Gebruiksvriendelijker maken van de verwerking van open overheidsgegevens

Gebruikerspaden voor het verlagen van de drempel bij portaalgebruik

---

Master thesis submitted to Delft University of Technology

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in **Complex systems Engineering and Management, specialization: Information Architecture**

Faculty of Technology, Policy and Management

by

Rafik Chelah

Student number: 4113136

To be defended in public on April 18th 2019

Graduation committee

Chair and First supervisor: Prof.dr.ir. M.F.W.H.A. Janssen, Section ICT
Second Supervisor : Dr.ir. I. Bouwmans, Section Energy and Industry

Voorwoord

Dit is rapport is tot stand gekomen in het kader van het scriptieonderzoek dat door de auteur van dit rapport is uitgevoerd.
Dit rapport is bedoeld voor lezers die interesse hebben in nieuwe kansen door middel van het analyseren van gegevens die door overheidsinstanties zijn gebruikt in de praktijk op bestuurlijk niveau. Het belang van de onderzoeksresultaten in dit rapport ligt bij het opstellen van bruikbare data-analyseprocessen voor datagebruikers.
De auteur is veel dank verschuldigd aan de begeleiders van de afstudeercommissie voor hun waardevolle adviezen: allereerst de heer Janssen als eerste begeleider en de heer Bouwmans als tweede begeleider. Daarnaast dank aan mevrouw Kam voor haar advies om het rapport taalkundig te verbeteren voor de lezer. Verder ook dank aan iedereen in het algemeen die mij advies heeft gegeven om mijn tekst te verbeteren na het lezen ervan. Verder ook dank aan de vijf geïnterviewde respondenten om de gebruikerspaden uit hoofdstuk 3 te evalueren. Dankzij de feedback die daaruit is voortgekomen zijn de gebruikerspaden beter gemaakt. Verder veel dank aan mijn ouders voor hun steun tijdens de studie.

Delft, maart 2019
Rafik Chelah

Samenvatting

De Nederlandse overheid heeft een opendataportaal (data.overheid.nl). Via dit portaal kunnen dataproviders open data beschikbaar maken voor hergebruik door datagebruikers. De datagebruikers kunnen onderverdeeld worden in *standaard eindgebruikers* en *ontwikkelaars*. Hierbij analyseren *standaard eindgebruikers* open data voor inzichten die hen in staat stellen om nieuwe kansen te herkennen, aan de hand van onder andere ontwikkelingen in de maatschappij. Bijvoorbeeld door aan de hand van cijfers die de netto bevolkingstoename weergeven binnen een regio, de behoefte van nieuwkomers tegemoet te zien in de vorm van nieuw aanbod aan producten. Hiervoor dienen namelijk in een vroeg stadium investeringen te zijn gedaan voor winstgevende activiteiten op de langere termijn. De *ontwikkelaars* zijn erop gericht om applicaties te ontwikkelen voor eindgebruikers met intuïtieve gebruikersinterfaces die een bepaalde hoofdfunctie vervullen. Hiervoor zijn ontwikkelaars geïnteresseerd in applicaties die samenvallen met ontwikkelingen in de praktijk die een onvervulde informatievraag genereren onder standaard eindgebruikers. Ontwikkelaars zijn bijvoorbeeld: applicatieontwikkelaars, data-analisten en databankontwerpers. Dit onderzoek richt zich op het gebruik van open data door standaard eindgebruikers. De gebruiksvriendelijkheid van het portaal is niet optimaal waardoor kansen onbenut blijven voor eindgebruikers.

Dit komt omdat hulpmiddelen ontbreken bij het effectief verwerken van open data. Ook is het lastig om strategische keuzes te maken bij het selecteren en gebruiken van open data, doordat het aanbod niet gestandaardiseerd is. Er is dus geen eenduidige manier waarop open data kunnen worden ingelezen en gebruikt voor de verrijking met andere datasets en informatie. De drempel voor het gebruik van het overheidsportaal is hoog waardoor het gebruik achterblijft bij de mogelijkheden die open data kunnen bieden aan standaard eindgebruikers.

Het overheidsportaal zou kunnen worden verbeterd op twee vlakken; (i) het aanbod van open data, en (ii) de strategieën vanuit eindgebruikers voor het gebruik van open data. Het uiteindelijk doel is daarbij om effectief gebruik van open data te stimuleren en daarmee te zorgen dat data gebruikt worden op een laagdrempelige manier met behulp van de creativiteit van eindgebruikers. Hierbij hoort de volgende hoofdvraag:

'Wat moeten portaaleigenaren doen om de drempel voor het gebruik van open data te verlagen?'

Om deze hoofdvraag te kunnen beantwoorden, zijn verschillende onderzoeksmethoden toegepast voor het onderbouwen van de problemen en het komen tot aanbevelingen voor specifieke maatregelen. Het probleem is dat er technische functies ontbreken voor het verwerken van open data op het portaal. Hierbij zijn er verschillen in hoe open data is aangeboden op het portaal. Met behulp van technische functies die door alle datagebruikers kunnen worden gebruikt hoeft een gebrek aan vaardigheden en kennis geen belemmering te zijn om gebruik te maken van open data, waarbij open data verwerkt kunnen worden op basis van een vooropgestelde onderzoeksvraag en/of hypothese. De datagebruikers kunnen hiermee worden gestimuleerd om creatief gebruik te maken van open data zonder dat de drempel voor het gebruik ervan te hoog is.

Allereerst is er een literatuuronderzoek ingezet om een aantal aspecten voor het portaal vast te kunnen stellen, namelijk de benoeming van functies waarmee open data kunnen worden verwerkt, de criteria waarmee verschillen in het aanbod van open data kunnen worden benoemd, en wettelijke richtlijnen uit officiële publicatiebladen op Europees niveau ten aanzien van beleid.

Ten tweede is een empirische analyse uitgevoerd op het huidige portaal voor het bepalen van ontbrekende functies die in de literatuur zijn benoemd.

De derde onderzoeksmethode is een experimentele aanzet tot gedefinieerde gebruikerspaden die nog niet op deze manier zijn gedefinieerd als leidraad voor eindgebruikers. Dit is het eerste onderzoek dat gebruikerspaden definieert en gebruikt voor open data. De gebruikerspaden kunnen worden hergebruikt door gebruikers voor data-analyse in verschillende domeinen en voor verschillende gebruikersdoeleinden.

De vierde gebruikte onderzoeksmethode omvat interviews met deskundigen die onderzoek doen aan de Technische Universiteit Delft. Voor dit onderzoek zijn in totaal zes opendatadeskundigen geïnterviewd. Eén opendatadeskundige is voor aanvang van het literatuuronderzoek geïnterviewd om suggesties te krijgen voor geschikte wetenschappelijke literatuur voor het literatuuronderzoek. De andere vijf deskundigen zijn geïnterviewd voor het evalueren van de ontbrekende functies op het portaal en voor evaluatie van en het verkrijgen van feedback op de gebruikerspaden.

De volgorde en samenhang waarin de verschillende onderzoeksmethoden zijn toegepast is als volgt.

In de literatuur zijn vier hoofdcategorieën van criteria gevonden waarin de verschillen in het aanbod van open data en daarmee gebruiksvriendelijkheid van open data kunnen worden uitgedrukt; (i) formaat, (ii) metadata, (iii) toegang, en (iv) kwaliteit (Dawes et al., 2016).

Tevens zijn functies geïdentificeerd die onder andere kunnen worden gebruikt bij het verzamelen, analyseren en visualiseren van open data. Met een empirische analyse van het huidige overheidsportaal is bepaald welke functies uit de literatuur ontbreken. Deze ontbrekende functies zijn vervolgens geëvalueerd na interviews met vijf deskundigen met behulp van enquêtevragen waarmee het belang van elke ontbrekende functie is vastgesteld. Verder zijn er EU-richtlijnen geanalyseerd om maatregelen te formuleren waarmee portaaleigenaren met de wettelijke richtlijnen kunnen omgaan. Deze richtlijnen bepalen de verplichtingen bij het aanbieden van open data op een portaal voor de verantwoordelijke overheden.

Vervolgens zijn op een experimentele manier gebruikerspaden ontworpen met behulp van data-analysetechnieken uit literatuur. Het gebruik van open data kan op een gestructureerde manier beschreven worden doordat strategische keuzes bij het analyseren van open data bepalend zijn voorgestelde doelstellingen. De gebruikerspaden geven een reeks van activiteiten aan om open data te verwerken voor een bepaald doel. De gebruikerspaden zijn gedefinieerd in de vorm van procedures bestaande uit te volgen stappen om vooropgestelde doelen te kunnen bereiken. De manier waarop de gebruikerspaden zijn gedefinieerd is experimenteel doordat ze nog verder moeten worden verbeterd aan de hand van onderzoek naar specifieke gebruikersbehoeften bij het analyseren van open data. Deze behoeften moeten vervolgens ook nog worden vertaald naar taken die vallen binnen een gebruikerspad. De initiële gebruikerspaden zijn, met behulp van enquêtes gericht aan opendatadeskundigen, onderzocht op gebruiksvriendelijkheid en verbeteringsmogelijkheden.

Met de volgende conclusies en aanbevelingen kan het aanbod van open data en het portaal gebruiksvriendelijker worden verbeterd.

Er zijn vier categorieën van criteria gevonden die kunnen worden gebruikt om het aanbod van open data op het portaal te verbeteren:

1. Formaat: verschillende voorwaarden voor het gebruik van open data, bijvoorbeeld waarbij software als gereedschap is benodigd om er toegang tot te krijgen. Voorbeeld: gelinkte en machine-leesbare open dataformats stellen datagebruikers in staat om direct in actie te komen met data-analyses zonder aparte preparatietaken voor het koppelen van datasets
2. Metadata: de exacte beschrijving van databronnen met een gezamenlijke context voor alle datagebruikers, in de vorm van een aparte informatiestructuur en relevante trefwoorden voor de vindbaarheid van datasets.
3. Toegankelijkheid: verschillende bestandsformaten waarin de open data worden aangeboden, inclusief de mogelijkheid om door datasets te navigeren via webpagina's, zonder de open data eerst te moeten downloaden.
4. Kwaliteit: omgang met kwaliteitsverschillen tussen datasets waarmee ook de kwaliteit van data-analyses wordt beïnvloed indien deze niet worden gecorrigeerd met behulp van beschikbare technieken of door de dataproviders.

De open data worden beschouwd in de vorm van datasets en kunnen hierbij worden onderverdeeld in *gestructureerd* on *ongestructureerd*. Voorbeelden van *gestructureerde* open dataformaten op het huidige portaal zijn: XML, JSON, CSV, en andere dataformaten. Voorbeelden van *ongestructureerde* open data zijn PDF-documenten, video, audio, foto's, brieven, beleidsnota's en andere gegevens die in verschillende dataformaten worden aangeboden. Aangezien het nut in open data gelegen is in de mogelijke koppeling tussen datasets, is er een voorkeur voor gestructureerde machine-leesbare datasets die snel elektronisch verwerkt kunnen worden. Daarmee kunnen datasets namelijk efficiënt gelinkt kunnen worden met andere datasets voor waardevolle data-analyse doeleinden. Bij gestructureerde datasets kunnen op basis van attributen subsets worden gefilterd uit een dataset. Dit kan aan de hand van een querytaal zoals SQL waarbij de datasets in een relationele database worden ingeladen en gekoppeld met andere gegevens. Hiermee kunnen opdrachten worden uitgevoerd met de wiskundige logica die de samenhang tussen gegevens representeert en die tot nieuwe inzichten kan leiden voor eindgebruikers. Oftewel, het analyseren en koppelen van open data is gemakkelijker indien open data gestructureerd worden aangeboden. Het kost meer moeite om ongestructureerde open data te analyseren en te koppelen aan andere gegevens doordat de open data moeten worden afgestemd op dezelfde gegevenstypen. De gebruikte gegevenstypen zijn afhankelijk van de verschillende taken die door datagebruikers worden uitgevoerd voor de gewenste inzichten. De vindbaarheid van datasets kan worden vergroot door aparte informatiestructuurbeschrijvingen toe te voegen, bijvoorbeeld in de vorm van een doelenboom, ter verduidelijking van de inhoudelijke betekenis ten opzichte van andere open data. De toegankelijkheid is hierbij de mogelijkheid om door open data te navigeren zonder het direct te hoeven downloaden naar een applicatie. Bijvoorbeeld door middel van gestandaardiseerde HTML-webpagina's. Ten slotte is de kwaliteit afhankelijk per situatie waarin gebruik wordt gemaakt van open data en moet er gelegenheid zijn om dit frequent te controleren door eindgebruikers via het portaal. Bijvoorbeeld met behulp van technieken die helpen om de kwaliteit van data te kunnen beoordelen voor het beoogde gebruik ervan.

De dataproviders en portaaleigenaren kunnen de EU-richtlijnen toepassen via open databeleid met behulp van de volgende drie maatregelen:

1. Privacygevoelige gegevens anonimiseren ter bescherming van personen.

2. Kostentoerekening op basis van de hoeveelheid gegevens toepassen via licenties. Bijvoorbeeld kosten in rekening brengen voor het gebruik van webservices, maar niet voor het gebruik van open data in het archief van het portaal.
3. Via onlinekanalen contact zoeken met gebruikers om te vragen wat men wil weten van de overheid en hierop inspelen door het aanbieden van nieuwe open data op het portaal.

Door middel van de hierboven genoemde maatregelen kunnen portaaleigenaren hun portaal veiliger en gebruiksvriendelijker maken. Met deze maatregelen kunnen portaaleigenaren het aanbod van open data beter laten aansluiten aan specifieke behoeften van datagebruikers.

De empirische analyse van het portaal op basis van de functies uit de literatuur laat zien dat de volgende functies (Zuiderwijk et al., 2014; Zuiderwijk, 2015), ontbreken in het huidige portaal:

1 Analyseren van open data
2 Gebruikerspaden kiezen
3 Visualiseren van data-analyse resultaten
4 Volgen van datagebruikers
5 Analysebevindingen delen via sociale media
6 Versiebeheer
7 Data kwaliteitsbeheer
8 Tutorial aanbieden
9 Interactiemechanisme

Deze ontbrekende functies kunnen een gebruiksvriendelijke verwerking van open data op het portaal stimuleren met gebruikerspaden als leidraad.

Er zijn vier gebruikerspaden gedefinieerd op een iteratieve manier als experiment voor het volgen van stappen op basis van vooropgestelde onderzoeksvragen en/of hypothesen. Dit is een experimentele en daardoor uitbreidbare reeks stappen die kan worden gevolgd om inzichten uit open data te verwerven. De gebruikerspaden stellen kennis paraat, waardoor datagebruikers gestructureerd te werk kunnen gaan. Het is geen garantie voor succes omdat dit ook afhangt van de vaardigheden met betrekking tot het verwerken en testen van de open data. De vier gebruikerspaden zijn gefocust op de volgende gebruikersbehoeften bij open datagebruik:

1. Beschrijven van samenvattende informatiewaarden van datasets, om beslissingen in de toekomst te onderbouwen.
2. Classificeren van gegevens, om in de vorm van voorspellende modellen, wat-als vragen te stellen voor het doorrekenen van verschillende scenario's.
3. Trendontwikkelingen onderzoeken met tijd als onafhankelijke variabele of index.
4. Groeperen of scheiden van deelpopulaties met gemeenschappelijke kenmerken uit een dataset met datapunten om relaties tussen variabelen te onderzoeken.

De belangrijkste aanbeveling is dat de portaaleigenaren de drempel kunnen verlagen door de negen ontbrekende functies te implementeren in het portaal. De gebruikerspaden kunnen gebruikt worden om de drempel bij het gebruik van open data verder te verlagen. Het is te verwachten dat datagebruikers, met behulp van de functies en gebruikerspaden, minder moeite en tijd kwijt zijn door leidend gebruik van open data.

Trefwoorden: Open data, Portaal, Gebruiksvriendelijkheid, Functionaliteit, en Gebruikerspaden.

Inhoudsopgave

Lijst van tabellen en figuren

Lijst van afkortingen en terminologie

API: Application Programming Interface, voor toegang tot een webservice waarmee een applicatie kan communiceren om toegang tot open data te krijgen (Wikipedia, 2018).

ATOM-Feed: voor automatische aanpassing van content op een website (Wikipedia, 2018).

CKAN: Comprehensive Knowledge Archive Network, is opensource software met een focus op dataopslag en datadistributie (Wikipedia, 2018).

Drupal: Een opensource software dat wordt gebruikt om websites en weblogs te beheren (Wikipedia, 2018).

HTML: HyperTextMarkup Language, is de standaard opmaaktaal van structuur en tekstuele inhoud van webpagina's in webbrowsers. Eventueel met directe koppelingen naar andere documenten en bestanden, of webpagina's met behulp van links (Wikipedia, 2018)

Javascript: een taal om webapplicaties te ontwikkelen en voor dynamische webpagina's (Wikipedia, 2018).

JSON: JavaScriptObjectNotation, is een gestandaardiseerd gegevensformaat met dataobjecten die bestaan uit attributen met bijbehorende waarden (Wikipedia, 2018).

Machine-learning: is een combinatie van algoritmen en statistische modellen om prestaties van computersystemen vooruitstrevend te verbeteren bij de uitvering van een taak. Hierbij zijn de modellen met behulp van de algoritmen gebouwd op basis van trainingsdata als input. De output zijn beslissingen of voorspellingen die initieel nog niet vaststonden zoals in een expliciet programma het geval is (Wikipedia, 2018).

MySQL: Een opensourcemanagementsysteem voor relationele databases, waarbij gegevens volgens een voorgeschreven relationeel model met elkaar samenhangen. Hierbij kunnen gegevens worden opgevraagd en aangepast volgens de SQL (Structured Query Language) taal (Wikipedia, 2018).

Opensource: Softwareproject dat vrij toegankelijk is en gebruikt mag worden voor elk doel door iedereen en die door verschillende onafhankelijke bijdragers tot stand is gekomen in een samenwerking (Wikipedia, 2018).

PDF: Portable Document Format, om publicatie in oorspronkelijke vorm, in digitaal formaat mogelijk te maken (Wikipedia, 2018).

RDF: Resource Description Framework, een driedelige structuurbeschrijving van databronnen op het web, voor het verdere gebruik ervan. Hierbij is de subject-predicaat-object structuurvorm, gebaseerd op taalkundige logica, voor respectievelijk de beschrijving van de bron, kenmerk ervan, en waarde van dat kenmerk uit een bron (Wikipedia, 2018).

RDF/XML: Dit is een machine-leesbare vorm van RDF om gegevensuitwisseling via RDF mogelijk te maken (Wikipedia, 2018).

SOAP: SimpleObjectAccessProtocol voor communicatie tussen verschillende applicatiecomponenten (Wikipedia, 2018).

URI: Uniform Resource Identifier (URI), voor het uniek benoemen van gegevens die via internet toegankelijk zijn voor anderen (Wikipedia, 2018).

URL: Uniform Resource Locator, is het adres naar een bestand op internet wat via de webbrowser kan worden gedownload. Dit kunnen ook webpagina's zijn in de vorm van bestanden (Wikipedia, 2018).

Usecasediagram: voor de specificatie van het gewenste gebruik van een systeem. Het is een gevisualiseerde weergave van actoren buiten een systeemgrens, en hun bedoelingen binnen de systeemgrens in de vorm van usecases. Hierbij worden de relaties tussen usecases binnen de gedefinieerde systeemgrens weergegeven (Wikipedia, 2018).

Webservice: de interface van een applicatiecomponent via webprotocol SOAP en datastructuur XML (Wikipedia, 2018).

XML: Extensible Markup Language, is gestandaardiseerde opmaaktaal om gestructureerde inhoud op te kunnen slaan, verzenden, en weer te kunnen geven dat zowel machine-leesbaar is als voor de mens (Wikipedia, 2018).

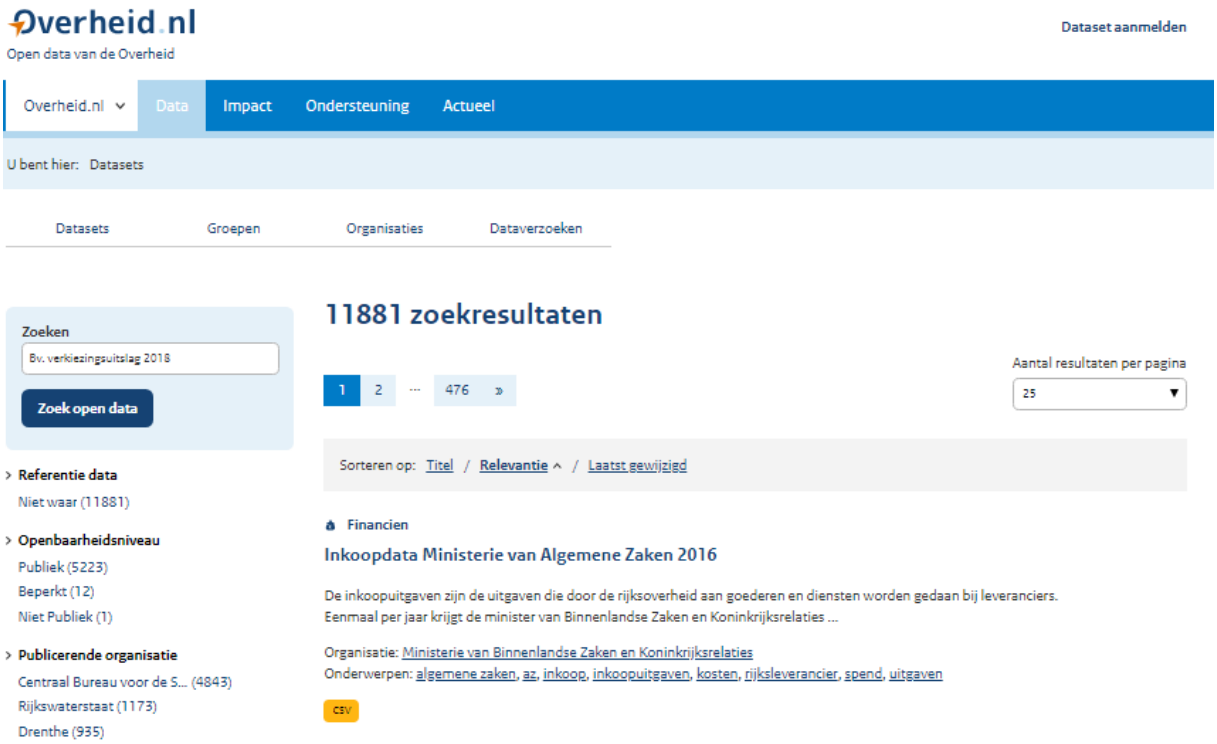
1 Inleiding

Het doel van dit hoofdstuk is om de aanleiding en daaruit volgende onderzoeksvragen voor het onderzoek te presenteren. Verder worden de gebruikte onderzoeksmethoden binnen het onderzoek beschreven, gevolgd door de structuurbeschrijving voor de rest van het rapport.

