antiklonter middel	antioioxidant	conserveemideel	ingediënten	kleurstofen	stabilasator	zuurregelaars
	antiklontermiddellen	antioixidan	conserveer middel	ingedriënten	kleurstoff	stabilatoren	zuurte regelaar
	antiklontermiddels	antioixidanent	conserveer middelen	ingrdiënten	kleurstoffe	stabilisatorenens	zuurte regelaars
	antiklontmiddel	antioixidantens	conserveerdmiddel	ingredient	kleurstoffens	stabilisators	zuurtegegerlaar
	antilklontermiddel	antioixidants	conserveermiddal	ingrediente	kleurstofs	stabilisor	zuurtegelaar
	antklontermiddel	antioixiddant	conserveermiddel e25	ingredienten	kleurstor	stabilistaor	zuurtegeregelaar
	s antiklontermiddel	antioxydant	conserveermiddleen	ingredientën	kleurtof	stabilistatoren	zuurteregalaar
	santiklontermiddel	antioxydanten	conserveermiddellen	ingredieten	kleustof	stabillisator	zuurteregalaars
		antioxydants	conserveermiddels	ingredinten	kleustoffen	stabilsator	zuurteregeaar
		antioxoxidant	conserveermiddelss	ingrediänten	kluerstoffen	stabilsatoren	zuurtereghelaars
		antooxydant	conserveermideel	ingrediënten	skleurstof	stablisator	zuurtereglaar
		artioxidant	conserveermidel	ingrediënten	skleurstoffen	stablisatoren	zuurteregulaar
		s antioxidant	conserveerniddel	ingrediënten		stailisator	zuurterelaar
		santioxidant	conserveersmiddel	ingrediënen		stailisatoren	zuurtregelaar
		santioxidanten	conseveermiddel	ingrediënte		stailisator	zuurtreglaar
			conseveermiddelen	ingrediëntenen			zuuteregelaar
			coserveermiddel	ingrediëntenp			zuuteregelaar
			konserveermiddelen	ingrediëntens			
			onserveermiddel	ingrediënts			
			s conserveermiddel	ingredriënten			
			sconserveermiddel	ingredënten			
			sconserveermiddelen	ingrediënten			
		ss conserveermiddel	ingrediënten				
		ssconserveermiddel	ingrenienten				
			ingrtedienten				
			ingrédiënten				
			ingrèdienten				
			inrediënten				
			ngrediënten				
			s ingrediënten				
			singrediënten				

Bijlage 2 Ingrediënttermen op volgorde van frequentie

De eerste 50 ingrediënttermen uit de ontologie op aflopende volgorde van frequentie.

Positie	Ingrediënt
1	zout
2	suiker
3	water
4	melk
5	specerijen
6	tarwebloem
7	citroenzuur
8	plantaardige olie
9	glucosestroop
10	dextrose
11	zonnebloemolie
12	tarwe
13	azijn
14	ui
15	gist
16	knoflook
17	soja
18	sojalecithine
19	cacaoboter
20	glucose-fructosestroop
21	E330
22	cacaomassa
23	gistextract
24	aardappelzetmeel
25	maltodextrine

Positie	Ingrediënt
26	gemodificeerd maiszetmeel
27	magere melkpoeder
28	E202
29	E471
30	weipoeder
31	lactose
32	selderij
33	maiszetmeel
34	mosterd
35	kaliumsorbaat
36	tarwezetmeel
37	gemodificeerd zetmeel
38	ascorbinezuur
39	volle melkpoeder
40	E500
41	xanthaangom
42	wortel
43	gerstemout
44	varkensvlees
45	E450
46	melkeiwit
47	raapzaadolie
48	tarwemeel
49	vanille
50	melkzuur

To explore
the potential
of nature to
improve the
quality of life

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
www.wur.nl/wfbr
info.wfbr@wur.nl

Rapport 1977
ISBN 978-94-6395-131-9

De missie van Wageningen University & Research is 'To explore the potential of nature to improve the quality of life'. Binnen Wageningen University & Research bundelen Wageningen University en gespecialiseerde onderzoeksinstituten van Stichting Wageningen Research hun krachten om bij te dragen aan de oplossing van belangrijke vragen in het domein van gezonde voeding en leefomgeving. Met ongeveer 30 vestigingen, 5.000 medewerkers en 10.000 studenten behoort Wageningen University & Research wereldwijd tot de aansprekende kennisinstellingen binnen haar domein. De integrale benadering van de vraagstukken en de samenwerking tussen verschillende disciplines vormen het hart van de unieke Wageningen aanpak.