Paragraaf 1.1 presenteert achtergrondinformatie over het portaal van de Nederlandse overheid. Vervolgens beschrijft 1.2 de belangen van portaaleigenaren, dataproviders, en datagebruikers, in het portaal. Daarna beschrijft 1.3 het belang van de koppeling tussen verschillende open data en informatieconcepten. Hierbij presenteert 1.4 een voorbeeld van een maatschappelijk relevante applicatie. Vervolgens beschrijft 1.5 problemen vanuit datagebruikers met behulp van concrete voorbeelden. Oplossingsrichtingen hierbij worden gepresenteerd in 1.6 voor nieuw beleid door portaaleigenaren. Paragraaf 1.7 introduceert gebruikerspaden als strategische oplossing voor datagebruikers en het gebruik ervan in andere domeinen. Paragraaf 1.8 presenteert de werkwijze voor het onderzoek naar oplossingsrichtingen aan de hand van subonderzoeksvragen en onderzoeksmethoden. Ten slotte beschrijft 1.9 de afbakening van het onderzoek met behulp van een onderzoeksmethodologie uit wetenschappelijke literatuur, inclusief structuurbeschrijving dat de rest van het rapport aankondigt.

1.1 Achtergrondinformatie over het portaal van de overheid

De Nederlandse overheid heeft een opendataportaal waarop open overheidsgegevens staan gepubliceerd (url: data.overheid.nl). Via een zoekmachine kan met behulp van zoektermen en filters worden gezocht naar open data. Figuur 1 toont een voorbeeld van een webpagina die toegang geeft tot beschikbare datasets.



Figuur 1: Screenshot van het portaal. Overgenomen uit "Data" van Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, z.d. (data.overheid.nl/dataset). z.d., Ministerie van Binnenlandse Zaken en Koninkrijksrelaties.

Het portaal geeft toegang tot zowel machine-leesbare gestructureerde datasets als ongestructureerde open data die niet direct machine-leesbaar zijn.

Voorbeeld voor *ongestructureerde* open data: een Pdf-rapport waarin een aanbeveling is gemaakt en gepresenteerd, voor een plaatselijke gemeente. Dit is een document dat inzicht kan verschaffen in verschillende vormen van overheidsbeleid in specifieke regio's. Maar het is ook bruikbare informatie voor particulieren en ondernemers die strategische keuzes maken voor investeringen binnen die gemeente (oorsprong van open data: data.overheid.nl).

Voor gestructureerde open data zijn er, behalve downloadbare databestanden, ook machine-leesbare informatiebronnen zoals webservices. Op het overheidsportaal is naar dergelijke webservices verwezen in de vorm van URL-links en met de oorspronkelijke dataprovider als versiebeheerder. Een specifiek voorbeeld van een webservice, heeft CBS als dataprovider, en die toegankelijk is via XML-Feed of JSON-API (CBS, 2018). Het gaat hierbij om een JSON-file Atom-Feed over de internationale handel in Nederland; in-, uit- en doorvoer. Deze gegevens zijn gesorteerd aan de hand van de Standard International Trade Classification (SITC), een lijst met goederen voor de periode van 2002-2012 (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2018). Dit is een webservice waarbij de open data kunnen worden verkregen met behulp van nog te ontwikkelen maatsoftware, bijvoorbeeld in de vorm van een applicatie die de open data verwerkt. Het vergt nog veel voorbereiding, vaardigheden en kennis van zaken bij het ontwikkelen van waardevolle software.

Er zijn hoge verwachtingen met betrekking tot het openstellen van overheidsgegevens (Dawes et al., 2016). Hierbij zijn overheidsgegevens beschikbaar gesteld aan het publiek, in de vorm van open data, voor individueel hergebruik. Een belangrijke eerste vraag bij het begrijpen van open data is steeds: *'Wat zijn open data?'.* Ubaldi (2013) definieerde open data als volgt: *"data that can be freely used, re-used and distributed by anyone, only subject to (at the most) the requirement that users attribute the data and that they make their work available to be shared as well" (Ubaldi, 2013: 6).*

Omdat dit onderzoek zich richt op overheidsgegevens die als open data beschikbaar kunnen worden gesteld, zal vanaf nu met de term 'open data', openbare overheidsgegevens worden bedoeld. Oftewel, overheidsgegevens die als open data beschikbaar zijn gesteld door de overheid. Ubaldi definieert overheidsgegevens als volgt: *"any data and information produced or commissioned by public bodies" (Ubaldi, 2013: 6).*

Het opendataportaal kan worden gebruikt om open data te vinden die beschikbaar is gesteld. Maar volgens de volgende definitie van een opendataportaal hoeft het zich niet te beperken tot het vinden van open data**:** *"A piece of software that has implemented the core features to manage open data. Those features include, but are not limited to, user management, data publishing, metadata management, dataset storage, access control, search and visualization, etc." (OpenDataMonitor, 2018: 1).*

De rol van de overheid als informatieverstrekker via open data is niet beperkt tot het beschikbaar stellen van datasets. Naast het aanbieden van open data kunnen koppelingen tussen verschillende datasets leiden tot nieuwe inzichten via het portaal. In het huidige portaal ontbreken echter technische functies om de open data te verwerken en betekenisvol te maken.

De informatieve presentatie van content op de website wordt gedaan met behulp van een Content managementsysteem Drupal (data.overheid.nl). Via het portaal worden links naar bestanden aangeboden via een zoekmachine met zoekfilters zoals thema, status, toegang etc. Het archiveren van de open data gebeurt via CKAN (Comprehensive Knowledge Archive Network) (data.overheid.nl). De portaaleigenaren hebben bij het gebruik van de hiervoor genoemde technologie, een belang bij goede technische prestaties van het portaal in werking

en snelheid. Daarnaast om frequent gebruik van het portaal, door eindgebruikers, te kunnen faciliteren op een gebruiksvriendelijke manier.

Datagebruikers zouden het portaal kunnen gebruiken om voorspellende modellen te produceren. Met behulp van voorspellende modellen kunnen 'wat-als-onderzoeksvragen' worden onderzocht, waarbij open data als middel kunnen worden gebruikt. Verschillende scenario's kunnen hiermee vooraf worden doorgerekend. Dit is bruikbaar om strategische keuzes te maken bij het ontwikkelen van activiteiten waar vraag naar is. Om die vraag te bepalen kunnen trends worden gemodelleerd, eveneens met open data als middel.

1.2 Belangen van stakeholders in het portaal

Het implementeren van informatietechnologie, zoals een portaal, is een kostbaar proces doordat er veel manuren nodig zijn door specialisten op het gebied van informatietechnologie. Bijvoorbeeld voor het koppelen van webservices met functies op het portaal zelf. De portaaleigenaren zijn hierbij verantwoordelijk voor de kosten, terwijl het portaal geen direct commercieel belang dient. Daarom is het budget beperkt en dienen stakeholders te worden geanalyseerd voor een duidelijke afweging tussen belangen en kosten voor portaaleigenaren. Daarmee zijn de mogelijkheden om het overheidsportaal verder te laten ontwikkelen, bijvoorbeeld door gespecialiseerde bedrijven op het gebied van informatietechnologie, ook beperkt. Om het ontwikkelperspectief van het portaal toch te onderbouwen zal rekening worden gehouden met de belanghebbende stakeholders. Oftewel, het begrijpen voor wie en waarom dit portaal belangrijk is voor hen.

Een definitie voor stakeholders is als volgt: *"Stakeholders are the people for whom we build systems" (Rozanski & Woods, 2005: 6).* Deze definitie kan worden gebruikt voor softwarearchitectuur, om softwaresystemen te ontwerpen aan de hand van de belangen vanuit stakeholders. Dit is daarom ook van toepassing op het portaal, voor het streven naar verbetersuggesties. Om rekening te houden met de belangen van stakeholders in het portaal, worden ze in de rest van deze paragraaf toegelicht.

De eerste stakeholders zijn de portaaleigenaren die voor het portaal betalen met een beperkt budget uit publieke middelen waar de belastingbetaler aan bijdraagt. Het overheidsportaal heeft het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties als opdrachtgever en het Kennis- en Exploitatiecentrum Officiële Overheidspublicaties voor het beheer van het portaal (Overheid.nl, 2018). Deze overheidsinstanties worden daarom binnen dit onderzoek als de portaaleigenaren beschouwd. Hierbij zijn de open data informerend en bovendien staat de overheid open voor innovatieve en nieuwe concepten, om ondernemerschap te stimuleren (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2015). Vernieuwing is ook mogelijk omdat de hoeveelheid informatie toeneemt door de invloed van technologische ontwikkelingen op het gebied van informatie en communicatie. Daarmee kunnen namelijk inzichten worden verworven voor nieuwe kansen in de maatschappij.

De dataproviders zijn verantwoordelijk voor het uploaden van open data waaronder: Gemeentes, RDW, Ministeries, CBS, provincies, onderzoeksinstellingen, netbeheerders etc. Het belang de dataproviders is hierbij beleidsdoelen behalen voor tevreden burgers met minimale risico's bij het beschikbaar maken van open data naar aanleiding van dataverzoeken. Om dit gebruiksvriendelijk te maken voor dataproviders, is het belangrijk dat het uploaden van open data op het portaal op een gemakkelijke manier gaat. Het mag dus niet teveel moeite of tijd kosten om het beschikbaar te maken, ook in het belang van de datagebruikers om bijvoorbeeld afwijzing van dataverzoeken om dit soort redenen te voorkomen.

Er zal een onderscheid worden gemaakt tussen standaard eindgebruikers en ontwikkelaars. Hierbij gebruiken ontwikkelaars open data voor data-analysetaken, met behulp van grotere hoeveelheden datastromen uit meerdere bronnen. Hierbij dient software als gereedschap om vernieuwde concepten te kunnen uitwerken in de praktijk, bijvoorbeeld in de vorm van applicaties die voor eindgebruikers toegankelijk zijn. Onder ontwikkelaars vallen: onderzoekers, data-analisten, applicatieontwikkelaars, en databankontwerpers. Het belang van ontwikkelaars is dat het portaal voorziet in gedetailleerde presentatie van metadata over werkende databronnen. Met hierbij een duidelijke en gebruiksvriendelijke weergave van de oorspronkelijke context van open data. De eindgebruiker maakt gebruik van een kleinere hoeveelheid gegevens in de vorm van datasets en/of rapporten om relaties tussen factoren te leren begrijpen. Onder de eindgebruikers vallen: burgers, ondernemers, journalisten, ambtenaren, wetenschappers, en studenten. Hun belangen zijn: nieuwe inzichten voor het maken van keuzes, bijvoorbeeld met behulp van interpreteerbare visualisatie van data-analyseresultaten. Maar het stimuleren van open datagebruik voor het oplossen van maatschappelijke vraagstukken gaat niet vanzelf wat blijkt in paragraaf 1.3.

1.3 Het belang van koppelingen tussen open data en informatieconcepten

Het alleen erkennen van de waarde van open data is niet waardevol, het gebruik ervan in combinatie met collectieve intelligentie kan wel leiden tot het genereren van waarde (Janssen et al., 2012). De aanname is dat voor het verkrijgen van collectieve intelligentie de verschillende invalshoeken van datagebruikers nodig zijn. Deze kunnen worden gebundeld als een collectief voor het vinden van consensus in maatschappelijke oplossingen. Op deze collectieve intelligentie kan indirect een beroep worden gedaan door maatschappelijke betrokkenheid via applicaties die gekoppeld zijn aan informatiebronnen (Ubaldi, 2013). Dit kan leiden tot inzichten die waardevol zijn voor de maatschappij in de vorm van nieuwe diensten die voorzien in een behoefte. Er is echter wel een community van getalenteerde en gemotiveerde ontwikkelaars nodig om het gebruik van open data tot een succes te maken (Chan, 2013). Daarom is het niet alleen belangrijk dat open data toegankelijk zijn, maar ook dat het platform erachter gebruiksvriendelijk is voor gemotiveerde talenten in het algemeen. In het huidige portaal is de drempel voor het portaalgebruik te hoog door nog aanwezige problemen. Een voorbeeld voor een probleem is de toegankelijkheid van open data op het overheidsportaal. Er is een onvriendelijke en onvolledige manier van opendata aanbieden waardoor er moeilijk onderscheid kan worden gemaakt tussen de verschillende open data op basis van de inhoud. Een lage gebruiksvriendelijkheid bij het onderscheiden kan het gebruik van open data verhinderen voor het effectief inzetten. Dit werkt demotiverend en laat slechte ervaringen achter bij eenieder die het portaal een kans heeft gegeven maar het niet als een realistische optie ziet voor het gebruik van open data waarvoor het bedoeld is. Laat staan voor ontwikkelaars die applicaties willen ontwikkelen en waarbij de eisen van consumenten ongekend hoog zijn door de wereldwijde concurrentiemarkt op het gebied van informatietechnologie. Alles moet kloppen en de applicatie moet bovenal vernieuwend waardevol zijn als optie in de eigen omgeving van consumenten.

1.4 Voorbeeld van een maatschappelijk relevante applicatie

Een ideaal voorbeeld van een dergelijke applicatie is: ‘Googlemaps’ door Google, waarbij gebruikers zowel kunnen zoeken op specifieke zoektermen, als ook navigeren over een gevisualiseerde landkaart die tot de verbeelding van alle gebruikers spreekt (zie www.googlemaps.com). Deze applicatie maakt gebruikt van verschillende

informatieconcepten en open data. Allereerst is er een visuele weergave van een gedigitaliseerde landkaart die verschillende locaties in Nederland en andere landen weergeeft. Daarnaast is er een geïntegreerde zoekmachine waarmee via een interactieve zoekbalk kan worden gezocht met behulp van zoektermen zoals een nader te bepalen locatie. Dit kan vervolgens worden opgezocht en de software zal dit met een symbool aanduiden op de gevisualiseerde landkaart. Om een dergelijke applicatie mogelijk te maken zijn verschillende datasets en informatieconcepten nodig die gelinkt zijn met elkaar. De zoekcoördinaten van de plaatsen op de landkaart zijn allemaal opgeslagen in een databank, en in de databank gelinkt aan specifieke zoektermen die de locaties kunnen representeren. De zoekmachine kan de zoektermen met plaatsen of straatnamen koppelen op de gevisualiseerde kaart met locaties. Hierbij kan rekening worden gehouden met fout gespelde straatnamen en/of onbestaande namen. Er is hier wel voorbereiding nodig van de datasets met de coördinaten en namen, en de visualisatie die presentatie van de informatie mogelijk maakt op de kaart. Door dit soort applicaties kunnen open data worden gebruikt in de vorm van nieuwe verbindingen tussen verschillende datasets en informatieconcepten (Ding et al., 2012). Maar voordat de ontwikkeling van dergelijke applicaties kan worden versneld dienen problemen van datagebruikers te worden opgelost, rekening houdend met wettelijke richtlijnen. Paragraaf 1.5 omvat de problemen inclusief bijbehorende voorbeelden. Vervolgens beschrijft 1.6 oplossingsrichtingen en waarbij ook zal worden ingegaan op wettelijke richtlijnen bij de ontwikkeling van open databeleid op Europees niveau.

1.5 Problemen met het portaal vanuit datagebruikers

Voorbeelden van datasetproblemen waar normale datagebruikers tegenaan lopen op het huidige portaal zijn als volgt. Dit is bepaald na een praktische analyse van het portaal voor een eerste indruk. Er zijn verschillende soorten datasets zoals JSON-bestanden, waarvoor specifieke technische vaardigheden nodig zijn om er, zonder menselijke tussenkomst, mee te kunnen communiceren door middel van softwareprotocollen. Om dit te leren is een aanzienlijke tijdsinvestering nodig. De datagebruiker moet namelijk verschillende technologieën leren gebruiken zoals XML en SOAP. Daarnaast zijn er kwaliteitsverschillen tussen verschillende datasets, waarbij onbekend is hoe hoog de kwaliteit is voordat een datagebruiker door de dataset kan navigeren. Dit moet later blijken nadat datagebruikers hebben gezocht en de moeite hebben genomen om het te openen, mits het geen webservice betreft waarbij de gebruiker software als gereedschap nodig heeft. De gegevens in een dataset kunnen bijvoorbeeld onoverzichtelijk en/of foutief zijn weergegeven.

Een overzicht van Voorkomende problemen met de databronnen zijn weergegeven in tabel 1.

Tabel 1: Problemen met databronnen (uit: Dawes et al., 2016: 26)

Link is niet permanent beschikbaar
Navigeren door dataset zonder de gegevens eerst te moeten downloaden niet mogelijk
Metadata is onduidelijk voor voldoende begrip van open data
Informatiestructuur over hoe de open data zijn opgebouwd is onduidelijk
Trefwoorden zijn niet representatief voor open data

De link naar een dataset of webservice, is niet permanent beschikbaar. Verder is het niet mogelijk om te navigeren door de dataset alvorens te besluiten om er gebruik van te maken

en te downloaden, al dan niet via een application programming interface (API). Om een dataset te openen moeten deze ergens lokaal worden opgeslagen op een computer of in een database op een server die ergens anders is gelokaliseerd waarvoor software als gereedschap is vereist. De metadata rondom de databron is vaag of onduidelijk, trefwoorden voor datasets komen niet overeen met de inhoud. De informatiestructuur is onduidelijk, met betrekking tot hoe de dataset is opgebouwd, dus met behulp van welke attributen de meetwaarden van elke variabele zijn gerepresenteerd. Deze attributen kunnen bijvoorbeeld in de vorm van een doelenboom worden gevisualiseerd om de relaties tussen de meetbare factoren en verhoudingen in de gehele dataset ten opzichte van andere attributen inzichtelijk weer te geven. Deze informatie helpt eindgebruikers bij het maken van strategische beslissingen bij het verzamelen en koppelen van open data uit verschillende bronnen. Ook is het bijwerken van de geschiedenis niet altijd duidelijk met betrekking tot de vraag hoe recent de open data zijn. Ook zijn de contactgegevens van de verantwoordelijken niet altijd bekend.

Voorbeelden van problemen door ontbrekende portaalfuncties zijn als volgt. Analyseren van datasets op het portaal is onmogelijk, zonder veel voorwerk te verrichten om open data te kunnen gebruiken met behulp van tools en zelf programmeren om de verwerking mogelijk te maken. De visualisatie is hier onderdeel van, omdat de resultaten dan effectief kunnen worden geïnterpreteerd op het portaal in plaats van via omwegen met andere tools. Tenslotte ontbreekt het aan interactiemogelijkheden met anderen op het portaal, om bijvoorbeeld van elkaars inzichten te kunnen leren op basis van dezelfde dataset. Daarnaast ontbreken er tutorials waarmee vaardigheden kunnen worden aangeleerd om het portaalgebruik mogelijk te maken voor iedereen die interesse heeft in open data. Kortom, het streven is om het portaal gebruiksvriendelijk te maken voor datagebruikers door middel van een technische invalshoek.

Een concreet voorbeeld is een wetenschapper die het portaal niet gebruikt en als gevolg hiervan een kans mist om zijn onderzoek sterker te onderbouwen. De wetenschapper doet onderzoek binnen een discipline waarvoor het portaal als een waardevolle bron wordt geschat. Om dit te bevestigen is een nadere beschouwing van de open data op bruikbaarheid voor het onderzoek vereist. De wetenschapper gespecialiseerd in economie, heeft bijvoorbeeld geen informaticakennis om bijvoorbeeld een API te gebruiken van een webservice waarvoor maatsoftware als gereedschap voor moet worden gebruikt. De API is hierbij de schakel tussen applicatie, die de open data gebruikt voor interne processen, en de databank of databron die als een webservice beschikbaar is gesteld door een overheidsinstantie. Ook heeft de wetenschapper beperkte middelen en tijd voor het onderzoek. Hierdoor besluit de wetenschapper geen gebruik te maken van een dataset over de in-, uit- en doorvoer van goederen in Nederland tussen 2002-2015 (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2018). Op basis van het stukje tekst over de open data kan de wetenschapper namelijk niet bepalen of het bruikbaar is voor het onderzoek. Oftewel, hoe de internationale handel van Nederland zich heeft ontwikkeld in de periode na de kredietcrisis met 2008 als hoogtepunt. Er is daarvoor meer informatie over de dataset nodig in de vorm van metadata of structuurbeschrijving van de dataset. Door het ontbreken van analytische vaardigheid en tijdgebrek ziet de wetenschapper af van het risico om ongeschikte gegevens te verwerven. Daarmee mist de wetenschapper een kans om met behulp van beschikbare open data een andere kant te belichten van de internationale handel.

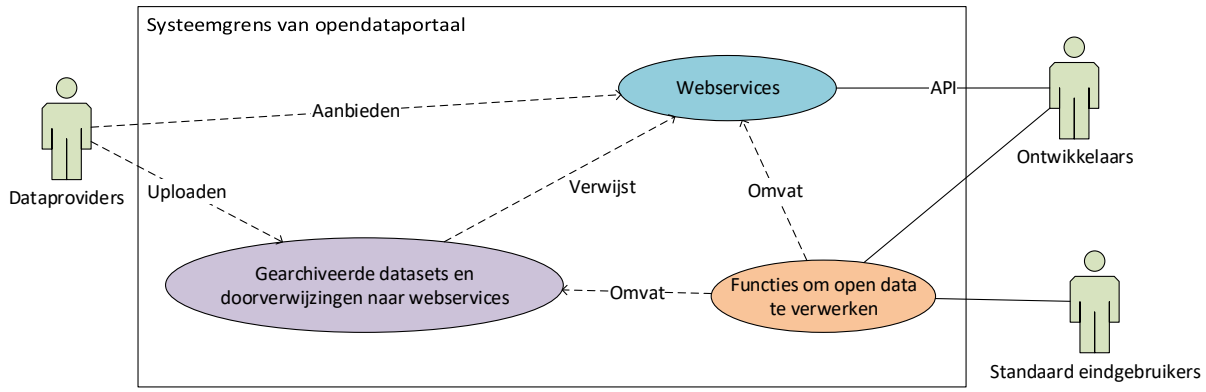
Voor meer gevorderde datagebruikers kan het portaal stimulerend zijn door het organiseren van samenwerkingsverbanden met andere ontwikkelaars. Voor minder gevorderde gebruikers kan het uitbreiden van de functies op het portaal tot een vooruitgang leiden bij het gebruik van open data.

Het probleem is dat in de huidige situatie de open data niet goed genoeg wordt gebruikt waardoor er kansen blijven liggen. Dit komt omdat er technische functies ontbreken voor het verwerken van open data en omdat de gebruiksvriendelijkheid van het portaal niet optimaal is. Hierbij zijn databronnen en datasets van invloed op de gebruiksvriendelijkheid doordat dit de waarde van het portaal voor datagebruikers representeert. Maar het huidige portaal heeft slechts een zoekfunctie naar datasets die door dataproviders zijn geüpload, waarbij de databronnen worden gesorteerd per thema en filters bevat om de zoekresultaten te verfijnen. Verder kan men dataverzoeken indienen voor specifieke open data via een online contactformulier. Er ontbreken functies die datagebruikers in staat stellen om de open data op het portaal te kunnen verwerken tot bruikbare informatie. In de volgende paragraaf zal daarom een usecasediagram worden gepresenteerd om systeemgrenzen aan te duiden. Daarmee kunnen beleidsoplossingen worden geformuleerd die rekening houden met de belangen van verschillende stakeholders buiten de systeemgrenzen van het portaal.

1.6 Formuleren van oplossingsrichtingen

Bij de problemen van het huidige portaal spelen alle stakeholders een rol. Er zijn in de huidige situatie meerdere portalen van gemeentes (Amsterdam, Utrecht, Haarlem, Breda etc.), Provincies (Gelderland, Zuid-Holland, etc.), Tweede kamer, CBS, RDW, etc. waarheen kan worden verwezen via het portaal van de overheid. Maar hiermee wordt open data verspreid over verschillende URL's. Er zijn inmiddels zelfs initiatieven waarmee open data van de overheid wordt geïndexeerd op basis van kenmerken en karakteristieken. Dit is niet gebruiksvriendelijk voor datagebruikers omdat het slechts gericht is op het vinden van open data op een centrale locatie. Het is de bedoeling om juist een centrale vindplek te hebben voor alle open data zodat koppelingen tussen datasets eerder kunnen worden gemaakt dan nu het geval is, voor zowel standaard eindgebruikers als ook ontwikkelaars. Dat is essentieel voor het verkrijgen van nieuwe inzichten en de ontwikkeling van vernieuwende concepten zoals de overheid wil.

Om oplossingsrichtingen te geven is het usecasediagram gebruikt om overzicht te creëren. Een overzicht van de bedoelingen in de vorm van usecases en relaties daartussen staan weergegeven binnen een gedefinieerde systeemgrens dat het opendataportaal representeert. Hiermee staan de belangen en relaties tussen belanghebbenden gevisualiseerd in een overzichtelijk diagram. De systeemgrens wordt gerepresenteerd door de rechthoek, en buiten de systeemgrens staan de stakeholders uit 1.2 weergegeven in figuur 2.



Figuur 2: Usecasediagram van het portaal

Wat we kunnen leren uit dit diagram is dat het portaal, als centrale bewaarplaats voor open data, een essentiële rol speelt voor alle stakeholders. Hiermee is gebruiksvriendelijkheid in ieders belang. De oplossingsrichtingen zijn als volgt. Om het gebruik van open data te structureren is het vereist dat de gebruikte open data aan criteria voldoen voor het effectief kunnen verwerken van open data. Met behulp van deze criteria kan de vindbaarheid, kwaliteit, en toegankelijkheid van open data worden verbeterd. Daarnaast zijn er technische functies die ontbreken in het huidige portaal, maar die wel nodig zijn om open data effectief te kunnen verwerken. Hierbij wordt rekening gehouden met Europese richtlijnen waarop het opendatabeleid moet worden afgestemd om te voldoen aan wettelijke kaders (PbEU, 2013; PbEU, 2014; PbEU, 2016). Uit figuur 2 blijkt dat het stimuleren van webservices, waarmee getalenteerde ontwikkelaars aan de slag kunnen om vernieuwende concepten te kunnen ontwikkelen, een stimulans is voor ontwikkelaars. Bijvoorbeeld door het aanbieden van tutorials, aan dataproviders, voor het operationeel maken van webservices vanuit overheidsinstellingen. Uiteraard is uitbreiding van open data in alle vormen gewenst, maar machine-leesbare formaten kunnen gebruikersvriendelijker worden gekoppeld aan andere datasets en zijn daarmee meer geschikt als open data.

Het gestructureerd gebruiken van open data voor analysedoeleinden helpt om technische keuzes bij de uitvoering van data-analyses te vergemakkelijken voor datagebruikers. Hierbij is het van belang om rekening te houden met verschillende soorten datagebruikers aangezien het gebruik van open data verschilt per datagebruiker. Met die verschillen kan rekening worden gehouden door gebruikerspaden te ontwikkelen.

1.7 Gebruikerspaden als strategische oplossingen voor datagebruikers

De definitie voor een gebruikerspad is als volgt: *"A way of achieving a specified result; a course of action" (Oxford, 2018: 1).*

Gebruikerspaden zijn relevant aangezien ze voor tijdswinst zorgen en er inspiratie kan worden opgedaan bij het gebruik van open data door potentiële datagebruikers op het portaal. Deze gebruikerspaden behoren tot een concept dat verder zal worden uitgewerkt aan de hand van de werkwijze die in de volgende paragraaf is gepresenteerd.

Voorbeelden van andere domeinen waarbinnen gebruikerspaden worden gebruikt zijn als volgt:

- Voor het beschrijven van transities in sociotechnische systemen, waarbij elk gebruikerspad een andere combinatie van mogelijke interacties representeert (Geels & Schot, 2007).
- Voor het modelleren en presenteren van informatie voor biologen die onderzoek doen naar biologische functies op verschillende niveaus van detail (Krishnamurthy et al., 2003).
- Voor het samenstellen en toewijzen van interdisciplinaire zorgplannen aan verschillende patiënten waarbij informatieve samenwerking tussen verschillende betrokken zorgprofessionals centraal staat (Atwall & Caldwell, 2002).

Bij de toepassing van gebruikerspaden in deze domeinen is tevens sprake van kritiek, aan de hand van de volgende redenen (Atwal & Caldwell, 2002):

- Bij tijdsdruk zal geen of maar gedeeltelijk gebruik worden gemaakt van gebruikerspaden.
- Het stellen van doelen voor het gebruik van gebruikerspaden kost doorgaans veel tijd.
- Vaardigheden en expertise blijven essentieel bij het gebruik van gebruikerspaden, naast de vereiste commitment om er iets van te willen maken.

Voor implementatie van gebruikerspaden zijn de volgende kritiekaspecten benoemd en die technisch van aard zijn (Krishnumurthy et al., 2003):

- Het koppelen van verschillende gedistribueerde databronnen is problematisch.
- Het formuleren van effectieve zoekopdrachten en visualiseren van resultaten op basis van samengestelde databanken op verschillende niveaus van abstractie, blijft een probleem.
- Het integreren van semantische zoekopdrachten met de mogelijkheid om te navigeren door gegevens in meerdere dimensies op een betekenisvolle manier is lastig.

Uit het voorgaande kan worden geconcludeerd dat het implementeren van taken binnen een gebruikerspad lastig is waardoor veel onderzoek naar de inrichting van databanken is vereist alvorens over te kunnen gaan naar technische implementatie met behulp van beschikbare technologie. Dit betekent dat de gebruikerspaden zullen worden ontworpen als een aanzet voor verder onderzoek en daarom flexibel zullen worden geformuleerd met ruimte voor aanpassingen en verbeteringen.

1.8 Werkwijze van het onderzoek

Om de oplossingsrichtingen uit de vorige paragraaf verder uit te werken zal een werkwijze worden gepresenteerd op basis van onderzoeksvragen. De werkwijze is aan de hand van de volgende onderzoeksvragen samengesteld met als hoofdvraag:

"Wat moeten portaaleigenaren doen om de drempel voor het gebruik van open data te verlagen?"

1. Welke bestaande oplossingen uit de literatuur kunnen het opendataportaal gebruiksvriendelijker maken, rekening houdend met wettelijke richtlijnen?

Om deze onderzoeksvraag te kunnen beantwoorden is een literatuuronderzoek uitgevoerd. Om relevante literatuur af te bakenen is een opendatadeskundige geïnterviewd over open data (zie bijlage 1). De focus van het literatuuronderzoek ligt op de functies om open data te kunnen verwerken op het portaal en op de criteria om het aanbod van open data gebruiksvriendelijker te kunnen maken. Bij de technische oplossingen horen ook wettelijke richtlijnen die komen uit publicatiebladen van de EU over opendatabeleid.

2. Wat ontbreekt er op het huidige portaal aan bestaande oplossingen voor datagebruikers?

Om deze onderzoeksvraag te kunnen beantwoorden zal gebruik worden gemaakt van verschillende analysemethoden om te kunnen bepalen welke bestaande oplossingen uit de literatuur ontbreken in het huidige portaal en dus kunnen worden hergebruikt. Eerst worden voorbeelden gegeven hoe de vier criteria hoofdcategorieën kunnen worden gebruikt om het aanbod van open data op het portaal te verbeteren. De methode die zal worden toegepast om te kunnen bepalen welke functies ontbreken in het huidige portaal van de overheid is als volgt. Eerst is een lijst van functies samengesteld die in de literatuur worden benoemd als onderdeel van een portaal. Vervolgens is voor elke functie uit deze lijst een waarneming gedaan op het portaal. De functies die ontbreken, blijven dan over. De functies uit de literatuur die ontbreken in het huidige portaal zijn geëvalueerd door vijf wetenschappers te bevragen met behulp van

enquêtevragen over het belang van de functies (zie bijlage 2). Ten slotte is met behulp van gebruikersscenario's onderzocht of er ruimte is voor ondersteuning van de datagebruikers bij het verwerken van open data.

3. Met welke vorm van ondersteuning kan het gebruik van open data meer doelgericht worden gestructureerd voor specifieke datagebruikers?

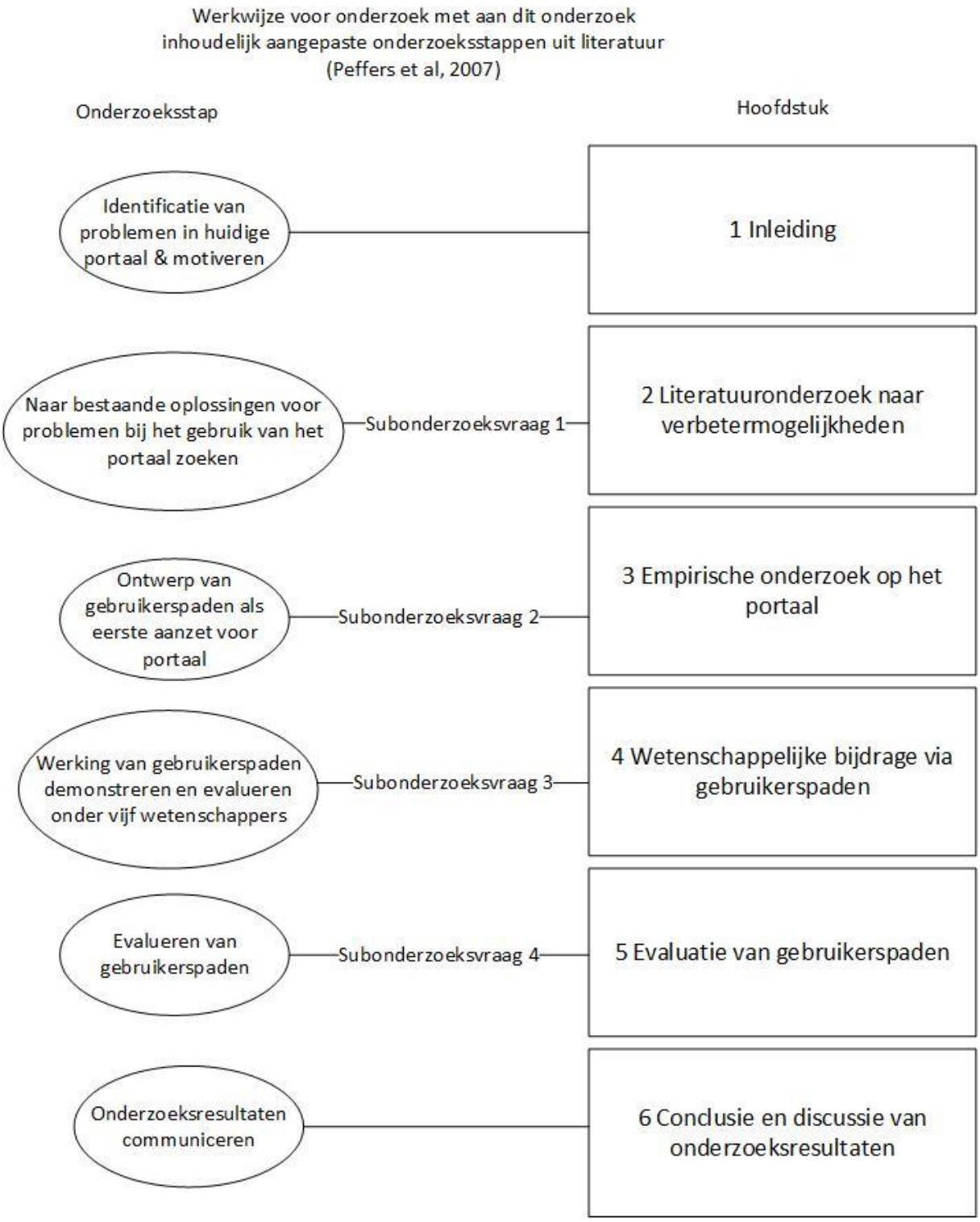
De mogelijkheden van functies, voor het verwerken van open data, worden gebruikt om gebruikerspaden te ontwikkelen. Met behulp van gebruikerspaden kunnen datagebruikers in actie komen om open data te analyseren op een gestructureerde en doelgerichte manier. Dit is een reeks te volgen stappen voor het structureel beantwoorden van onderzoeksvragen en/of hypothesen met behulp van functies. De experimentele gebruikerspaden zullen worden gedefinieerd als eerste aanzet voor verdere ontwikkeling bij gebruikersbehoeftes van datagebruikers.

4. Kloppen de volgorde en het aantal stappen als onderdeel van ieder gebruikerspad ter ondersteuning van datagebruikers?

De poule van respondenten bestaat uit vijf wetenschappers aan wie de gebruikerspaden zijn voorgelegd ter evaluatie. Met behulp van enquêtes is het belang van gebruikerspaden gemeten. Daarnaast is hun ook gevraagd of de volgorde van gebruikerspaden klopt en of er ontbrekende stappen zijn. Op basis van de antwoorden hierop zijn de gebruikerspaden aangepast.

1.9 Afbakening en structuurbeschrijving van vervolgonderzoek

Aan de hand van onderzoekstappen door Peffers et al. (2007) zal het vervolg van de onderzoekswerkwijze worden toegelicht en gelinkt aan hoofdstukken. Het overzicht van de werkwijze voor het verdere onderzoek staat in figuur 3.



Figuur 3: Onderzoekstappen gelinkt aan vervolghoofdstukken (Peffers et al., 2007)

In figuur 3 staat het overzicht van onderzoekstappen en de relatie met hoofdstukken. Deze relaties zijn gerepresenteerd door middel van onderzoeksvragen voor de uiteindelijke aanbeveling in het laatste hoofdstuk. De onderzoekstappen hebben ook een relatie met elkaar die door Peffers et al. (2007) is gedefinieerd als “Design Science Research Methodology (DSRM) Process Model” (Peffers et al., 2007:44). Dit model is oorspronkelijk bedoeld voor onderzoek van informatiesystemen, waar dit onderzoek ook onder valt, aangezien het portaal een informatiesysteem betreft met aanbod van open data. Daarom zijn voor de rest van het

rapport de verschillende onderzoekstappen zijn gelinkt aan respectievelijk de deelvragen en hoofdstukken waarin de resultaten zijn uitgewerkt. In hoofdstuk 2 zullen technische verbetermogelijkheden van het portaal worden weergegeven na een literatuuronderzoek. Ten eerste zullen de criteria voor de databronnen en datasets worden gedefinieerd. Tevens worden wettelijke richtlijnen en maatregelen, die van toepassing zijn om hiermee om te gaan, gepresenteerd. Daarna worden in hoofdstuk 3 de functies uit de literatuur die ontbreken in het huidige portaal gepresenteerd. Vervolgens worden ze geëvalueerd door het bevragen van vijf wetenschappers. Tevens zijn maatregelen geformuleerd om het aanbod aan open data te verbeteren met behulp van criteria. Daarmee kan de tweede deelvraag worden beantwoord. Hoofdstuk 4 is voor beantwoording van de derde deelvraag. Hierbij worden gebruikerspaden geïntroduceerd als nieuwe oplossing. In hoofdstuk 5 zijn de eerste vier gebruikerspaden geëvalueerd onder vijf deskundigen en verbeterd als antwoord op de vierde deelvraag. In hoofdstuk 6 worden de conclusies en een discussie van de onderzoeksresultaten gepresenteerd waarbij tevens suggesties voor toekomstig onderzoek zijn gegeven.

2 Literatuuronderzoek naar verbetermogelijkheden

Het doel van dit hoofdstuk is om het aanbieden en verwerken van open data op het portaal gebruiksvriendelijker te maken met behulp van bestaande oplossingen uit wetenschappelijke literatuur. Daarom is de onderzoeksvraag bij dit hoofdstuk als volgt: *'Welke bestaande oplossingen uit de literatuur kunnen het portaal gebruikersvriendelijker maken, rekening houdend met wettelijke richtlijnen?'* Criteria om het aanbod van open data op het portaal te verbeteren zijn toegelicht in 2.1. Vervolgens zijn in 2.2 Europese richtlijnen bij het aanbieden van open data gepresenteerd en opties om hiermee om te gaan. Daarna is in 2.3 een overzicht van functies voor het portaal gepresenteerd. In 2.4 zijn de conclusies van dit literatuuronderzoek gepresenteerd en is een aanzet naar het volgende hoofdstuk gegeven.

2.1 Aanbod van open data verbeteren met behulp van criteria

Doordat open data vaak op verschillende vormen en manieren kunnen worden aangeboden en gebruikt, is verwerking nodig voordat datasets uit verschillende open databestanden met elkaar kunnen worden gekoppeld. Dat is nodig om toepasbaar en ondersteunend te zijn voor een goedwerkende toepassing in de analysetaak (Dawes et al., 2016). Dit impliceert dat de bruikbaarheid van datasets essentieel is voor de verwerkingstaken in goedwerkende applicaties. Door de beoordeling van open data aan de hand van criteria, kan worden onderzocht hoe de bruikbaarheid van het aanbod kan worden verbeterd. Vervolgens kan het aanbod worden verbeterd door middel van technische methoden die open databestanden transformeren, bijvoorbeeld naar het gewenste formaat. De open data op het portaal kunnen worden beoordeeld door middel van toekenning van scores voor elke criteriumassociatie. De totaalscores die horen bij deze hoofdcategorieën bepalen wat de potentie is om het aanbod gebruiksvriendelijker te maken met behulp van technische methoden. De technische methoden komen later aan bod en maken onderdeel uit van de wetenschappelijke bijdrage van dit onderzoek in de vorm van gebruikerspaden. Deze vier hoofdcategorieën zijn: formaat, metadata, toegang, en kwaliteit (Dawes et al., 2016). Ze bestaan uit bijbehorende criteria die in subparagrafen worden gepresenteerd nadat een overzicht van de hoofdcategorieën is gegeven:

- Formaat voor beoordeling van de manier waarop datasets worden aangeboden
- Metadata voor het beoordelen van de vindbaarheid van datasets
- Toegang voor de beoordeling van toegankelijkheid tot datasets
- Kwaliteit voor de beoordeling van de gegevens in een dataset

Met behulp van de vier hoofdcategorieën kan het aanbod worden beoordeeld op gebruiksvriendelijkheid voor datagebruikers.

2.1.1 Formaat van open data beoordelen

Volgens het 5-sterren model van Berners-Lee (2019) kunnen open data op de volgende manieren verbeterd worden aangeboden op het portaal (Berners-Lee, 2019): Het beschikbaar maken van open data in welk formaat dan ook onder een open licentie met een star als score. De open data in gestructureerde vorm beschikbaar stellen met twee sterren als totaalscore. Het in gestructureerde vorm beschikbaar maken van open data in een niet-gepatenteerd open dataformaat met drie sterren als score. Het gebruik van URI's om gegevensconcepten aanwijsbaar te maken, daarmee kunnen de open data makkelijk in een andere context worden gebruikt met vier sterren als totaalscore. Ten slotte het linken van open data naar andere gegevens voor context met vijf sterren als score. Het koppelen van open data tussen

verschillende informatieconcepten is uiteindelijk de bedoeling bij het effectief gebruik maken van open data. Dawes et al. (2016) heeft een overzicht gemaakt van criteria en geassocieerde scores om het aanbod van open data te kunnen beoordelen. Het heeft een sterk verband met het eerdere 5-sterren model. Om de vorm van de gegevens te beoordelen, is er de hoofdcategorie Formaat. De lijst van criteria en scores voor de categorie Formaat staan in tabel 2.

Tabel 2: Scores met betrekking tot gegevensformaten. Aangepast overgenomen uit (Dawes et al., 2016: 26),

Score	Formaat van gegevens
20%	Open voorwaarden voor de gegevens, maar niet gestructureerd (bijvoorbeeld: JPG-, PDF-, TXT-, Doc-, Open XML-indelingen).
40%	Datasets zijn gestructureerd, maar met gesloten (gepatenteerd) formaat (bijvoorbeeld: XLS)
60%	Datasets zijn gestructureerd en gepresenteerd in open formaten (bijvoorbeeld: XML, CSV in plaats van XLS)
80%	Datasets worden gepresenteerd in Semantische Web Formaten (Open standaard van W3C) (Bijvoorbeeld: RDF, OWL, SPARQL)
100%	Gegevens worden gepresenteerd in gekoppelde open data-indelingen (Semantische Web-indelingen en URI-entiteiten).

De scores zijn gesplitst in het aantal criteria die evenredig zijn verdeeld op basis van cumulatieve scoretoenames met 100% als maximale score en 20% als laagste score. Doordat er vijf criteria zijn benoemd, zijn de stappen cumulatief oplopend in stappen van 20%. Formaat is representatief voor hoe open data wordt gepresenteerd aan datagebruikers. Het betekent dat er een aantal preferenties is voor het aantrekkelijk maken van datagebruik met behulp van de formaattechnologie waarin de open data zijn aangeboden. Bijvoorbeeld in de vorm van een XML-bestand als een hiërarchisch gestructureerde representatie van een datadocument van de overheid. Het formaat is van invloed op het succesvol gebruik van aangeboden open data. Dit komt doordat XML platformonafhankelijk is waardoor het bruikbaar is voor dataoverdracht tussen verschillende applicaties van verschillende platforms. Daarmee kunnen open data waardevol worden voor verschillende groepen eindgebruikers die gebruik maken van verschillende platformapplicaties.

2.1.2 Metadata van databestand beoordelen

Voor het verbeteren van de vindbaarheid van open data is er de hoofdcategorie Metadata. Een lijst van criteria en scores van Metadata staat in tabel 3.

Tabel 3: Metadata voor dataset. Aangepast overgenomen uit (Dawes et al., 2016: 26),

Score	Metadata van databestand

12,5%	De revisiegeschiedenis van de datasets is gegeven
12,5%	Het is mogelijk om eerdere versies van datasets te downloaden
12,5%	De opgegeven frequentie voor het bijwerken van de dataset wordt gepresenteerd
12,5%	De dataset wordt zoveel mogelijk bijgewerkt (subjectief omdat de frequentie afhangt van het type gegevens)
12,5%	De verantwoordelijke persoon is aangegeven
12,5%	De contactgegevens van de verantwoordelijke persoon worden gepresenteerd
12,5%	Het is mogelijk om de structuur van de dataset te bepalen. Als de structuur van de dataset in een afzonderlijk bestand wordt gepresenteerd, komt deze overeen met de dataset
12,5%	De trefwoorden die overeenkomen met de inhoud van de dataset worden gepresenteerd

De scores zijn evenredig verdeeld over de acht criteria en zijn onafhankelijk van elkaar. De totale score wordt bepaald door de toegekende scores van criteria die van toepassing zijn op een dataset bij elkaar op te tellen. Hoe meer metadata des te hoger de totale score. Metadata bestaan uit drie verschillende soorten informatie over open data (Jeffery, z.d.), namelijk: *schema*, *navigatie*, *en associatie*. Met schema wordt gedoeld op het beschrijven van de informatiebron. De *navigatie* gaat over het pad naar een informatiebron. Met *associatie* worden aanvullende informatiegegevens verstrekt in de vorm van een domeinbewuste gebruikersinterface. *Metadata* zijn gegevens over open datasets, voor het beschrijven van datasets, wat belangrijk is voor bruikbare opslag van informatie (Duval et al., 2002). Deze informatie moet kunnen worden opgevraagd of in ieder geval duidelijk worden gemaakt zodat datagebruikers weten wat de gegevens eigenlijk meten in de praktijk. Dit kan belangrijk zijn met het oog op mogelijke links tussen verschillende datasets die verschillen van metadata, maar wel strategisch in een informatiewaardeketen kunnen worden geplaatst (Duval et al., 2002).

2.1.3 Toegankelijkheid van open data beoordelen

Om de toegankelijkheid te kunnen beoordelen is er de hoofdcategorie Toegang. Een lijst van criteria en scores voor toegankelijkheid staat in tabel 4.

Tabel 4: Toegang voor datagebruik. Aangepast overgenomen uit (Dawes et al., 2016: 26)

Score	**Toegankelijkheid van open data**
25%	Mogelijkheid om dataset te downloaden zonder foutmeldingen, via een werkende link naar dataset
25%	Permanente link naar datasets (of API) is beschikbaar

25%	De dataset kan in verschillende formaten worden gedownload
5%	Mogelijkheid om de dataset in HTML-vorm te bekijken
10%	Extra navigatiehulpmiddel aanwezig (voor zoeken, pagina-einden, facetten)
10%	Dataset wordt gepresenteerd in een kaart of andere manieren om te bekijken (niet in tabellen)

De scores zijn verdeeld over meerdere en onafhankelijke criteria. De totale score die de toegankelijkheid van gegevens aanduidt wordt bepaald door de optelling van scores van de criteria die van toepassing zijn op een specifieke databron. De toegankelijkheid van open data is essentieel voor het succes van het portaal. De gegevens moeten foutloos, en navigeerbaar zijn op de plek waar ze staan of waarheen wordt verwezen, met het oog op gebruiksvriendelijkheid. Dit kan ook ergernissen bij toekomstige datagebruikers vermijden en daarmee ook de negatieve ervaringen bij portaalgebruik verminderen.

2.1.4 Kwaliteit van open data beoordelen

Voor het kunnen beoordelen of verbeteren van de kwaliteit van open data is er de hoofdcategorie Kwaliteit. Een lijst van criteria en scores voor de kwaliteit van open data staat in tabel 5.

Tabel 5: Kwaliteit van open data. Aangepast overgenomen uit (Dawes et al., 2016: 26),.

Score	**Kwaliteit van open data**
0, 10, of 20%	Mate van volledigheid van de dataset (percentage lege cellen)
0, 10, of 20%	Aanwezigheid van ongestructureerde informatie (grote stukken tekst in cellen)
0, 10, of 20%	De afwezigheid van niet-gestandaardiseerde gegevens (heterogeniteit van informatie in cellen, bijvoorbeeld het adres, de naam van de persoon en het telefoonnummer of een lijst van waarden die door komma's zijn gescheiden in één cel).
0, 10, of 20%	Het gebruik van verschillende gegevensformaten in dezelfde kolom
0, 10, of 20%	Fouten in de gegevens, bijvoorbeeld verkeerde of vreemde telefoonnummers, spelfouten, afval in de cellen in plaats van zinvolle informatie

De scores zijn optioneel, waarbij er een score kan worden toegekend door de datagebruiker van 0, 10, of 20% bij elke criterium. De totale score voor datakwaliteit is bepaald door optelling van toegekende scores. De kwaliteit van datasets die worden aangeboden is van invloed op de bruikbaarheid van de open data. De datasets moeten compleet worden gepresenteerd of aangeboden. Een dataset die wordt aangeboden als gestructureerd mag geen *ongestructureerde* informatie (tekst) bevatten in de cellen. Ook is een standaard die aangeeft hoe de informatie wordt gepresenteerd een plus, evenals vermijding van heterogene

informatie. Dit komt de overzichtelijkheid en analyseerbaarheid ten goede waarmee waardevolle resultaten kunnen worden verkregen. Fouten en verschillende technologieformaten binnen een dataset komen de kwaliteit van de aangeboden gegevens niet ten goede.

2.1.5 Aanbevelingen om gebruikersvriendelijkheid van open data te verbeteren

De beoordeling voor het gebruik van open data kan worden gedaan door het Kennis- en Exploitatiecentrum Officiële Overheidspublicaties die de opendatapublicaties op het portaal beheert. De kwaliteit van de open data moet op een open en transparante manier op het portaal worden gepresenteerd voor gebruikersvriendelijkheid. Dit komt omdat de beoordeling, op basis van de criteria, het mogelijk maakt om datasets van hoge kwaliteit te onderscheiden van minder bruikbare gegevens, voorafgaand aan toepasbare analyses. Dit bespaart tijd voor gebruikers en resulteert dus in een hogere gebruikersvriendelijkheid van het portaal, omdat het de zekerheid biedt dat op kwalitatieve resultaten kan worden vertrouwd voor doeltreffende analyses. Hiermee kunnen datagebruikers hun inspanningen op het ontwerp van gegevensanalyses concentreren in plaats van op de kwaliteit van open data. Een overzicht van de besproken criteria staat gesorteerd per hoofdcategorie in tabel 6.

Tabel 6: Overzicht van criteria voor het verbeteren van open data. Aangepast overgenomen uit (Dawes et al., 2016: 26)

Formaat	**Metadata**	**Toegang**	**Kwaliteit**
1. Open voorwaarden 2.Gestructureerde gegevens 3. Open indelingen 4.Semantic Web Formats, open standaarden door W3C 5. Gelinkte open dataformaten inclusief Uniform Resource Identifier (URI)	1. Herzienings-geschiedenis 2. Eerdere reeksen gegevens 3. Gespecificeerde periodiciteit van gegevensupdates 4. Data is up-to-date 5. Opgegeven verantwoordelijke persoon 6. Contactdetails van verantwoordelijke persoon 7. Structuur van de dataset in een afzonderlijk bestand 8. Relevante trefwoorden gepresenteerd	1. Werkende link naar dataset 2. Permanente link naar gegevens 3. Downloaden in verschillende formaten 4. Te bekijken in HTML-vorm 5. Extra navigatietool om door gegevens te navigeren 6. Dataset wordt gepresenteerd op verschillende manieren om te bekijken	1. Volledigheid van dataset zoals lege cellen 2. Ongestructureerde informatie zoals tekst in cellen 3. Niet-gestandaardiseerde gegevens vanwege heterogeniteit van informatie in cellen 4. Verschillende gegevensformaten in dezelfde kolom 5. Fouten in gegevens

De criteria zijn aan de hand van de vier hoofdcategorieën aangeduid. In tabel 6 is er een duidelijk verschil tussen meetbare intervalvariabelen te zien. De toegankelijkheid en de metadata kunnen worden gebruikt voor het beoordelen van de gebruiksvriendelijkheid voordat de open data zijn geopend, terwijl het formaat en kwaliteit voor de gebruikersvriendelijkheid

van datasets kunnen worden gebruikt. De volgende aanbevelingen zijn toegelicht en bedoeld om de gebruiksvriendelijkheid van open data via het portaal te kunnen verbeteren:

1. Formaat: open dataformaten met URI's verdienen de voorkeur voor de uniforme en unieke benaming van databestanden. Oftewel, open standaarden voor de databestanden om gebruikt te worden als open data en om de koppeling met andere gestructureerde open data gemakkelijker te kunnen maken.
2. Metadata: met behulp van relevante trefwoorden die representatief zijn voor de databestanden of webservices. Dit kan de vindbaarheid van open data via de zoekmachine op het portaal verbeteren. Ook de structuur van de dataset in een afzonderlijk bestand kan hierbij helpen omdat daarmee kan worden bepaald welke informatie uit de open data kan worden gehaald.
3. Toegankelijkheid: de mogelijkheid om open data via verschillende werkende links en dataformaten te kunnen downloaden. Daarnaast is er de mogelijkheid om door het aanbod van open data te navigeren, bijvoorbeeld via een overzichtelijke kaart. Hierbij worden de gegevens in HTML-formaat gepresenteerd, als de standaard web-indeling. Hiermee hebben datagebruikers de mogelijkheid om open data te observeren alvorens het al dan niet te gaan gebruiken voor verdere data-analyses, zonder er software als gereedschap voor te hoeven gebruiken of een externe tool.
4. Kwaliteit: aangezien de kwaliteit van open data kan verschillen in volledigheid, structuur, standaardisatie, of heterogeniteit van informatie in cellen is beoordeling nodig. Ook een verkeerd aanbod van open data, bijvoorbeeld met meerdere formaten in een kolom, dubbele of foutieve gegevens, is ongewenst voor doeltreffende data-analyses. Om de kwaliteit te verhogen moet daarom worden nagegaan hoe de kwaliteit van datasets is voordat het kan worden gebruikt voor analyses. Bijvoorbeeld met behulp van frequentieanalyse, om uitschieters te identificeren, die de gemiddelden beïnvloeden zonder valide reden. Door correlaties tussen variabelen te onderzoeken, hoeven alleen de variabelen uit een dataset te worden geselecteerd die interessant zijn voor analyses en de rest van de gegevens hoeven dan verder geen rol van betekenis te spelen. Dit maakt de kwaliteit van gelinkte datasets beter door middel van selectie. Verder kunnen ook beschrijvende statistische informatiewaarden van de datasets worden bestudeerd om in te schatten of ze geschikt zijn voor specifieke analysedoeleinden. Bijvoorbeeld het aantal dataobjecten oftewel rijen, mediaan, gemiddelde, modus, standaarddeviatie, minimale- en maximale waarden, en verdeling van de waarden (kansberekening). Maar ook overeenkomstige karakteristieken van groepen observaties in een dataset behoren tot informatiewaarden in een dataset. Dit kan met behulp van technische functies worden toegepast om datasets te onderzoeken.

Het gebruik van open data is ook gebonden aan wettelijke kaders die juridische implicaties kunnen hebben voor de langere termijn. Daarom moet hiermee rekening worden gehouden bij het aanbieden van open data op het portaal. In de volgende paragraaf zijn drie EU-richtlijnen geanalyseerd voor de afwegingen die de overheid moet maken bij het ontsluiten van open data en de aanbevelingen om hiermee om te gaan.

2.2 Wettelijke richtlijnen voor het aanbieden van open data

In deze paragraaf worden de wettelijke kaders geanalyseerd met behulp van EU-richtlijnen om te begrijpen welke afwegingen dataproviders moeten maken bij het aanbieden van open data. De vraag vanuit dataproviders om wel of niet open data te publiceren op het overheidsportaal wordt bepaald door een aantal factoren, te weten: eigenaarschap van data, embargoperiode, transparantie, datatoegankelijkheid & licenties, datagevoeligheid, datakwaliteit, en metadata. Dit komt omdat er consequenties zijn verbonden aan de ontsluiting van gegevens die van invloed zijn op het succes van open data. Zo kan het voorkomen dat een overheidsorganisatie gegevens ontsluit die resulteren in negatieve publiciteit voor een overheidsorganisatie met als oorspronkelijk doel meer transparantie. Verder kan het beperken van gegevens voor bepaalde stakeholders in minder creatieve hergebruik van gegevens resulteren omdat ze zich gelimiteerd voelen en dus minder voordelen van de ontsluiting ervan als gevolg ondervinden. Door een lage kwaliteit van gegevens zullen er ook minder goede beslissingen genomen en conclusies getrokken worden door datagebruikers als de doelgroep van dataproviders (Zuiderwijk & Janssen, 2015).

De gebruikerservaring bij het portaal is hierbij niet erkend of goed beschreven. Om dit te verbeteren zal de focus op de gebruikersvriendelijkheid voor datagebruikers worden gelegd. Hierbij wordt gestreefd naar een selfservicesysteem waarmee datagebruikers zichzelf kunnen bedienen om vooropgestelde onderzoeksvragen en/of hypothesen te beantwoorden. De gebruikersvriendelijkheid van de aangeboden open data kunnen worden gemeten op basis van de criteria die in tabel 6 staan beschreven. Hierbij moeten open data beschikbaar zijn gemaakt op het portaal voor permanent hergebruik. Er zijn drie Europese richtlijnen van toepassing op de openstelling van informatie, als open data, door de overheid. Deze juridische documenten zijn oorspronkelijk in publicatiebladen van de Europese Unie (EU) gepresenteerd. Ze hebben betrekking op het beleid door nationale lidstaten binnen de EU, waaronder ook Nederland, over het hergebruik van open data. Deze richtlijnen vormen het wettelijk kader voor het gebruik van open data.

Ten eerste, naar aanleiding van Richtlijn (EU) 2016/680 van het Europees Parlement en de Raad van 27 april 2016 over de verwerking van persoonsgegevens voor strafvervolging en onderzoek na strafbare feiten, waar bescherming van natuurlijke personen centraal staat (PbEU 2016, L119/89). Privacygevoelige informatie vraagt om zorgvuldigheid bij het beschikbaar stellen van open data omdat de veiligheid van personen in het geding is. Hiervoor moeten dataverzoeken zorgvuldig worden behandeld en daarmee onbedoelde consequenties voorkomen door risico's te minimaliseren van onrechtmatig lekken van privacygevoelige informatie. De ontsluiting van open data mag niet ten koste gaan van de veiligheid van personen. Om privacygevoelige gegevens te beschermen kan worden overgegaan tot het anonimiseren van gegevens waardoor het niet gebonden is aan natuurlijke personen.

Ten tweede, naar aanleiding van Mededeling (EC) 2014/C 240/01 van de Commissie van 27-7-2014 over exploitatie van hergebruikte overheidsgegevens via gestandaardiseerde licenties (PbEU 2014, C240/1). Voor het beschikbaar stellen van open data worden kosten gemaakt aangezien de gegevens door in ieder geval een ambtenaar moet worden geüpload op het portaal. De voorwaarden die de overheid stelt aan het beschikbaar stellen van open data kan via gestandaardiseerde licenties transparant worden gemaakt als achtergrondinformatie bij het behandelen van dataverzoeken. Met behulp van open licenties voor open data op het portaal en kostentoerekening voor het gebruik van webservices kunnen de kosten voor het operationeel houden van webservices op externe servers worden gedekt.

Ten derde, naar aanleiding van Richtlijn (EU) 2013/37/EU van het Europees Parlement en de Raad van 26 juni 2013 tot wijziging van Richtlijn 2003/98/EG wat betreft het hergebruik

van overheidsgegevens (PbEU 2013, L175/1). Het beschikbaar stellen van open data kan de overheid open en transparant maken. Maar geslotenheid brengt minder risico's met zich mee voor de overheid, waar onbedoelde consequenties mee voorkomen kunnen worden. De overheid wil zo min mogelijk risico's lopen, waarbij het initiatief bij de aanvragers van open data is gelegd. Met behulp van sociale mediacampagnes met de vraag 'Wat wil ik weten van de overheid?' kan aan de hand van de verkregen reacties worden bepaald in welke open data het publiek is geïnteresseerd om dat vervolgens te ontsluiten voor dit publiek.

De verschillen in het aanbod van open data kunnen worden verwerkt met behulp van beschikbare functies. Deze functies kunnen het gebruik van open data gebruikersvriendelijker en daarmee ook aantrekkelijker maken voor nieuwe datagebruikers. Het huidige beleid voorziet niet in de behoefte en stimulering van kansen uit open data. De functies die datagebruikers in gelegenheid stellen om open data te verwerken staan in de volgende paragraaf beschreven.

2.3 Analyseren van functies uit literatuur

In Zuiderwijk et al. (2013) is een raamwerk gepresenteerd waarbinnen een aantal functionaliteiten zijn opgenomen om open data infrastructuren te vergelijken. In het geval van het raamwerk zijn dit: De opendatahub van de Europese Unie (EU), het commerciële open dataplatform Junar met haar oorsprong in Argentinië, en ENGAGE 2.0 open data-infrastructuur (Zuiderwijk et al., 2013).

De EU open data infrastructuur ondersteunt hoofdzakelijk: het downloaden, analyseren, en visualiseren van gegevens. ENGAGE 2.0 focust op het gebruiken en bediscussiëren van data, en Junar ondersteunt het publiceren van open data. Het open dataproces bestaat volgens de auteurs uit het proces waarbij de creatie van data, publicatie van data, vinden van data, gebruiken van data, linken van data, hergebruik van data, en discussie daarvan deel uitmaakt. Op basis hiervan kunnen verschillende open data infrastructuren worden vergeleken, als gekeken wordt naar de functionaliteiten die het open dataproces mogelijk maken via open data infrastructuur.

De belangrijkste functionaliteiten die door Zuiderwijk et al. (2013) benoemd worden om open data infrastructuren te vergelijken zijn: registratie/login, zoekfunctie met zoekwoorden, kwalitatieve feedback op open data, kwantitatieve feedback op open data, gebruikerssupport, samenwerkingsomgeving, en versiemanagement. Verder zijn de volgende tools geformuleerd: formaatverandering, opschonen, analyseren, curatie, visualisatie, linken, monitoren van gegevensgebruik, en monitoren van feedback op gegevens (Zuiderwijk et al., 2013). Onder andere deze functionaliteiten kunnen gebruikt worden om de gebruikersvriendelijkheid van het portaal te verbeteren indien een of meerdere van de functionaliteiten ontbreken in het huidige portaal. De volgende functies kunnen een onderdeel zijn van het portaal van de overheid (Zuiderwijk et al., 2014; Zuiderwijk, 2015):

F1. Zoekfunctie met behulp van gebruikersinvoer via trefwoorden, filters en data attributen

F2. Downloaden van dataset via link

F3. Registreren op portaal als nieuwe gebruiker

F4. Data kunnen opschonen, analyseren, verrijken en koppelen

F5. Gebruikerspaden van open data raadplegen inclusief het gebruik en links tussen verschillende datasets

F6. Visualiseren van resultaten na toegepaste data-analysetechnieken

F7. Feedback geven aan gegevensproviders en beleidsmakers

F8. Datasets kunnen uploaden, door dataproviders

F9. Metadata kunnen toevoegen, door dataproviders

F10. Gebruikers en datasets kunnen volgen

F11. Bevindingen aan de hand van uitgevoerde analyses en links naar datasets delen via sociale mediakanalen

F12. Versiemanagement voor datasets, door dataproviders

F13. Datakwaliteit kunnen beoordelen via het portaal, door datagebruikers

F14. Tutorials of zelfstudiemodulen over het aanbieden en analyseren van open data

F15. Het kunnen bediscussiëren van datasets of wat er kan worden geleerd van het gebruik van open data

De associaties van de lijst van functies met auteurs uit de literatuur staan in tabel 7.

Tabel 7: Overzicht van verschillende auteurs en benoemde functionaliteiten

	Auteur						
Functie	Zuiderwijk (2015)	Zuiderwijk et al. (2014)	Zuiderwijk et al. (2013)	Thorsby et al. (2016)	Colpaert et al. (2013)	Alexopoulos et al. (2014)	Iamamphai et al. (2016)
F1. zoeken	+	+	+	+	+		+
F2. downloaden	+	+	+			+	+
F3. registreren	+	+	+		+		
F4. analyseren	+	+	+	+		+	
F5. gebruikerspaden	+	+		+			
F6. visualiseren	+	+	+			+	+
F7. terugkoppeling	+	+	+		+	+	+
F8. uploaden	+		+		+	+	
F9. metadata	+	+			+	+	
F10. gebruikers volgen	+				+		
F11. analysebevindingen delen	+				+		+
F12. versiebeheer	+		+		+	+	
F13. datakwaliteitsbeheer	+	+	+			+	
F14. tutorial	+	+		+			
F15. interactie	+	+	+		+	+	+

De functies uit de literatuur zijn benoemd door verschillende auteurs. Hiervan zijn de relaties tussen de auteurs en functies gepresenteerd. In tabel 7 zijn in totaal zeven publicaties geanalyseerd en vijftien functies gepresenteerd die onderdeel kunnen zijn van een portaal dat open data publiceert. Wat hierbij opvalt, is dat het aantal publicatieassociaties verschilt per functie. De meest associaties (6) zijn met de volgende functies: F1, F7, en F15. Deze functies betreffen de volgende functionaliteiten: zoeken, feedback geven, en discussie. De functies met vijf publicatieassociaties betreffen: F2, F4, en F6. Deze functies zijn representatief voor: downloaden, analyseren, en visualiseren. De zes functies met de meeste publicatieassociaties hebben een hoge prioriteit als onderdeel van het portaal. Uiteindelijk zijn alle functies F1-F15 geëvalueerd onder deskundigen (zie bijlage 2). Van deze functies ontbreekt een aantal in het huidige portaal. Deze zijn in de volgende paragraaf gepresenteerd.

2.4 Conclusie van onderzoeksresultaten na literatuuronderzoek

Aan de hand van de resultaten na uitvoering van het literatuuronderzoek kan het volgende worden geconcludeerd. Voor het monitoren van gebruikersvriendelijkheid zijn vier criteria gevonden in de literatuur die kunnen worden gebruikt om open data gebruikersvriendelijker aan te bieden. Verder zijn drie wettelijke richtlijnen gevonden waarmee portaaleigenaren rekening dienen te houden bij het implementeren van functies. Op basis van het literatuuronderzoek zijn vijftien relevante functies naar voren gekomen die onderdeel kunnen zijn van een portaal. In het volgende hoofdstuk zal verder worden onderzocht welke ruimte voor verbetering er is om het huidige portaal gebruiksvriendelijker te maken voor datagebruikers.

3 Empirisch onderzoek op het portaal

Het doel van dit hoofdstuk is om de volgende onderzoeksvraag te kunnen beantwoorden: *'Wat ontbreekt er op het huidige portaal aan bestaande oplossingen voor datagebruikers?*'. Om deze onderzoeksvraag te beantwoorden is onderzocht welke resultaten van het literatuuronderzoek geschikt zijn voor implementatie. Zo zijn in 3.1 via elk van de vier hoofdcategorieën van criteria suggesties gedaan om het aanbod van open data op het portaal gebruikersvriendelijker te maken. Vervolgens is onderzocht welke functies ontbreken in het huidige portaal in 3.2. Deze functies zijn geëvalueerd onder vijf wetenschappers als de respondenten in 3.3. Verder is de huidige opzet van het portaal in de vorm van gebruikersscenario's uitgewerkt om te onderzoeken waar ruimte is voor verdere ondersteuning van datagebruikers in 3.4. De conclusie van de onderzoeksresultaten zijn toegelicht in 3.5 met een opzet naar hoofdstuk 4 voor ontwerp van gebruikerspaden.

3.1 Suggesties om het aanbod van open data te verbeteren

In deze paragraaf worden de verschillen in het aanbod van open data aangepakt voor verbetering van de gebruikersvriendelijkheid van de open data op het portaal. Dit is gedaan aan de hand van vier hoofdcategorieën waaronder het aanbod van open data kan worden beoordeeld op gebruikersvriendelijkheid. Er kunnen vier hoofdcategorieën van criteria worden aangeduid: formaat, kwaliteit, toegang en metadata (Dawes et al., 2016). Ten eerste moeten datagebruikers om kunnen gaan met de verschillende dataformaten van de bestanden die open data bevatten. Bijvoorbeeld: verschillende gegevenstypen dienen op elkaar afgestemd te worden wat kan door middel van conversie naar een database bijvoorbeeld. Hier kunnen bij het koppelen met andere datasets kolommen worden toegevoegd die met behulp van conversie vanuit andere datasets worden ingeladen in dezelfde database.

Ten tweede moet worden omgegaan met kwaliteitsverschillen tussen gestructureerde datasets. Hierbij kunnen, door het toepassen van frequentieanalyse, uitschieters worden verwijderd. Verder kan door middel van correlaties tussen variabelen worden besloten om relevante datasets samen te voegen. Of met behulp van beschrijvende statistiek, uitgedrukt in informatiewaarden over de dataset, bepalen hoe de gegevens verdeeld zijn en of dit overeenkomt met toepassing van bepaalde technieken waarvoor bijvoorbeeld vereist is dat de dataset statistisch normaal verdeeld is. Ook het vinden en verwijderen van dubbele gegevens en corrigeren van fouten in de dataset kunnen met behulp van analysetechnieken worden geïdentificeerd en uitgevoerd. Verder kunnen ontbrekende waarden worden toegevoegd, bijvoorbeeld op basis van het uitgerekende gemiddelde. Daarnaast het standaardiseren op gegevensformaten die worden gebruikt binnen de dataset met behulp van eenheden, zoals datums, maten, en geldwaarden.

Ten derde, het omgaan met de toegang tot de bestanden, bij voorkeur via de aanwezigheid van permanent werkende links naar de bestanden en het toegang verlenen via een webservice. Datagebruikers moeten gemakkelijker door de open data kunnen navigeren en waarnemen via webpagina's als standaard met een overzicht van URI's voor verschillende datasets die relevant zijn binnen een domein.

Ten vierde, het omgaan met metadata verschillen op het portaal. De beschrijving van de beschikbare datasets klopt niet altijd en is niet op dezelfde wijze geformuleerd, waardoor het onduidelijk is of de dataset wel geschikt is voor de datagebruikers zonder verder onderzoek te doen. Bijvoorbeeld door de gegevens te interpreteren voor zover dit mogelijk is of door contact op te nemen met de dataprovider over de metadata. Daarom moeten de zoektermen

representatief zijn voor de dataset. Dit kan met behulp van de informatiestructuur van de dataset die de context van de dataset kan representeren op een interpreteerbare manier.

Behalve het kunnen verbeteren van het aanbod van open data op het portaal is gelegenheid om open data te kunnen verwerken ook belangrijk voor datagebruikers die open data effectief willen inzetten voor hun gebruiksdoelen. Hiervoor zal in de volgende paragraaf worden bepaald welke functies, die in de literatuur zijn benoemd, ontbreken in het huidige portaal.

3.2 Functies die ontbreken op het huidige portaal

Het huidige portaal van de overheid zal worden geanalyseerd op basis van de functies die uit de literatuur naar voren zijn gekomen. De methode die zal worden toegepast om te kunnen bepalen welke functies ontbreken in het huidige portaal van de overheid is als volgt. Aan de hand van de volgende hypothese zal voor elke functie uit de lijst worden getoetst of deze ontbreekt op het portaal van de overheid. H0= 'Functie X is ontbrekend op het portaal?'. Met alternatief H1= 'Functie X is niet ontbrekend op het portaal' voor functie 1 tot en met 15. De toetsing is door middel van waarneming op het portaal (data.overheid.nl). De functies uit de literatuur die ontbreken in het huidige portaal staan in tabel 8.

Tabel 8: Ontbrekende functies in huidige portaal

Functies	**Ontbrekend in huidige portaal (data.overheid.nl)**
F1. zoeken	Nee
F2. downloaden	Nee
F3. registreren	Nee
F4. analyseren	Ja
F5. gebruikerspaden	Ja
F6. visualiseren	Ja
F7. terugkoppeling	Nee
F8. uploaden	Nee
F9. metadata	Nee
F10. gebruikers volgen	Ja
F11. conclusies delen	Ja
F12. versiebeheer	Ja
F13. data kwaliteitsbeheer	Ja
F14. tutorial	Ja
F15. interactie	Ja

De functies uit de literatuur zijn vergeleken met de functies in het huidige portaal. Hieruit is gebleken dat een aantal functies uit de literatuur ontbreken in het huidige portaal. De functies: F4, F5, F6, F10, F11, F12, F13, F14, en F15 ontbreken in het huidige portaal. Dit is bepaald door het overzicht aan mogelijke functies voor het portaal in tabel 7 te vergelijken met de beschikbaarheid van functies op het portaal van de overheid. Hieruit kunnen we beredeneren dat er in totaal negen functies zijn waarmee het portaal kan worden verbeterd om het aantrekkelijk te maken voor toekomstige datagebruikers.

Voor elke functie is toegelicht wat de toegevoegde waarde is voor de datagebruikers (Zuiderwijk et al., 2014; Zuiderwijk, 2015):

1 Analyseren van open data: met deze functie kunnen open data worden geprepareerd om het te kunnen verrijken door het te linken met andere datasets bijvoorbeeld, waarna gekozen data-analyses kunnen worden uitgevoerd.
2 Gebruikerspaden kiezen: dit is een zoekfunctie waarmee naar een geschikt gebruikerspad kan worden gezocht op basis van karakteristieken zoals de manier waarop gegevens kunnen worden verworven. Een gebruikerspad is afhankelijk van welke specifieke datasets een datagebruiker wil gebruiken.
3 Visualiseren van data-analyse resultaten: deze functie kan worden gebruikt om analyseresultaten interpreteerbaar te maken voor een breed publiek. Een reeks getallen en cijfers zegt minder dan een visualisatie waarin de resultaten betekenisvol staan gesorteerd en waarmee afwegingen kunnen worden gemaakt voor bijvoorbeeld strategische keuzes. Ook zijn de analyseresultaten beter deelbaar met behulp van visualisatie met anderen zonder voorkennis van de open data.
4 Volgen van datagebruikers: via deze functie kan een datagebruiker kiezen wie hij/zij gaat volgen op het portaal voor degenen die volgers willen in plaats van anoniem gebruikmaken van het portaal. Men kan bijvoorbeeld inspiratie opdoen, leren van elkaar en samenwerking starten door taken te verdelen via het portaal.
5 Analysebevindingen delen via sociale media: deze functie maakt interactie via meerdere bestaande kanalen mogelijk waardoor het bereik van bevindingen groter is. Ter onderbouwing kan worden verwezen naar uitgevoerde analyses op het portaal voor bijvoorbeeld feedback of interactie.
6 Versiebeheer: deze functie behoort tot de metadata van datasets aangezien het de geschiedenis van de totstandkoming van de dataset weergeeft. Hiermee kunnen datasets worden aangepast door geautoriseerde dataproviders en kunnen datagebruikers eventuele aanpassingen meenemen in hun analyses, bijvoorbeeld wanneer de oorspronkelijk gebruikte dataset verouderd is.
7 Data kwaliteitsbeheer: deze functie maakt het mogelijk om de kwaliteit van de dataset te beoordelen door datagebruikers een score te laten geven die ontleend is van het kwaliteitscriterium voor datasets. Hiermee kan de dataprovider de datasets verbeteren waarbij de scores worden bijgehouden.
8 Tutorials aanbieden: voor het gebruik van open data kunnen tutorials worden aangeboden waarmee vaardigheden worden aangeleerd die ontbreken. Doel is het gebruik van datasets te bevorderen middels interactieve onlinecursussen. Bijvoorbeeld een model maken, prepareren van open data of software ontwikkelen om het API-gebruik te bevorderen.
9 Interactiemechanisme: op het portaal kan met andere datagebruikers worden gecommuniceerd, bijvoorbeeld bij het uitvoeren van data-analyses. Hiermee kan worden samengewerkt via het portaal op basis van dezelfde datasets. Ook kunnen taken worden verdeeld als meerdere datagebruikers dezelfde inzichten hebben, kan men elkaar vinden via het portaal en samenwerkingen starten.

Deze functies zijn bruikbaar voor het portaal, maar er zijn wel nadelige kanttekeningen te plaatsen aan het gebruik ervan voor dit onderzoek:

- Implementatie van functies vraagt om investeringen van portaaleigenaren in softwareontwikkeling.

- De functies maken het portaal complexer in omvang en daarmee lastiger om te onderhouden.
- Door de functies nemen datastromen op het portaal toe in het geval ze gebruikt worden, waardoor hardware-middelen zoals extra servers, benodigd zijn voor extra verwerkingscapaciteit. Dit laatste is ook afhankelijk van het aantal bezoekers die tegelijk gebruik maken van het portaal.
- De functies zijn niet allemaal even relevant voor eindgebruikers omdat gebruikersbehoeften kunnen verschillen. Bovendien zijn niet alle functies bedoeld voor data-analyse maar juist als ondersteuning bij het proces van waardecreatie met behulp van verschillende databronnen, waaronder extern verkregen informatie.

De voorgestelde functies worden geëvalueerd in de volgende paragraaf door wetenschappers te interviewen, om hun waarde voor datagebruikers te bevestigen.

3.3 Evalueren van ontbrekende functies onder deskundigen

Om de ontbrekende functies te valideren is een enquête gehouden onder vijf deskundigen. De deskundigen zijn geselecteerd omdat zij allemaal experts zijn op het gebied van open data met verschillende expertises of specialisaties, die PhD-onderzoek doen op het gebeid van open data aan de Technische Universiteit Delft. Daarmee kunnen ze als representatief worden beschouwd bij het evalueren van gebruikerspaden als nieuw hulpmiddel voor eindgebruikers van open data. De verschillende expertises van de respondenten zijn als volgt:

1. Het produceren en testen van modellen voor de beantwoording van onderzoeksvragen en/of hypothesen.
2. Open data gebruiken voor wetenschappelijke analyses, om trends te ontdekken in verkeersstromen voor logistieke inzichten.
3. Open data gebruiken als een bron voor van vertrouwen en transparantie voor de overheid en bedrijfsleven, voor nieuwe bedrijfslocaties en waar de maatschappij van kan profiteren.
4. Open data gebruiken als middel voor het creëren van meer voordelen.
5. Raamwerk ontwikkelen waarin verschillende databronnen zijn gebundeld om het analyseren van big data gebruiksvriendelijker te maken.

Allereerst is de relevantie van elke functie bepaald door deze deskundigen te interviewen. De opinie van deskundigen over het belang van de ontbrekende functies staan in tabel 9. Het belang van functies is gemeten op een schaal van 1 tot 5, met 1 onbelangrijk en 5 heel erg belangrijk. De mening van de respondenten over het belang van de aangegeven ontbrekende functies staat in tabel 9.

Tabel 9: Belang van ontbrekende functie, Schaal: 1= niet belangrijk en 5= erg belangrijk

Functie	**Deskundige 1**	**Deskundige 2**	**Deskundige 3**	**Deskundige 4**	**Deskundige 5**	**Gemiddelde**
F4	4	5	4	5	5	4,6

F5	3	4	1	5	4	3,4
F6	3	5	5	5	4	4,4
F10	3	4	3	5	4	3,8
F11	4	3	2	2	2	2,6
F12	5	5	5	5	5	5
F13	4	4	5	5	5	4,6
F14	4	5	5	5	4	4,6
F15	4	4	5	1	4	3,6

Bij elk van de functies zijn de toegekende scores door de vijf respondenten opgeteld en gedeeld door vijf. Dit is om in de meest rechterkolom de gemiddelde score uit te rekenen als meting voor het functiebelang. In tabel 9 valt F11 op door de lage score van 2,6. Met uitzondering van deze zwak scorende functie, onder de respondenten, scoren de functies F12, F4, F13, F14, en F6 op volgorde van hoog naar minder, bovengemiddeld op belangrijkheid. Met uitschieter van F12 dat versiebeheer voor datasets (F12) representeert. De laatstgenoemde functie wordt dus essentieel gevonden voor het portaal met de maximale score van 5. Dit is anders met functie F11, wat onder de experts gemiddeld het laagste scoort (2,6). Dus F11 heeft een lage prioriteit voor wat betreft implementatie van ontbrekende functies volgens het normenkader. Over het algemeen is er geprioriteerd, naar aanleiding van de ranking op basis van de scores, van hoog naar minder. Met uitzondering van functies F4, F13 en F14 die even belangrijk zijn met een score van 4,6. Er is dus geen verschil in prioriteit voor de implementatie van deze drie functies ten opzichte van de rest van de gemeten functies, ze zijn dus even belangrijk bevonden. Op basis van de scores zijn de functies geprioriteerd, en dit is representatief binnen dit onderzoek doordat de evaluatie op basis van de vijf deskundigen is gedaan. Maar hierbij is geen rekening gehouden met de portaaleigenaren op basis van budget bijvoorbeeld. In 6.4 zal bij bespreking van de beperkingen van dit onderzoek in het laatste deel van de conclusie hier nader op worden ingegaan. De prioriteit van hoog naar laag voor implementatie van de functies is als volgt, en waarbij de volgorde van de functies F4, F13, F14 verder niet uitmaakt na F12 en voor F6 in het rijtje van functies:
F12, F4, F13, F14, F6, F10, F15, F5, F11.

3.4 Analyseren van gebruikersscenario's in huidige opzet van het portaal

Er zijn momenteel drie hoofdgebruikersscenario's voor het gebruik van het portaal van de Nederlandse overheid (zie data.overheid.nl). Ze zijn als volgt gedefinieerd (Open data van de overheid, 2018). De volgende stap geldt voor alle drie gebruikersscenario's: Stap 0: ga naar de URL-locatie (data.overheid.nl), dit portaal is online beschikbaar voor gebruik, via Data. ➞

Gebruikersscenario 1: indienen van dataverzoek

Stap 1: ga naar het tabblad 'Overzicht alle dataverzoeken' en zoek naar een gegevensverzoek dat vergelijkbaar is met de huidige wensen. Deze lijst kan worden gesorteerd op inhoud met

behulp van trefwoorden, status (open of behandeld), fase, resultaten worden overhandigd (zie lijst), thema (van gegevensverzoek) en actiehouder.
Stap 2: ga naar het tabblad 'Dataverzoek indienen' en vul de specificaties van het gegevensverzoek op het online formulier in: titel, gevraagde gegevens, thema, gewenste formaat, gewenste start- en einddatum van gegevens, mogelijke gegevensbron: overheidsorganisatie, of er problemen waren bij het zoeken, zoekpogingen, waar je de data voor wilt gebruiken, doelstelling van het dataverzoek, of het een commercieel doel betreft, en persoonlijke gegevens die niet publiekelijk zichtbaar worden getoond (naam, organisatie, e-mail, telefoonnummer).
Stap 3: ga naar 'Toelichting bij dataverzoeken' raadpleeg eventueel aanvullende informatie over de afhandeling van dataverzoeken.

Gebruikersscenario 2: dataset downloaden van het portaal
Stap 1: relevant thema kiezen
Stap 2: zoeken naar dataset via zoekbalk, invoeren van trefwoorden en klikken op de zoekknop.
Stap 3: verfijnen van de zoekresultaten op basis van onder andere: status (beschikbaar/ in onderzoek/ gepland), toegang (publiek/ beperkt), data-eigenaar, thema (geselecteerd), subthema, hoogwaardige dataset (ja/ nee), licenties (CC-0, CC-BY 4.0, Public Domain, CC-BY-SA, Onbekende Licentie, CC-BY 3.0), Labels, Taal, Bronformaat: CSV, XLS, PDF, JSON, Atom Feed, ESRI Shape bestand, ZIP, XML, WebMap-service, HTML, XLSX, WebmappingTiling-service, webfunctiedienst, broncatalogus.
Stap 4: kiezen van bestand en downloaden voor gebruik

Gebruikersscenario 3: toegang tot webservice
Stap 1: specifieke software of tool is vereist voor gebruik van de webservice.
Stap 2: selectie van beschikbare webservice, wat gewenst is voor de databaseapplicatie. Voorbeelden van bestaande webservices die worden beheerd door overheidsorganisaties zijn:
- Basis wettenbestand (applicatie/ Soap (Simple Object Access Protocol) + XML), met betrekking tot wet- en regelgeving in Nederland op basis van verschillende attributen zoals: datacreatie, datum voorlopige aanvraag, datum_introductie, datum_materiaalaanvraag, datum_invoer, datum_replicatie, reikwijdte_terug etc.
- RDW-certificeringen in NL, geregistreerde voertuigen met de volgende voorbeeldattributen: registratienummer, voertuigtype, merk, handelsnaam, APK-vervaldatum, registratiedatum, bruto BPM, aantal stoelen, kleur, cilinders, massa leeg voertuig, maximaal toegestane massa van het voertuig, massa van de verkeerswaardigheid enz.
Stap 3: webservice-aanvraag indienen via onlineapplicatie.
Stap 4: onlineapplicatie maakt gebruik van ontvangen data, voor verdere verwerking van zijn eigen proces, of voor presentatie over de toepassing van relevantie voor applicatiegebruikers.

Na het analyseren van de drie huidige gebruikersscenario's die representatief zijn voor het huidige portaal van de overheid, kunnen we concluderen dat de huidige scenario's onvolledig voorzien in de verwerking van open data. Ze zijn meer gericht op het aanbieden van open data in plaats van op het gebruikersproces. Gebruikerspaden kunnen hierin verandering brengen aangezien ze vanaf het begin op het gebruikersdoel van de gebruikers zijn gefocust door middel van onderzoeksvraag en/of hypothese bijvoorbeeld. De open data dienen wel op de juiste manier te worden aangeboden, en dit is de taak van de dataproviders, die open data aanleveren. De dataproviders kunnen daarom een essentiële rol spelen bij het verbeteren van

de gebruikersvriendelijkheid van het portaal via het aanbieden van meer en betere open data. Het structureren van het open datagebruik kan dit stimuleren met een hogere vraag naar open data.

3.5 Conclusie na empirisch onderzoek op het portaal

Aan de hand van de vier hoofdcategorieën van criteria voor aanbod van open data zijn suggesties gedaan om dit aanbod gebruiksvriendelijker te maken. Verder is de lijst van functies gereduceerd. Want van de oorspronkelijk vijftien functies blijken er negen te ontbreken in het huidige portaal, terwijl deze in de literatuur door meerdere auteurs zijn benoemd als onderdeel van een portaal met open data. Daarnaast zijn ze van belang voor het portaal volgens vijf wetenschappers met een volgorde van prioriteit, van hoog naar laag: F12, F4, F13, F14, F6, F10, F15, F5, F11. Ten slotte is na het analyseren van de huidige gebruikersscenario's met het portaal gebleken dat er ruimte is voor ondersteuning van datagebruikers bij het verwerken van open data met behulp van gebruikerspaden. Om de functies effectief in te zetten voor datagebruikers zal er in het volgende hoofdstuk gebruikerspaden worden uitgewerkt om het gebruik van open data te structureren.

4 Wetenschappelijke bijdrage via gebruikerspaden

Het doel van dit hoofdstuk is om de volgende onderzoeksvraag beantwoorden: *'Met welke vorm van ondersteuning kan het gebruik van open data meer doelgericht worden gestructureerd voor specifieke datagebruikers?*'. Paragraaf 4.1 presenteert daarom gebruikerspaden als oplossing voor het verwerken van open data op een portaal als uitgangspunt. Deze gebruikerspaden zijn een initiële aanzet voor verder onderzoek naar en de ontwikkeling van betere gebruikerspaden die werken in de praktijk. Om verschillende gebruikerspaden te kunnen onderscheiden met behulp van karakteristieken, zijn categorieën voor deze karakteristieken gedefinieerd in 4.2. Vervolgens presenteert 4.3 voorbeeldtoepassingen voor de gebruikerspaden. Ten slotte is in 4.4 de conclusie van dit hoofdstuk gepresenteerd.

4.1 Formuleren van initiële gebruikerspaden

De volgende gebruikerspaden zijn gedefinieerd voor het uitvoeren van analyses op basis van open data. Het doel hierbij is om het nut van gebruikerspaden te bepalen en een eerste aanzet voor verdere ontwikkeling ervan te geven. De gebruikerspaden zijn namelijk voorbeelden die nog verder doorontwikkeld en verbeterd kunnen worden. Eerst is een overzicht gepresenteerd van de gebruikerspaden, hoe ze bepaald zijn en wat de verschillen zijn tussen de gebruikerspaden. De gebruikerspaden zijn als volgt bepaald, inclusief verschillen daartussen.

- Het eerste gebruikerspad gaat ervan uit dat er relaties zijn tussen variabelen kunnen worden benaderd en dit mogelijk maakt om continue variabelen te voorspellen. De sterkere relaties tussen de variabelen kunnen worden gebruikt om de meest invloedrijke factoren in een probleemveld te bepalen. Het hoofdmodel kan worden gebruikt om structuur te leren herkennen uit open data en daarmee voorspellingen te doen bijvoorbeeld.
- Het tweede gebruikerspad is geschikt indien je als eindgebruiker wilt focussen op het doorrekenen van scenario's. Hierbij zijn classificatiemodellen geschikt doordat ze classificaties kunnen bepalen voor alle variabelen en bijbehorende waarden. Indien bij bepaalde scenario's alle waarden vallen binnen de toegestane waardegrens van de classificatie, kan er een label geplakt worden op het scenario en kan met deze informatie beslissingen worden genomen in praktische situaties.
- Het derde gebruikerspad is gefocust op open data voor toepassing van tijdserie-analyses. Dit zijn open data met tijd als onafhankelijke variabele of index voor de afhankelijke variabele. Hiermee kunnen statistische inzichten worden verkregen bij verschillende toestanden in de tijd voor praktische verbetermogelijkheden.
- Het vierder gebruikerspad is gefocust op exploratieve cluster data-analyse van open data waarmee subgroepen met gedeelde kenmerken kunnen worden gevormd. Om bijvoorbeeld inzicht te krijgen in groepen mensen in een stad waarvoor je een nieuwe dienst wilt ontwikkelen.

De gebruikerspaden verschillen op basis van gebruikte modelleertechnieken, naast het perspectief op hoe open data gebruikt kunnen worden gebruikt om beslissingen onderbouwd door te voeren in de praktijk.
Daarnaast zijn er voorwaardelijke condities gevonden waaraan de gebruikerspaden moeten voldoen. De volgende condities zijn van toepassing bij de ontwikkeling van gebruikerspaden (Geels & Schot, 2007):

- Ze zijn niet deterministisch, dus de reeks stappen in een gebruikerspad kunnen worden beïnvloed van buitenaf, bijvoorbeeld door nieuwe informatie.
- De ideale type gebruikerspaden vragen om zorgvuldige toepassing omdat het belangrijk is dat keuzes binnen een gebruikerspad beargumenteerd dienen te worden gebalanceerd ten opzichte van het doel voor het gebruik van een gebruikerspad.

Deze condities voor de ontwikkeling van gebruikerspaden maken duidelijk dat geen ontwikkeling mogelijk is in het geval dat input van eindgebruikers geen rol vervult in gebruikerspaden. Juist de keuzes om open data te koppelingen en de manier waarop een gebruikerspad wordt doorlopen maken de open data waardevol voor eindgebruikers, met creativiteit als unieke toevoeging bij de uitvoering van data-analyses. Anders zou een gebruikerspad een applicatie met functies zijn, wat weinig ruimte overlaat voor het verkrijgen van creatieve inzichten vanuit eindgebruikers met specifieke informatiebehoeftes.

Ook wanneer de gebruikersbehoeftes onbekend zijn en niet kunnen worden vastgesteld is het niet mogelijk om ze te kunnen doorontwikkelen per doelgroep.

Daarom zal het ontwerp gericht zijn op een oplossing die realistisch is en vrijblijvend is voor eindgebruikers die gebruik willen maken van het aanbod van open data. Dit betekent dat de keuzes en uitkomsten vooraf niet vaststaan maar onderdeel blijven van de creativiteit door datagebruikers.

De volgende gebruikerspaden zijn gedefinieerd doordat ze gericht zijn op de gebruikersbehoeftes die onder andere vallen onder de standaard eindgebruikers. De initiële gebruikerspaden zijn als volgt:

1 Ondernemerpad

Stap 1: definieer de onderzoeksvragen of definieer de hypothese die u wilt bevestigen of afwijzen.

Stap 2: verkrijg data van het open dataportaal. Dit kan handmatig worden gedaan door de gegevens (set) te downloaden. Of automatisch, als u wilt dat het een herhaalbaar proces op een URL is.

Stap 3: ontleedt de gegevens voor kwaliteitsbeoordeling, met betrekking tot het doel ervan. Als dit acceptabel is, gaat u verder, anders herhaalt u de vorige stap met andere gegevens op het portaal.

Stap 4: filter relevante subsets van datasets, link naar andere (externe) gegevens, bereiken, categorieën enz.

Stap 5: de gegevens analyseren, te beginnen met beschrijvende statistieken. Totdat dit acceptabel en voldoende informatie betreft, bijvoorbeeld in de vorm van lineaire regressielijnen met de juiste variabelen. Dit is parametrisch als u aannames maakt over de functie van het gegevensmodel dat u representatief voor de dataset wilt maken door middel van parameters (James et al., 2013: 61). Bijvoorbeeld met behulp van support vectormachines om gegevens gecontroleerd te classificeren (Berthold & Hand, 2007: 181). Om de fitheid van een gekozen model te bepalen kunnen verschillende maten worden gebruikt, bijvoorbeeld door middel van de gemiddelde kwadratische fout (James et al., 2013: 29). Dit is om te kwantificeren in hoeverre de voorspelde responswaarde voor een gegeven responswaarde dichtbij de werkelijke respons in de dataset ligt.

Stap 6: presenteer de resultaten in de vorm van visualisaties die in de vorm van een onderzoeksrapport kunnen worden gepresenteerd.

Stap 7: verfijn op basis van visualisaties naar het relevante niveau van gegevens voor inzichten om onderzoeksvragen te beantwoorden. Als dit geen aanvaardbaar antwoord geeft op de onderzoeksvragen en/of hypothesen, herhaal dan de vorige stap.
Stap 8: deel visualisaties met andere gebruikers op het portaal en/of sociale mediakanalen.

2 Wetenschapperpad

Stap 1: onderzoeken, zoeken naar datasets op thema's, filters en gegevensattributen.
Stap 2: hypotheses definiëren op basis van de dataverkenning die moet worden bevestigd of geweigerd.
Stap 3: data-analysemethoden selecteren. Bijvoorbeeld een voorspellend model, in de vorm van een random forest, een techniek die beslissingsstructuren decoreert (James et al., 2013). Voor de classificatie of regressie van gegevens.
Stap 4: gegevens verkrijgen van portaal, handmatig of automatisch herhaalbaar vanaf een URL.
Stap 5: ontleed de gegevens voor kwaliteitsbeoordeling voor het gebruikersdoel of onderzoeksdoel. Indien aanvaardbaar, ga door, anders herhaal voorgaande stap.
Stap 6: filter relevante attributen en koppel eventueel aan andere (externe) open data.
Stap 7: het verzamelen van de datasets, het maken van het model en het testen van de prestaties. Bijvoorbeeld met neuraal netwerk als de modelleertechniek (Berthold & Hand, 2007: 269). Het valideren door middel van K-voudige kruisvalidatie om de beste configuratie te kiezen (Berthold & Hand, 2007: 58). Dus het scheiden van de dataset in meerdere subsets en het testen van de prestaties van modelconfiguraties op elke subset.
Stap 8: visualiseer de resultaten en presenteer de klassen die relevant zijn om de hypothesen te beantwoorden.
Stap 9: verfijn de data-analyseresultaten op basis van inzichten uit de visualisaties.
Stap 10: deel de resultaten met anderen voor feedback.

3 Ontwikkelaarpad

Stap 1: zoeken naar datasets of API/ URL's tussen het aanbod van open data op het portaal
Stap 2: verkrijgen of downloaden van open data met een geautomatiseerd proces of het opslaan van de gegevens in een database. Dit moet een database zijn met een vooraf gedefinieerde structuur voor tijdreeksgegevens.
Stap 3: het gebruik van ongecontroleerde machine-learning, in de vorm van clusteranalyses, om subgroepen met gedeelde kenmerken binnen de gegevens te ontdekken (James et al., 2013: 386).
Stap 4: onderzoeksvraag en/of hypothese definiëren volgens de ontdekte clusters.
Stap 5: vooraf verwerken van de gegevens, koppelen aan andere datasets voor het verrijken van de gegevens, om zo in staat te zijn om relevant onderzoek te doen.
Stap 6: gebruik de resultaten voor het voorspellen van tijdseries. Er zijn meerdere staten en technieken die kunnen worden gebruikt om het gedrag te modelleren. De toestanden zijn als volgt (Ragsdale, 2008: 486):

- Stationair, zonder trends in de tijd
- Niet-stationaire, inclusief opwaartse en neerwaartse trends in de gegevens
- Seizoensgebonden, inclusief patronen in zowel stationaire als niet-stationaire tijdreeksgegevens

Stap 7: ontleed de gegevens voor kwaliteitsbeoordeling om onderzoeksvragen en/ of hypothesen te beantwoorden. Als dit onacceptabel is, herhaalt u de vorige stap.

Stap 8: om een goed voorspellingsmodel voor de gegevens te maken, kunnen de volgende technieken worden gebruikt (Ragsdale, 2008):

- Voortschrijdend gemiddelde: voor het markeren van lange termijntrend over kortetermijn willekeurige fluctuaties.
- Gewogen voortschrijdend gemiddelde: hetzelfde als met voortschrijdend gemiddelde maar dan met variabele gewichten.
- Exponentiële afvlakking: voor exponentieel afnemende gewichten in de loop van de tijd, die zijn toegewezen met exponentiële functies, om het gegevensgedrag in de loop van de tijd representeren.
- Holt-Winters' Methode: voor additieve seizoensgebonden effecten door middel vaneen prognosemodel dat bestaat uit een voorspellingsvergelijking, samen met drie additieve smoothing-vergelijkingen. Eén voor het niveau, voor de trend en een seizoenscomponent.

Stap 9: ontleed de gegevens voor kwaliteitsbeoordeling, indien aanvaardbaar ga naar volgende stap. Anders herhaal de vorige stap.
Stap 10: pas de fitheidtest toe voor het beoordelen van prestaties van modelvoorspelling.
Bijvoorbeeld met een of meerdere van de volgende technieken (Ragsdale, 2008: 487):

- Gemiddelde absolute afwijking: om de variabiliteit van univariate steekproef van kwantitatieve dataset te meten
- Gemiddelde kwadratische fout: om het gemiddelde gekwadrateerde verschil tussen geschatte waarden en waarnemingen in de gegevens te meten
- Standaardafwijking: om de voorspellingsfouten voor verschillende modellen te vergelijken op basis van dezelfde dataset
- Gemiddelde absolute percentuele fout: om de nauwkeurigheid van modellen in termen van percentages te meten

Stap 11: de resultaten inzichtelijk presenteren in de vorm van visualisaties.
Stap 12: de vorige stappen verfijnen, volgens verkregen inzichten uit de visualisaties. Als dit acceptabel is, bevestig dan de onderzoeksresultaten door de onderzoeksvragen te beantwoorden en/of hypothesen te bevestigen of af te wijzen.
Stap 13: deel de onderzoeksresultaten in de vorm van een onderzoeksrapport.

4 Journalistpad

Stap 1: kies een dataset op basis van een interessant thema dat de voorkeur heeft.
Stap 2: verkrijg de dataset handmatig door deze naar de harde schijf van een lokale computer te downloaden.
Stap 3: filter relevante subsets en link naar andere gegevens van externe bronnen. Selecteer waar nodig reeksen van waarden voor verschillende gegevensattributen.
Stap 4: clusteranalyse toepassen op de dataset, voor het ontdekken van groepen met gedeelde kenmerken (James et al., 2013: 386).
Stap 5: definieer de hypothese volgens gedefinieerde clusters.
Stap 6: analyseer de gegevens, mogelijk met statistische gevolgtrekking of beschrijvende statistieken, afhankelijk van de hypothese.
Stap 7: test de prestaties van het model door middel van fitheid (Ragsdale, 2008). Deze meting kan worden gebruikt voor het testen van statistische hypotheses, waarbij bijvoorbeeld twee monsters worden getrokken uit dezelfde verdeling die de dataset vertegenwoordigt.
Stap 8: representeren van de resultaten in vorm van visualisaties voor inzichten.

Stap 9: de analyses volgens afgeleide inzichten uit visualisaties verfijnen tot acceptabel onderzoeksresultaten.
Stap 10: deel inzichten in de vorm van een online beschikbaar onderzoeksrapport.
De vier gebruikerspaden hebben verschillende karakteristieken waarmee ze zich kunnen onderscheiden van elkaar. Dit is de reden dat voorgenoemde gebruikerspaden een aanzet zijn voor verdere ontwikkeling van meer en kwalitatief betere gebruikerspaden voor specifieke datagebruikers. In de volgende paragraaf zijn de hoofdcategorieën van karakteristieken gedefinieerd die aan de hand van de vier gebruikerspaden zijn vastgesteld.

4.2 Onderscheiden van verschillende gebruikerspaden

Aan de hand van de volgende categorieën kunnen gebruikerspaden worden onderscheiden. Gevolgd door een toelichting van alle categorieën:

- Analysetechnieken
- Gegevensvariabelen
- Gegevensmodellering
- Onderzoeksaanpak
- Gegevensverwerving
- Modelprestatiemeting

De categorieën kunnen worden gebruikt voor het ontwikkelen van de zoekfunctie naar verschillende gebruikerspaden:

Analysetechnieken: bij deze eerste categorie wordt er onderscheidt gemaakt tussen beschrijvende en voorspellende modellen. Beschrijvende modellen zijn voor het samenvatten van gegevens en herkennen van groepen met gedeelde kenmerken bijvoorbeeld. De voorspellende modellen zijn er voor het leren van historische gegevens door middel van patroonherkenning voor toekomstige beslissingen.

Gegevensvariabelen: een specifieke dataset kan gestructureerde gegevens, in de vorm van gesorteerde kwantitatieve waarden bevatten. Of ongestructureerde kwalitatieve gegevens in de vorm van een rapport bijvoorbeeld.

Gegevensmodellering: in het geval van parametrische modellen wordt een functie gedefinieerd en worden de waarden benaderd op basis van de veronderstelde functie. In het geval van niet-parametrische modellen, worden de gegevens door middel van meer complexe modellering geanalyseerd. Om onderzoeksvragen te beantwoorden of om statistische gevolgtrekkingen te gebruiken voor het testen van hypotheses zal er moeten worden geleerd van modellen. Het gecontroleerd leren met parametrische aannames, is nuttig voor het testen van hypothesen met kwantitatieve benadering. Ongecontroleerd leren met niet-parametrische modellen kan worden gebruikt bij het beantwoorden van onderzoeksvragen die gaan over het ontdekken van structuren op basis van de gegevens. Bijvoorbeeld groepen van observaties en/of klassen die aan de gegevens kunnen worden toegewezen op basis van perspectieven die dieper liggen dan de oorspronkelijk samengestelde gegevens door koppeling met andere datasets.

Onderzoeksaanpak: bij deze categorie is het mogelijk om verschillende vertrekpunten te kiezen voor onderzoek op basis van open data. Dat kan met onderzoeksvragen of hypothesen. Men dient de gegevens te verzamelen nadat hypothesen en/of onderzoeksvragen zijn gedefinieerd. Een combinatie van onderzoeksvraag en hypothese is ook mogelijk. Een voorbeeld om open data te analyseren is om een functie op een dataset aan te nemen en de hypothesen te bevestigen of te verwerpen door onafhankelijke variabelen

te vergelijken met een gekozen afhankelijke variabele. Hierbij worden de gegevens gebruikt als basis voor toepassing van analysetechnieken. Het is ook een optie om te beginnen met onderzoeksvragen en vervolgens te zoeken naar gegevens waarbij men de variabelen aanpast met behulp van modificatieregels die worden toegepast op basis van aannames uit beschrijvende statistische analyses bijvoorbeeld. Men moet dus aannames doen om leren mogelijk te maken. Om patronen in een branche of interessegebied ontdekken die niet noodzakelijk heel specifiek hoeven te zijn, zoals in het geval bij het testen van een hypothese nodig is. Maar men wilt wel structuur uit de open data leren voor de toekomst.

Gegevensverwerving: het is mogelijk om afzonderlijke datasets handmatig te downloaden of gebruik te maken van software om er een geautomatiseerd en hierdoor herhaalbaar proces van te maken. Toegang tot een continue datastroom is interessant voor het voorspellen van trendreeksen in de tijdreeksen, voor het leren van patronen in de loop van de tijd. De eerste manuele optie voor het downloaden is discreet. Terwijl de tweede automatische optie continu is.

Modelprestatiemeting: bij deze categorie wordt een onderscheidt gemaakt tussen modelprestatie-optimalisatie en het kiezen van het niveau van modelcomplexiteit. Bij de laatstgenoemde gaat het om de fractie van correcte voorspellingen van een functie, uitgedrukt in een percentage, op basis van gegeven dataset. Maar voor modelprestatie-optimalisatie is het mogelijk om verschillende configuraties testen van een model waarin een dataset opsplitst is in verschillende subsets voor verschillende configuraties. Deze resulteren in modelprestaties en worden vergeleken om vervolgens de beste configuratie te kiezen op basis van het grootste aantal correcte voorspellingen. Is het doel om een groot aantal correcte voorspellingen te doen over een dataset of juist het leren van trends? Als dit laatste het geval is, wordt het optimaliseren van de modelprestaties op veronderstelde aannames beschouwd als schijnnauwkeurigheid. Omdat het geen ruimte laat voor interpretatie van asymmetrische waarnemingen in de dataset aangezien het in de functie zou kunnen worden vastgelegd. Dus voordat modelprestaties worden geoptimaliseerd, is het de vraag hoe nauwkeurig men dat wilt doen en welke sleutelindicatoren kunnen worden gebruikt, naast bijvoorbeeld 'een fractie van de juiste voorspellingen'. Andere configuraties zijn misschien interessanter voor onderzoeksvragen, afhankelijk van wat en hoe men van beschikbare open data wilt leren. Dit is ook het geval bij het testen van hypotheses. Als er een verschil is in uitkomsten bij dezelfde hypotheses op verschillende subsets als gevolg van modelconfiguratieverschillen, is dit ongewenst. Dit komt omdat het model zo eenvoudig mogelijk moet worden gehouden bij het verwerpen of bevestigen van vooraf gedefinieerde hypothesen. In dat geval moet het niveau van modelcomplexiteit zo eenvoudig mogelijk zijn zonder voorspelbaarheid te verliezen.

Om de verschillen tussen gebruikerspaden te demonstreren zijn de volgende paragraaf voorbeeldtoepassingen van de gebruikerspaden uitgewerkt.

4.3 Voorbeeldtoepassingen van gebruikerspaden

Er zijn vier gebruikerspaden geproduceerd, ieder in de vorm van een reeks te volgen stappen. Dit is een initiële presentatie van de toepassing van de gebruikerspaden die nog doorontwikkeld moeten worden door ze specifieker te maken op basis van de wensen van verschillende soorten datagebruikers. Ze kunnen verder worden verfijnd conform de behoefte naar gebruikerspaden voor de beantwoording van onderzoeksvragen en/of hypothesen. De voorbeeldtoepassingen voor verschillende doelgroepen zijn als volgt:

De doelgroep voor het eerste gebruikerspad zijn ondernemers:

Dit eerste gebruikerspad bestaat uit een reeks te volgen stappen waarmee de voorspellende onderzoeksvraag en/of hypothese kan worden beantwoord. Een voorbeeld onderzoeksvraag is: *'Met hoeveel zal het inwonersaantal in 2020 toegenomen zijn en welke diversiteit zal deze verwachtingsgewijs opleveren in Den Haag?'*. De open data die van toepassing zijn, zijn in de vorm van een Pdf-document beschikbaar gesteld als een onderzoeksrapport. In dit rapport staat de ontwikkeling van het inwonersaantal in de vorm van een prognose tussen 2012-2020. Er is geen maatsoftware nodig om de dataset te kunnen downloaden. De analyses zijn namelijk al uitgevoerd en in het document zelf gepresenteerd. Vervolgens zal de kwaliteit van de open data worden beoordeeld door de datagebruiker, in dit geval blijkt de kwaliteit hoog te zijn omdat het een officieel uitgebracht beleidsrapport is. De ondernemer is geïnteresseerd in een bepaalde leeftijdscategorie en zal daarom een subset tussen 25-30 jaar nemen. Daarom zullen de resultaten worden gesorteerd op alle 25-30-jarigen per afkomst. Het staafdiagram visualiseert de prognose, waarmee de onderzoeksvraag kan worden beantwoord en besluit de ondernemer een supermarkt te openen met producten die passen bij de culturele diversiteit in de desbetreffende regio.
Karakteristieken voor het eerste gebruikerspad zijn: Pdf-document, Manuele Gegevensverwerving, Onderzoeksvraag, Analyse.

Het tweede gebruikerspad met als doelgroep de wetenschapper:
Een wetenschapper wil onderzoek doen naar de kredietcrisis die in 2008 tot een hoogtepunt kwam en economische consequenties had voor de wereld waaronder Nederland. De wetenschapper wil de internationale handel van- en naar Nederland onderzoeken in de periode voor en na 2008. De wetenschapper zoekt en vindt een dataset van in-, uit- en doorvoer van handel, aan de hand van de Standard International Trade Classification (SITC) 2002-2015 in de JavaScript Object Notation (JSON) ATOM-feed dataformaat (CBS, 2018). De wetenschapper wil de open data in een MySQL database importeren waarna de dataset kan worden geëxploreerd op basis van query's waarmee subsets kunnen worden geselecteerd uit de gevulde database. De hypothese die is opgesteld naar aanleiding van de gegevens is als volgt: *'Na de kredietcrisis is de invoer van bewerkt hout en constructiewerken afgenomen in Nederland, en daarnaast is ook de uitvoer van vlees van runderen afgenomen'.* De wetenschapper kan er ook voor kiezen om eerst te analyseren welke goederen significant zijn veranderd na de kredietcrisis als onderzoeksvraag. Vervolgens gebruikt de wetenschapper, randomforest classificatietechniek om met behulp van meerdere algoritmen, de variantie uit de middelen en daarmee tot een passende classificatie van in-, uit- en doorvoercijfers te komen. De gegevens worden met behulp van maatsoftware geïntegreerd in een database op een externe server. De kwaliteit van de dataset is beoordeeld conform de score van het bijbehorende criterium. Vervolgens zijn specifieke productcategorieën geselecteerd en op een neuraal netwerk als invoervariabelen gebruikt met de tijd als uitvoervariabele. Hiermee kunnen relaties tussen goederen die zijn beïnvloed door de kredietcrisis worden bepaald.
Karakteristieken van het tweede gebruikerspad zijn: webservice, onderzoeksvraag, random forest, verificatie, configuratieoptimalisatie.

Het derde gebruikerspad heeft de ontwikkelaar als doelgroep:
Een ontwikkelaar wil een dienst ontwikkelen met behulp van open data over voertuigen in Nederland. Het type auto's wil de ontwikkelaar koppelen aan de levering van auto-onderdelen binnen een dag na de bestelling. Daarvoor is het vereist dat de websitebezoeker op kenteken naar onderdelen kan zoeken in de vorm van een webservice, die maatsoftware vereist en de website waarop de content wordt gepresenteerd. De volgende onderzoeksvraag wordt gesteld

door de ondernemer: *'Van welke soort onderdelen is de vraag hoger in bepaalde periodes?'*. Om deze onderzoeksvraag te kunnen beantwoorden worden tijdsreeksanalyses uitgevoerd om fluctuaties in de vraag naar bepaalde soort onderdelen te observeren in een vroegtijdig stadium. De fluctuaties en trends zijn weergegeven in de vorm van een lijndiagram. De informatie kan eventueel worden gecombineerd met weerdata om de invloed van het weer te meten op de besteding aan auto-onderdelen bijvoorbeeld. Maar daarvoor moet maatsoftware worden ontwikkeld om de datasets in te laden in een database, om erdoorheen te navigeren, zodat de relevante datasets kunnen worden gekoppeld. Het model erachter is geverifieerd door middel door een formule waarmee de verschillen tussen voorspellingswaarden uit het model en de daadwerkelijke observaties worden gekwadrateerd en opgeteld. Waarna het wordt gedeeld door het aantal observaties in de gebruikte dataset. Met het kwadraat worden alle waarden positief gemaakt ten opzichte van de lijn die het model representeert. Hiermee kan de prestatie van het model worden uitgedrukt in een gemiddelde gekwadrateerde fout of de Mean Squared Error (MSE).
Karakteristieken van het derde gebruikerspad zijn: webservice, onderzoeksvraag, tijdreeksanalyse, verificatie, gemiddelde kwadratische afwijking.

Doelgroep voor het vierde gebruikerspad is de journalist:
De journalist wil onderzoek doen naar trends op het gebied van fusies en overnames. Deze hebben economie als thema. Een exploratieve cluster data-analyse is uitgevoerd om groepen met gedeelde kenmerken te ontdekken. Bijvoorbeeld op basis van de attributen bedrijfstak en bedrijfsgrootte. Naar aanleiding van de gepresenteerde clusters kan een hypothese worden gedefinieerd: *'Het verschil in bedrijfsgrootte is bepalend of twee bedrijven fuseren of dat het grote bedrijf de kleinere overneemt'*. Om deze hypothese te testen, is een lineair regressiemodel gemaakt waarbij de relaties tussen de indicatoren worden geanalyseerd om de hypothese te bevestigen of juist te ontkrachten naar aanleiding van de analyseresultaten. Om dit uit te rekenen is een specifieke strategie nodig om tot de gewenste informatie te kunnen komen. Bijvoorbeeld door eerst het gemiddelde van de bedrijfsgroottes uit te rekenen, om vervolgens twee subsets te maken met bedrijfsgrootte als index, oftewel de onafhankelijke variabele. De eerste subset van de waarden liggen boven het gemiddelde van bedrijfsgrootte, en de bijbehorende observaties voor het aantal fusies. De tweede subset van de waarden zijn gelijk aan of onder het gemiddelde van bedrijfsgrootte, inclusief bijbehorende observaties van het aantal fusies. Van beide subsets wordt de gemiddelde absolute afwijking uitgerekend. Indien deze gemiddelde absolute afwijking groter is in de tweede subset met bovengemiddelde waarden van bedrijfsgrootte dan de andere subset, kan worden beredeneerd dat een toenemende grootte van een bedrijf toenemende invloed heeft op fusie als optie. Dit kan ook worden uitgerekend met het aantal overnames als afhankelijke variabele en bedrijfsgrootte als index, of de onafhankelijke variabele. De gemiddelde absolute afwijking heet ook Mean Absolute Deviation (MAD).
Karakteristieken van het vierde gebruikerspad zijn: webservice, cluster data-analyse, hypothese, lineaire regressiemodel, verificatie, gemiddelde absolute afwijking.

4.4 Conclusie

Er zijn vier gebruikerspaden gedefinieerd die structuren representeren voor de efficiënte verwerking van open data. Dit verschilt per datagebruiker en daarom kunnen gebruikerspaden in de toekomst kwalitatief beter gemaakt worden door meer onderzoek te doen naar de gebruikersbehoeften. Tevens zijn er categorieën gedefinieerd waarmee onderscheid kan worden gemaakt tussen specifieke gebruikerspaden. Met deze informatie kan een effectieve

en gebruiksvriendelijke zoekfunctie naar gebruikerspaden mee worden ontwikkeld. Andere kunnen leren hoe ze een data-analyseproces kunnen aanpakken aan de hand van iteratieve en stapsgewijze taken. Deze gebruikerspaden zijn een aanzet voor verdere ontwikkeling van dit concept. Het gaat hierbij niet om de compleetheid en volledigheid, maar hoe gebruikerswensen bij open data het best kunnen worden ondersteund op een structurele manier.

5 Evaluatie van gebruikerspaden

De gebruikerspaden die in het vorige hoofdstuk zijn opgesteld, zullen in dit hoofdstuk worden geëvalueerd en verbeterd naar aanleiding van de feedback door vijf deskundigen. Doel van dit hoofdstuk is het beantwoorden van de volgende onderzoeksvraag: *'Klopt de volgorde en het aantal stappen als onderdeel van ieder gebruikerspad ter ondersteuning van datagebruikers?'*. In 5.1 zijn de gebruikerspaden geëvalueerd op belang om te meten of ze, op een schaal van 1 tot 5, een realistische en toepasbare oplossing zijn vanuit de respondenten bezien. Vervolgens is aan de deskundigen gevraagd of zij ontbrekende stappen herkennen in de gebruikerspaden, en of de volgorde van de stappen in de gebruikerspaden correct is. De uitkomsten hiervan zijn beschreven in 5.2. Na beantwoording van de twee hiervoor genoemde vragen zijn de suggesties voor aanpassing van de gebruikerspaden gepresenteerd in 5.3. De voor- en nadelen voor gebruikerspaden als toepassing staan beschreven in 5.4. Tenslotte presenteert 5.5 de conclusie van dit hoofdstuk.

5.1 Gebruikerspaden evalueren

In deze paragraaf zijn de gebruikerspaden geëvalueerd op belang door te vragen hoe belangrijk ze worden gevonden op een schaal van 1 tot 5 met 1=onbelangrijk en 5=erg belangrijk. Dit is gevraagd aan vijf PhD-studenten die deskundig zijn op het gebied van open data binnen verschillende toepassingsgebieden. Waaronder het verkrijgen van logistieke inzichten in verkeersstromen, het ondersteunen bij de analyse van big datastromen, het verfijnen van geconstrueerde voorspellingsmodellen, en genereren van voordelen met open data. Deze deskundigen doen onderzoek aan de Technische Universiteit Delft op het gebied van open data met een focus op open overheidsgegevens. De vraag aan deze respondenten is tot stand gekomen doordat hun visie op de gebruikerspaden, als een bevestiging kan worden beschouwd voor de goede richting naar toekomstig onderzoek. Met belang bedoelen we bruikbaarheid van gebruikerspaden in de praktijk voor datagebruikers. De vijf bevraagde respondenten zijn hierbij representatief om het belang van gebruikerspaden te onderstrepen, doordat ze werken met open data in verschillende richtingen, terwijl het hoofdthema van dit onderzoek het gebruik van open data betreft.

In tabel 10 staat hoe belangrijk ieder gebruikerspad is vanuit de deskundigen bezien, op een schaal van 1 tot 5. Voor elk gebruikerspad zijn vijf deskundigen gevraagd een score toe te kennen. Van deze scores is het gemiddelde van elk gebruikerspad weergegeven in de meest rechtse kolom van tabel 10.

Tabel 10: Belang van gebruikerspaden onder de deskundigen, Schaal 1= niet belangrijk en 5= erg belangrijk

Gebruikers pad	Deskundige 1	Deskundige 2	Deskundige 3	Deskundige 4	Deskundige 5	Gemiddelde
Een	3	4	5	5	5	4,4
Twee	4	5	5	5	5	4,8
Drie	4	5	5	5	4	4,6
Vier	4	5	5	5	4	4,6

Aan de hand van de weergegeven resultaten in tabel 10 is te zien dat alle gebruikerspaden een hoge score hebben die gemiddeld hoger ligt dan 4, bezien vanuit alle respondenten. Volgens de resultaten uit tabel 10 kan daarom geconcludeerd worden dat de gebruikerspaden potentieel waardevol zijn voor open data-analyses. De volgende paragraaf presenteert antwoorden op vragen over mogelijke tekortkomingen van de gebruikerspaden.

5.2 Verbetersuggesties voor gebruikerspaden

De vier gebruikerspaden zijn in eerste instantie opgesteld met het idee dat, om open data te analyseren, je stapsgewijs te werk moet gaan om overzicht te houden op taken die moeten worden voldaan voor het uiteindelijke gebruiksdoel. Daarmee kunnen de gebruikerspaden een leidende rol vervullen voor toekomstige datagebruikers. Het gebruik van de gebruikerspaden vereist wel dat bestaande technologie, in de vorm van functies, beschikbaar moeten zijn om de taken binnen een realistisch tijdsbestek te kunnen voltooien. Daarom wordt rekening gehouden met mogelijke functies voor data-analyses en criteria over de gebruiksvriendelijkheid van open data. Om de gebruikerspaden te verbeteren zijn vijf respondenten geïnterviewd en gevraagd of ze ontbrekende stappen herkennen in 5.2.1. Vervolgens is nagegaan of de stappen in de gebruikerspaden een onjuiste volgorde hebben in 5.2.2.

5.2.1 Ontbrekende stappen in gebruikerspaden

De vijf respondenten zijn gevraagd of ze ontbrekende stappen herkennen in de gedefinieerde gebruikerspaden. De antwoorden zijn gepresenteerd in tabel 11.

Tabel 111: Ontbrekende stappen in de gebruikerspaden, Legenda: Y=yes (0 punt) en N=no (1 punt)

Gebruikers pad	**Deskundige 1**	**Deskundige 2**	**Deskundige 3**	**Deskundige 4**	**Deskundige 5**	**Score**
Een	Y	Y	Y	Y	Y	0
Twee	Y	N	Y	Y	Y	1
Drie	Y	Y	N	N	Y	2
Vier	Y	N	Y	N	Y	2

Bij de optellingen in de meest rechtse kolom zijn het aantal ontkennende antwoorden van de respondenten, oftewel dat er geen stappen ontbreken in de gebruikerspaden, bepalend voor de totaalscore. Deze scores zijn voor elk onafhankelijk gebruikerspad gemeten en staan in de meest rechtse kolom van tabel 11.

Uit de tabel kan er het volgende worden opgemaakt. Alle respondenten zien ontbrekende stappen in het eerste gebruikerspad. Dit deskundigen vinden dus dat het eerste gebruikerspad nog onvolledig is voor het gebruik ervan bij open data-analyses. Bovendien zijn de andere drie deskundigen minder kritisch in termen van ontbrekende stappen, maar hebben ze wel aangegeven. Daardoor kan er worden geconcludeerd dat er ruimte is voor verbetering van alle gebruikerspaden.

5.2.2 Onjuiste ordening van stappen in gebruikerspaden

Aan de respondenten is vervolgens gevraagd of de stappen die de verschillende gebruikerspaden vormen naar hun mening de juiste volgorde hebben. Hun antwoorden en totaalscores staan in tabel 12.

Tabel 122: Juiste ordening voor stappen in gebruikerspad. Legenda: Y=yes (1 punt) en N=No (0 punt)

Gebruiker spad	Deskundige 1	Deskundige 2	Deskundige 3	Deskundige 4	Deskundige 5	Score
Een	Y	Y	Y	Y	Y	5
Twee	N	N	Y	Y	Y	3
Drie	N	Y	Y	Y	Y	4
Vier	N	N	Y	Y	Y	3

Voor elk onafhankelijk gebruikerspad zijn vijf respondenten gevraagd of er sprake is van de juiste volgorde van stappen. De scores bestaan uit het aantal positieve antwoorden voor elk gebruikerspad en staan in de meest rechtse kolom van tabel 12 gepresenteerd. Wat we kunnen concluderen uit tabel 12 is dat het eerste gebruikerspad het beste scoort onder de vijf respondenten. Over het algemeen zijn de deskundigen dus van mening dat de ordening van stappen klopt. Maar er zijn verschillende opvattingen over het ordenen van stappen met betrekking tot gebruikerspad 2 tot en met 4. Daarom kan de ordening van stappen in deze gebruikerspaden worden verbeterd. In de volgende paragraaf wordt de feedback van deskundigen verwerkt voor implementatie in herziene versies van de verschillende gebruikerspaden. Elk van de geïnterviewde respondenten heeft feedback gegeven bij het ontbreken van stappen en/of niet-kloppende volgorde daarvan. De gebruiksdoelen bij open data en feedback op de gebruikerspaden zijn per respondent beschreven in bijlage 3. De suggesties voor verbetering van de gebruikerspaden zijn per gebruikerspad behandeld in de volgende paragraaf.

5.3 Feedback verwerken voor verbeterde gebruikerspaden

De deskundigen hebben feedback gegeven op basis van twee enquêtevragen die in 5.2 zijn beantwoord. In bijlage 3 zijn achtereenvolgens de gebruiksdoelen en suggesties gepresenteerd voor verbetering van de gebruikerspaden. De nieuwe versies van de gebruikerspaden, na het verwerken van de feedback door vijf deskundigen, staan in bijlage 4. Deze verbeterde gebruikerspaden zijn het resultaat van een aantal additionele stappen welke zijn aangegeven met ‘extra stap’ of ‘extra uitleg’ bij de elk gebruikerspad, naast aanpassingen dat de volgorde betreft van stappen in elk gebruikerspad. In deze paragraaf is beschreven welke mogelijke aanpassingen zijn vastgesteld per gebruikerspad.

Feedback voor aanpassingen aan het ondernemerpad

Er ontbreekt een terugkoppeling naar de dataproviders. Dit is belangrijk om de datakwaliteit onder de loep te nemen. Ook om in staat zijn om het proces van datakwaliteit in een vroeg

analysestadium te delen met andere gebruikers. Daarom is het belangrijk om feedback te geven aan dataproviders tussen stap 6 en stap 8.

Er ontbreekt een stap: ten slotte moet de onderzoeksvraag en/of hypothese worden beantwoord. De stappen: 3, 4, en 5 zijn afhankelijk van welke open data worden verkregen. Ook het gebruik van grote datasets moet meestal worden gescheiden in subsets om verspilling van tijd bij het analyseren van de gegevens te voorkomen. Als er voor een onderwerp slechts kleine gegevensreeksen beschikbaar zijn, moeten de gegevens vaker worden bijgewerkt.
Verder mist er bij het eerste gebruikerspad, de presentatie van de resultaten in de vorm van een peer-reviewed publicatie. Verder moeten afzonderlijke stappen van het gebruikerspad meer specifiek op het domein zijn van het onderzoek waarvoor het gebruikerspad is gebruikt binnen individueel onderzoek. Er zijn verschillende regio's, en domeinen voor onderzoek, waarbij het opzetten van experimenteeronderzoek hoort. Daarmee zouden de gebruikerspaden specifieker kunnen worden gemaakt.

Extra stappen naar aanleiding van de feedback op het ondernemerpad

- Extra stap A: de verkregen gegevens hebben invloed op de manier waarop de volgende stappen: 3, 4 en 5 worden uitgevoerd en/of gedefinieerd. Als er een grote dataset is, moet deze worden gescheiden in subsets, anders worden gegevens inefficiënt geanalyseerd, wat een verspilling van tijd is. Als er slechts een kleine dataset voorhanden is, kan deze na feedback aan de dataprovider worden bijgewerkt. Of gelinkt met andere mogelijk relevante datasets naar aanleiding van constructieve feedback.
- Extra stap B: geef feedback aan de dataprovider voor het onderzoeken van de datakwaliteit. Indien de kwaliteit van dataset niet naar wens is, kan dit kenbaar worden gemaakt in de vorm van constructieve feedback, ter verbetering ervan.
- Extra stap C: problemen met datakwaliteit delen met anderen vergelijkbare datagebruikers om interactie over de datakwaliteit mogelijk te maken met constructieve feedback tot gevolg. De datakwaliteit is een probleem wanneer er foute conclusies worden getrokken uit foutieve resultaten.
- Extra stap D: geef feedback aan dataproviders voor het onderzoeken van de datakwaliteit. Bij het uitvoeren van de analyses kan aan de dataprovider fouten kenbaar worden gemaakt, in de vorm van constructieve feedback ter verbetering van datakwaliteit.
- Extra stap E: beantwoord de onderzoeksvragen en hypothesen volgens afgeleide resultaten en publiceer de resultaten in een peer-reviewed journaal. Met peer-reviewed bedoelen we dat verschillende deskundigen binnen het vakgebied waar de data-analyse onder valt, een kritische blik werpen op het geschreven werk in de vorm van feedback.

Feedback voor aanpassingen aan het wetenschapperpad

Wat het tweede gebruikerspad betreft, eerst definiëren van de onderzoeksvraag om vervolgens de relevante gegevens te verzamelen. Ook ontbreekt de terugkoppeling over de datakwaliteit naar de dataproviders.

Tussen stap 8 en stap 10, is er behoefte aan een terugkoppeling naar de dataprovider. Ook zou er een extra stap kunnen worden toegevoegd: uiteenzetten van doelstellingen. Het onderzoeksdoel zou aan het begin, bijvoorbeeld in de vorm van een

hypothese moeten worden aangeduid. Dit hangt ook af van de klant voor wie de open data zou moeten worden geanalyseerd bijvoorbeeld.

Extra stappen naar aanleiding van de feedback op het wetenschapperpad

- Extra stap A: begin met het definiëren van onderzoeksvragen en zet de onderzoeksdoelen uiteen.
- Extra stap B: geef feedback aan de dataproviders als er iets ontbreekt of fout is met de open data, of doe een datakwaliteitsbeoordeling. Er is hier geen standaard voor aangezien dit afhankelijk is van de data zelf en dient met een kritische blik te worden geëvalueerd. Zijn de metingen geschikt voor het gebruiksdoel is hierbij de hoofdvraag die verder kan worden uiteengezet in subvragen voor verschillende perspectieven op de geschiktheid van de dataset voor specifieke gebruiksdoelen.
- Extra stap C: uitleggen wat voor soort informatie is gewenst, en voor welke verwerkingstaken, om te begrijpen welke aanvullende gegevens nodig zijn.
- Extra uitleg voor stap 5: de criteria voor datakwaliteit moeten worden gedefinieerd om te beslissen of de gegevens nuttig zijn voor het beantwoorden van de onderzoeksvragen en/of het bevestigen of ontkrachten van de vooraf gedefinieerde hypothesen.
- Extra stap D: geef feedback aan dataproviders over de kwaliteit van gebruikte gegevens.

Feedback voor aanpassingen aan het ontwikkelaarpad

Wat het derde gebruikerspad betreft, zou het moeten beginnen met de onderzoeksvraag en vervolgens de verzameling van gegevens.

Verder is vóór stap 4, tot stap 3 behoefte aan resultaten. Daarnaast is er behoefte aan een extra stap tussen stap 3 en stap 4, die gedetailleerder moet worden beschreven met behulp van aanvullende informatie. Ook is er behoefte aan meer uitleg over stap 5 en stap 7.

Het opslaan van gegevens is niet verplicht, aangezien aggregatie van gegevens en daarmee meerdere bronnen worden gebruikt voor de data-analyse. De kwaliteit van de open data moet worden beoordeeld alvorens data-analysetechnieken toe te passen in de vorm van modellen die worden gefit op datasets. Daarnaast zijn evaluatietechnieken voor de modellen afhankelijk van de soort data-analysetechnieken die worden gebruikt. Zo is voor een clusteranalyse de gemiddelde gekwadrateerde fout niet geschikt terwijl voor regressiemodellen dit juist wel geschikt is om de modelprestaties te evalueren. De datakwaliteit moet worden beoordeeld voordat de data-analysemodellen kunnen worden gebouwd in het algemeen. Hiermee kan worden voorkomen dat het gemiddelde te sterk wordt beïnvloed door uitschieters die het gemiddelde meetrekken zonder valide reden, maar door slechte datakwaliteit. Alle data-analysemethoden zijn eveneens afhankelijk van de doelstellingen die aan het begin van de gebruikerspaden zijn geformuleerd.

Extra stappen naar aanleiding van de feedback op het ontwikkelaarpad

- Extra stap A: begin met het definiëren van de onderzoeksvragen.
- Extra stap B: interpreteer de resultaten in het licht van verder relevant onderzoek.
- Extra stap C: vertaal de onderzoeksvragen en hypothesen in een gegevensmodel dat alle vragen kan beantwoorden. Dit model representeert de waarden van de dataset die gebruikt is voor produceren van het model. Bijvoorbeeld om voorspellingen te doen

over toekomstige ontwikkelingen door hypotheses te toetsen met behulp van geproduceerde voorspellingsmodellen.

- Extra stap D: definieer de criteria die u wilt gebruiken voor de beoordeling van de gegevenskwaliteit, dit hangt af van wat voor soort onderzoeksvragen u wilt beantwoorden en de hypothesen die u wilt bevestigen of juist ontkrachten. Ga vervolgens door naar de kwaliteitsbeoordeling om deze te verbeteren door feedback te geven aan dataproviders.
- Extra stap E: kies het juiste model evaluatiemethode die past bij de modelleer techniek.

Feedback voor aanpassingen aan het journalistpad

Wat het vierde gebruikerspad betreft, moet de kwaliteitsbeoordeling de eerste stap zijn. Ook de verstrekking van feedback aan dataproviders en de datakwaliteitsbeoordeling delen met andere datagebruikers mist. Dit is nodig om de datakwaliteit te verbeteren of om de koppeling met andere datasets te herkennen. Bijvoorbeeld als een identificatiekolom ontbreekt die wel onderdeel was van een dataset die was geüpload.

Ook is er behoefte aan een terugkoppeling naar de dataproviders, omdat dit van groot belang is om de kwaliteit van open data te verbeteren.

Stap 5 kan het beste naar het begin van de gebruikerspad worden gehaald. Eerst begint men namelijk met het formuleren van een hypothese of onderzoeksvraag, en daarna kunnen data-analysetechnieken worden gekozen voor de beantwoording. Omdat alle data-analysemethoden afhankelijk zijn van de onderzoeksdoelen die aan het begin van een gebruikerspad zijn geformuleerd.

Extra stappen naar aanleiding van de feedback op het journalistpad

- Extra stap A: doe een kwaliteitsbeoordeling van mogelijke gegevens.
- Extra stap B: geef feedback aan gegevensproviders over de gekozen gegevens.
- Extra stap C: deel data kwaliteitsbeoordeling met andere gebruikers, voor feedback om het te combineren met andere gegevens om die gegevenskwaliteit te verbeteren. Hierbij is het belangrijk dat koppelingen tussen gegevens waar mogelijk worden herkend en toegevoegd. De identificatienamen van de kolommen zijn handige informatie bij het maken van keuzes voor de koppeling van verschillende datasets.

De verbetersuggesties hebben tot inzichten die geleid die niet alleen de gebruikerspaden hebben verbeterd (zie bijlage 4). Maar ook wel voor- en nadelen kunnen worden geassocieerd met het gebruik van open data door standaard eindgebruikers. In de volgende paragraaf zijn deze gepresenteerd.

5.4 Voor- en nadelen van gebruikerspaden

Door het evalueren van de gebruikerspaden zijn een aantal voor- en nadelen vastgesteld met betrekking tot de inzet van gebruikerspaden. De volgende voordelen en nadelen kunnen daarom worden beschouwd als argument voor verdere ontwikkeling van gebruikerspaden:

Voordeel 1: met gebruikspaden kan er efficiënter aan de slag worden gegaan met het gebruik van open data, dit bespaart tijd en kosten voor datagebruikers.

Voordeel 2: de potentie van open data is niet meer onbekend aangezien er structuur is gegeven voor het gebruik ervan. Het kiezen van bewezen technieken en ontwikkelen van vaardigheden staan hierdoor open voor de ontwikkeling van collectieve intelligentie.

Nadeel 1: de creativiteit van datagebruikers is minder doordat een voorgeschreven structuur de datagebruikers minder kritisch laat nadenken over een aanpak voor het analyseren van open data. Dit laatste hangt ook samen met het gebruik maken van Informatie- en communicatietechnologie van andere disciplines waarbij analyse van gegevens een belangrijke rol speelt, zoals bij softwareontwikkeling.

Nadeel 2: het is niet bekend wat de waarde van een gebruikerspad is alvorens er gebruik van te maken. Want het analyseren van gegevens verschilt per achterliggend gebruikersdoel en daarmee is een gebruikerspad geen garantie voor succes. Bij onbedoeld of onjuist gebruik kunnen resultaten tegenvallen. Hierbij is aansturing vermoedelijk nodig om het juiste gebruik ervan te verzekeren.

5.5 Conclusie

Gebruikerspaden maken het gemakkelijker om een overzicht te hebben over de taken die dienen te worden uitgevoerd voor verschillende data-analyses. Deze gebruikerspaden zijn een eerste aanzet om effectief aan de slag kunnen gaan met open data, te bevorderen. Er zijn meer gebruikerspaden mogelijk die kunnen leiden tot gewenste resultaten voor specifieke doelgroepen. Het definiëren van gebruikerspaden is oorspronkelijk onderdeel van het creatieve proces om te komen tot het gebruiksdoel met creatieve keuzes vanuit de datagebruikers zelf. Met de vier gebruikerspaden kunnen data-analyseresultaten worden behaald op een creatieve maar gestructureerde manier. De gebruikerspaden zijn verbeterd met de feedback van vijf deskundige open datagebruikers die verschillende visies hebben voor het gebruik van open data. De aanzet zit in de goede richting voor toekomstige ontwikkeling en inzet in de praktijk. Ze zijn niet af doordat nog niet is onderzocht of ze gebruikt worden in de praktijk. Daarvoor moeten ze eerst worden aangeboden op een portaal via een aantrekkelijke zoekfunctie. Dit geeft datagebruikers de mogelijkheid om gebruikerspaden te kiezen die door anderen zijn gebruikt bij de uitvoering van data-analyses. De gebruikerspaden structureren het gebruik van open data via beschikbare functies om met behulp van analyseresultaten uiteindelijk de onderzoeksvraag en/of hypothese te kunnen beantwoorden. Dit onderzoek is de aanzet voor de ontwikkeling van meer en uitgebreidere gebruikerspaden. De gebruikerspaden in dit onderzoek kunnen als basis dienen voor verder onderzoek.

6 Conclusies en aanbevelingen

In dit hoofdstuk worden de conclusies aanbevelingen gegeven naar aanleiding van de hoofdvraag uit het eerste hoofdstuk: ***'Wat moeten portaaleigenaren doen om de drempel voor het gebruik van open data te verlagen?'***

In paragraaf 6.1 is er antwoord gegeven op de hoofdvraag door middel van specifieke maatregelen die portaaleigenaren moeten nemen om de drempel te verlagen voor het gebruik van open data. Vervolgens is in 6.2 beargumenteerd met welke methoden de subonderzoeksvragen zijn beantwoord en wat de methoden specifiek voor het onderzoek hebben opgeleverd. Daaropvolgend is in 6.3 zowel de wetenschappelijke als praktische contributie van dit onderzoek uiteengezet. Hierbij wordt beargumenteerd dat de ontwikkeling van gebruikerspaden complex is, doordat vele iteraties zijn vereist naast het begrijpen van de gebruikersbehoefte. Wat betreft de praktische contributie wordt toegelicht welke veranderingen er aan het portaal nodig zijn. 6.4 zet de beperkingen van de onderzoeksopzet uiteen. Aansluitend presenteert 6.5 mogelijke toekomstige onderzoeksvragen. Ten slotte omvat 6.6 een reflectie op het onderzoek gegeven vanuit een algemene beschouwing op de gemaakte keuzes dit de richting van het onderzoek bepalen.

6.1 Antwoord op de hoofdvraag in de vorm van maatregelen voor portaaleigenaren

Met behulp van de volgende hoofdcategorieën van criteria kan het aanbod van open data gebruiksvriendelijker worden gemaakt:

1. Formaat: door middel van gelinkte open data-indelingen waarbij de databron met een betekenisvolle naam is gepresenteerd. Dit maakt datagebruikers mogelijk om door de open data te navigeren met de context als achtergrondbasis.

2. Metadata: om de vindbaarheid van open data via de zoekmachine van het portaal te verbeteren, betere zoektermen te definiëren die de inhoudelijke relevantie dekken en een apart bestand bij te voegen met de informatiestructuur van het open databestand. Hiermee kan sneller worden gestart met het uitvoeren van de taken voor een data-analyseproject.

3. Toegankelijkheid: om open data toegankelijker en navigeerbaar te maken voor datagebruikers. Dit kan bijvoorbeeld door middel van de informatiestructuur van datasets aan te passen, of webpagina's te ontwikkelen in gestandaardiseerd HTML webformaat, waarop open data kunnen worden gepresenteerd.

4. Kwaliteit: datasets worden verbeterd naar aanleiding van fouten die worden gevonden, zoals uitschieters zonder valide reden, dubbele gegevens, en onvolledigheid. Ook kan een terugkoppeling naar dataproviders over de datakwaliteit leiden tot feedback voor een verbetering in de kwaliteit van datasets.

Analyseren van datasets op het portaal is onmogelijk, zonder veel voorwerk te verrichten om open data te kunnen gebruiken met behulp van tools en zelf programmeren om de verwerking mogelijk te maken. Visualisatie is hier onderdeel van, omdat de resultaten dan effectief kunnen worden geïnterpreteerd op het portaal in plaats van via omwegen met andere tools. Tenslotte

ontbreken er twee belangrijke zaken: (i) interactiemogelijkheden met anderen op het portaal, om bijvoorbeeld van elkaars inzichten te kunnen leren op basis van dezelfde dataset en (ii) tutorials waarmee vaardigheden kunnen worden aangeleerd om het portaalgebruik mogelijk te maken en kansen niet onbenut blijven voor iedereen die interesse heeft in open data. Kortom, het streven is om het portaal gebruiksvriendelijk te maken voor datagebruikers door middel van een technische invalshoek maar wel rekening houdend met wettelijk richtlijnen.

Met de volgende maatregelen kan rekening worden gehouden met wettelijke richtlijnen, bij het aanbieden van open data op het portaal:

1. Privacygevoelige open data anonimiseren.

2. Open gestandaardiseerde licenties voor datasets op het portaal en kostentoerekening voor het gebruik van webservices bij het aanbieden van grotere hoeveelheden open data.

3. Proactief aanbieden van open data op het portaal.

- Implementeren van negen ontbrekende functies zodat datagebruikers open data gebruikersvriendelijker kunnen verwerken op het portaal.

- Het concept van gebruikerspaden verder ontwikkelen en uitbreiden voor doelgerichte data-analyses.

Deze maatregelen zijn tot stand gekomen aan de hand van verkregen onderzoeksresultaten die naar aanleiding van de specifiekere subonderzoeksvragen staan toegelicht in de volgende paragraaf.

6.2 Onderzoeksresultaten per subonderzoeksvraag

Het antwoord op de hoofdvraag is na beantwoording van vier subonderzoeksvragen tot stand gekomen. De relatie tussen deze onderzoeksvragen is functioneel voor de onderzoeksresultaten. De bestaande oplossingen aan de hand van de eerste onderzoeksvraag worden gebruikt voor de daaropvolgende tweede onderzoeksvraag om te bepalen welke missen in het huidige portaal. Ter beantwoording van de derde onderzoeksvraag zijn aan de hand van een empirische gebruiksanalyse op het portaal gebruikerspaden gedefinieerd. Dit is een initiële aanzet die het gebruik van open data structureert voor eindgebruikers. Ten slotte zijn deze gebruikerspaden geëvalueerd door vijf respondenten te interviewen als antwoord op de vierde onderzoeksvraag.

1 Welke bestaande oplossingen uit de literatuur kunnen het opendataportaal gebruiksvriendelijker maken, rekening houdend met wettelijke richtlijnen?
De volgende hoofdcategorieën aan criteria kunnen worden gebruikt om open data te monitoren op gebruiksvriendelijkheid:

- Formaat: er zijn verschillende formaten waarin databronnen kunnen worden aangeboden aan datagebruikers

- Metadata: de context waarvoor of waaruit de dataset is samengesteld kan worden gerepresenteerd op basis van relevante termen en de informatiestructuur van de open data in een apart bijgevoegd bestand
- Toegankelijkheid: open data kunnen op verschillende manieren worden gepresenteerd, afhankelijk van de context waarop open data zijn gefundeerd
- Kwaliteit: datasets verschillen in kwaliteit en daarmee ook validiteit van modellen waarmee beslissingen worden onderbouwd

Wettelijke richtlijnen waarmee rekening moet worden gehouden door portaaleigenaren:

- Privacy-gevoeligheid van open data over natuurlijke personen
- De kosten en baten bij het ontsluiten van open data
- Geslotenheid of openheid bij het ontsluiten van open data

Ten slotte zijn er vijftien mogelijke functies gevonden als onderdeel van het portaal.

2 Wat ontbreekt er op het huidige portaal aan bestaande oplossingen voor datagebruikers?

Van initieel geselecteerd vijftien functies zijn er negen functies die ontbreken in het huidige portaal (Zuiderwijk et al., 2014; Zuiderwijk, 2015). De generaliseerbaarheid van de prioriteiten van de functies is beperkt doordat slechts vijf respondenten zijn gevraagd over het belang van de functies voor het huidige portaal:

- F12: Versiebeheer
- F4 : Analyseren
- F13: Data kwaliteitsbeheer
- F14: Tutorial aanbieden
- F6 : Visualiseren
- F10: Datagebruikers volgen
- F15: Interactiemechanisme
- F5 : Gebruikerspaden aanbieden
- F11: Analysebevindingen delen via sociale media

3 Met welke vorm van ondersteuning kan het gebruik van open data meer doelgericht worden gestructureerd voor specifieke datagebruikers?

Met behulp van een representatief gebruikerspad voor de aanpak van een data-analyseproject. Een gebruikerspad bestaat uit stappen die elk weer een serie taken representeren die onderdeel zijn van de stap. Hierbij kan doelgericht en efficiënt aan de slag worden gegaan met projecten waarbij data-analyse als methode een belangrijke rol vervult. Er zijn vier initiële gebruikerspaden gedefinieerd in de vorm van prototypes. Gebruikerspaden verschillen van elkaar aan de hand van karakteristieken die in de volgende categorieën zijn onderverdeeld:

- Analysetechnieken
- Gegevensvariabelen
- Gegevensmodellering
- Onderzoekaanpak
- Gegevensverwerving
- Modelprestatiemeting

Met deze categorieën kunnen nieuwe strategische aanpakken, in de vorm van gebruikerspaden, worden ontwikkeld voor datagebruikers met verschillende gebruikersbehoeften bij het gebruik van open data.

4 Klopt de volgorde en het aantal stappen als onderdeel van ieder gebruikerspad ter ondersteuning van datagebruikers?

Alle gebruikerspaden scoren hoger op basis van gemiddelde scores die zijn bepaald op een schaal van 1 tot en met 5, waarbij 5 erg belangrijk is en 1 onbelangrijk. De scores zijn: voor ondernemerpad (4,4), wetenschapperpad 2 (4,8), ontwikkelaarpad (4,6), en journalistpad (4,6).

Voor het verbeteren van de gebruikerspaden is de volledigheid in stappen geëvalueerd door de respondenten te vragen of er ontbrekende stappen zijn in de gebruikerspaden. Er zijn ontbrekende stappen aangegeven waardoor er ruimte is voor verbetering van de initieel gedefinieerde gebruikerspaden. Daarnaast is voor elk gebruikerspad geëvalueerd of de stappenreeks in elk gedefinieerde gebruikerspad de juiste volgorde hebben. Hier heeft elk gebruikerspad hoog gescoord, ondernemerpad (5), wetenschapperpad (4), ontwikkelaarpad (4), en journalistpad (3). Aangezien voor volgorde niet de maximale score is behaald bij de tweede derde en vierde gebruikerspad, is er ruimte voor verbetering van de volgorde van de stappen in de gedefinieerde gebruikerspaden. Met behulp van enquêtevragen zijn de gebruikerspaden op een kwantitatief-analytische manier beschouwd. Onder de respondenten is vastgesteld dat de gebruikerspaden relevant zijn voor data-analyse met ruimte voor verbetering. De gebruikerspaden zijn afzonderlijk door de respondenten geëvalueerd op basis van kwalitatief-analytische beschouwing. Met behulp van interviewvragen is vastgesteld wat de verschillende gebruikersdoelen en uitdagingen de respondenten ervaren bij het gebruik van open data en of er eventuele verbetersuggesties zijn voor de gebruikerspaden.

De volgende feedback is door de respondenten gegeven op de gebruikerspaden waaronder ook kritische kanttekeningen:

- Om gebruikerspaden effectief te maken zijn feedbackloops tussen verschillende datagebruikers en dataprovider nodig om de datakwaliteit te kunnen verbeteren. Daarnaast om de koppeling met andere datasets te herkennen.
- Een essentiële laatste stap voor een gebruikerspad is de beantwoording van de onderzoeksvraag en/of hypothese. Hierbij is bij de uitvoering van de stappen uitleg en tussenresultaten benodigd.
- Het is ook van belang dat datagebruikers open data opvragen en gebruiken met een afhankelijkheid van open data. In het geval van grote datasets, kan het meer tijd kosten als alles moet worden gescheiden in subsets. Of bij een te kleine gegevensreeks, die vaker moet worden bijgewerkt. Het ligt aan praktische termen wat er uit de gegevens moet worden gehaald.
- De huidige gebruikerspaden zijn ontoereikend voor wetenschappers die meer diepgaande data-analyses willen uitvoeren. Misschien zijn de huidige gebruikerspaden voor journalisten wel relevant. Dat ligt eraan of de gebruikerspaden daadwerkelijk in de praktijk zullen worden gebruikt in hun huidige vorm.
- De afzonderlijke stappen in een gebruikerspad zouden meer kunnen worden gespecificeerd op basis van het domein waar het individuele onderzoek onder valt. Er is namelijk sprake van verschillende regio's en domeinen van onderzoek, waar het opzetten van experimenteeronderzoek bij hoort. Hiervoor moeten wel de doelstellingen worden uiteengezet.

- Het is niet nodig om datasets op te slaan voor de analyse aangezien gegevens kunnen worden geaggregeerd en daarmee meerdere databronnen worden gebruikt bij het analyseren van datasets. De evaluatietechnieken voor de modellen zijn afhankelijk van de soort data-analysetechnieken die zijn gebruikt. Bijvoorbeeld voor de evaluatie van een toegepaste clusteranalyse, is gemiddelde kwadratische fout niet geschikt, maar voor regressiemodellen juist wel geschikt om modelprestaties uit te drukken. In het algemeen moet de datakwaliteit worden beoordeeld voordat data-analysemodellen kunnen worden gebouwd. Hiermee kan worden voorkomen dat de gemiddelden in datasets te sterk worden beïnvloed door uitschieters zonder valide reden, maar door de slechte datakwaliteit. Hierdoor zijn alle gebruikte data-analysemethoden afhankelijk zijn van de onderzoeksdoelen die aan het begin van een gebruikerspad zijn geformuleerd. De datagebruiker moet namelijk aan het eind ook een hypothese en/of onderzoeksvraag beantwoorden.

6.3 Contributie van het onderzoek aan wetenschap en maatschappij

De wetenschappelijke contributie van dit onderzoeksproject is gelegen in de gebruikerspaden. Dit is het eerste onderzoek dat gebruikerspaden definieert voor open data in de vorm van grijpklare strategische aanpakken. De gebruikerspaden generaliseren de aanpak om open data te analyseren op basis van vooropgestelde onderzoeksvragen en/of hypotheses. Hiermee kunnen datagebruikers direct in actie komen na het kiezen van een gebruikerspad om zich vervolgens te focussen op de open data-analyses zelf. Datagebruikers kunnen bij implementatie van de ontbrekende functies een gebruiksvriendelijker portaal tegemoetzien met meer en betere functies. Daarbij kunnen gebruikerspaden worden gebruikt als leidraad ter ondersteuning voor specifieke gebruikersdoelen van datagebruikers. Het ontwikkelen van gebruikerspaden voor het analyseren van open data op een gestructureerde manier is een complex proces. Dit komt doordat vele iteraties zijn vereist naast het gedetailleerd begrijpen van specifieke gebruikersbehoefte.

Het voordeel van gebruikerspaden is dat er tijd kan worden bespaard als er gelijk aan de slag wordt gegaan met een strategische aanpak. Dit onderzoek heeft een eerste aanzet gedaan tot de ontwikkeling van gebruikerspaden met de intentie om datagebruikers te ondersteunen zonder diepgaande analytisch onderzoek te doen naar de gebruikersbehoeften van datagebruikers.

De praktische contributie van dit onderzoeksproject bestaat uit de volgende zaken. Naast de negen ontbrekende functies in het huidige portaal is het mogelijk om het aanbod van open data te verbeteren door eerst te onderzoeken waar verbetermogelijkheden zijn met behulp van criteria en scores. Daarnaast is het van belang om rekening te houden met wettelijke richtlijnen, waarvoor maatregelen zijn gepresenteerd. Tenslotte is het ontwikkelen van gebruikerspaden voor specifieke datagebruikers vernieuwend en maakt het open data gebruiksvriendelijker. Met behulp van gebruikerspaden worden doelgerichte analyses namelijk gestructureerd aangeboden voor hergebruik door datagebruikers die drempelloos in actie kunnen komen. Dit nieuwe concept kan helpen om de aantrekkelijkheid van het gebruik van open data te verbeteren met behoud van ruimte voor creativiteit bij toepassing van data-analyses.

6.4 Beperkingen van de onderzoeksopzet

In deze paragraaf zullen de aspecten van het onderzoeksproject en die de generalisatie van de uitkomsten van het onderzoeksproject belemmeren worden besproken. Er zijn drie van deze belemmeringen te benoemen.

Allereerst is er de enquête die is gehouden onder vijf PhD-studenten, aan de Technische Universiteit Delft, die onderzoek verrichten op het gebied van open overheidsgegevens. De scope voor deze onderzoeksmethode is beperkt tot deze deskundige respondenten maar had ook naar onderzoekers op andere universiteiten kunnen worden uitgebreid. Tevens hadden eindgebruikers kunnen worden geïnterviewd om gebruikersbehoeften bij het gebruik van open data vast te kunnen leggen. Er is namelijk meer onderzoek nodig om specifieke gebruikersbehoeften buiten het wetenschappelijke domein vast te kunnen stellen. De vijf respondenten zijn op basis van expertise op het gebied van open data, en open overheidsgegevens als specifiek segment in het aanbod van open data. Ze hebben diverse gebruiksdoelen bij het gebruik van open data. Onder andere het testen van modellen, en open data analyseren voor inzichten in praktijksituaties zoals verkeer of uitgevoerde overheidsprojecten.

De tweede beperking is dat een opendatadeskundige invloed gehad op de richting van het literatuuronderzoek via suggesties voor wetenschappelijke literatuur. De opendatadeskundige is gespecialiseerd in open overheidsgegevens, oftewel open data die door de overheid beschikbaar zijn gesteld voor hergebruik. Daarom is verbreding van het huidige literatuuronderzoek mogelijk indien een opendatadeskundige met een andere perspectief wordt geïnterviewd als leidraad voor het literatuuronderzoek. Bijvoorbeeld een opendatadeskundige die focust op alle vormen van open data en de koppeling daartussen.

Verder is de ontwerpaanpak beperkt in de zin van hoe er tot de selectie van relevante functies is gekomen en de gebruikerspaden evenals de beschrijving van de karakteristieken ervan. Het vergelijken van het huidige portaal met de functies uit de literatuur is een grillige aanpak in de zin dat het niet duidelijk is hoe de ontwerpvolgorde van het portaal is geweest en hoe de huidige functies tot stand zijn gekomen. Zou dat zijn door een vraag vanuit het bedrijfsleven? Of door beschikbare archiveringstechnologie (CKAN) voor open data die ook door andere landen zijn gebruikt? Vermoedelijk is het door wetgeving waarbij de belastingbetaler recht heeft op data, tot stand gekomen. De overheid kan voor het bevorderen van transparantie, openheid geven van zaken via open data waardoor beleid controleerbaar blijft voor burgers. Hierdoor zou het portaal niet alleen als een samensmelting van functies kunnen worden gezien maar vooral hoe het beoogt een doel te behalen voor verschillende belanghebbenden.

Een alternatieve methode voor de bepaling van functies op het portaal, is het interviewen van eindgebruikers voor de bepaling van de gebruikersbehoefte bij het portaal. Dit kan aan de hand van een requirementsanalyse bij stakeholders, waarna de gebruikersbehoefte kan worden vertaald naar eisen voor het portaal in de vorm van usecases.

Ook is er een beperking voor hoe de functies zijn geprioriteerd. Dit is namelijk gedaan op basis van de scores die zijn toegekend door de verschillende respondenten met verschillende expertises. Dit zijn vijf experts op het gebied van open data. Maar de functies zijn bedoeld voor niet-experts en daarmee is er een beperking aan de relevantie van de prioriteitsbepaling van de functies bij het implementeren ervan in het portaal. Om goed te prioriteren zouden eindgebruikers moeten worden geïnterviewd.

De functies waarmee het huidige portaal kan worden verbeterd zijn in het belang van potentiële eindgebruikers en portaaleigenaren. Ontwikkelaars kunnen het portaal beschouwen als een

databron en gebruiken mogelijk eigen software om hun gebruiksdoelen te behalen. Voor alle datagebruikers die online open data willen analyseren zijn de functies een toegevoegde waarde omdat ze gericht zijn op het individueel uitvoeren van data-analyses.

De gebruikerspaden structureren het gebruik van open data aan de hand van een reeks te volgen stappen als eerste aanzet. Dit limiteert de betrouwbaarheid van de gebruikerspaden doordat lastig is na te gaan of daarmee gebruikersbehoeftes zijn gerepresenteerd. Door evaluatie met de vijf respondenten is dit deels opgevangen. Maar meer onderzoek is vereist die de gebruikersbehoefte beter in kaart brengt voor gebruik van open data.

Daarnaast zijn de karakteristieken van de gebruikerspaden gerepresenteerd door verschillende hoofdcategorieën. Dit is bij het ontwerpen van de gebruikerspaden veelzeggend voor de strategische vorm bij het analyseren van open data. Maar dit hoeft niet zo te zijn voor doelgroepen die gebruik willen maken van open data vanuit een geheel andere insteek dan gebruikerspaden als het initiatief. Een wetenschapper kan bijvoorbeeld onderzoek doen, en nog niet hebben besloten of hij/zij gebruik wilt maken van open data als bron van informatie terwijl de onderzoeksvragen al zijn gedefinieerd. Dit is een beperking van de gebruikerspaden, die uitgaan van een compleet onderzoeksaanpak en niet uitgaan van het voorwerk van datagebruikers.

Voor het ontwikkelen van een zoekfunctie bij de verschillende gebruikerspaden kan gebruik worden gemaakt van categorieën die de verschillende gebruikerspaden karakteriseren. Dat laatste kan flexibiliteit voor het gebruik van gebruikerspaden ook bevorderen aangezien niet van elke categorie of stap gebruik hoeft te worden gemaakt binnen een gekozen gebruikerspad.

6.5 Toekomstige onderzoeksvragen

De volgende toekomstige onderzoeksvragen zijn gepresenteerd, als suggestie voor nieuw onderzoek naar aanleiding van de hiervoor genoemde beperkingen van dit onderzoeksproject. Het bevragen van respondenten in dit onderzoek was beperkt tot een doelgroep met een wetenschappelijke achtergrond – de vijf PhD-studenten. Andere groepen zouden ook moeten worden ondervraagd om een volledig beeld te krijgen van mogelijke datagebruikers. De volgende onderzoeksvraag behoort bij dit vervolgonderzoek: *Welke verschillende gebruikerspaden voor verschillende typen gebruikers zijn relevant?*

Met relevant wordt bedoeld dat het gebruikerspad een waardevolle aanpakstructuur is voor een datagebruiker bij het beantwoorden van een onderzoeksvraag en/of hypothese met behulp van open data. De doelgroepen die kunnen worden gebruikt om het aantal gebruikerspaden uit te breiden zijn de volgende:

- Journalisten op basis van beschrijvende statistiek, open en transparante overheid.
- Burger als eindgebruiker; om te leren begrijpen en toepassen van analysetechnieken voor beter volgen en begrijpen van overheidsbeleid door ontsluiting van open data.
- Ambtenaren, bij de overheidsinstelling als gebruiker, via geavanceerde analysetechnieken.
- Ontwikkelaars of infomediairs van applicaties naar aanleiding van verkregen analytische inzichten, met behulp van open data.

Ook zou het literatuuronderzoek kunnen worden uitgebreid naar literatuur die wordt aanbevolen door onderzoekers uit het open dataveld met verschillende achtergronden. Voor dit onderzoek is de volgende onderzoeksvraag van toepassing: *Welke type datagebruikers maakt gebruik van open data en welke achtergrond hebben zij?*

Tot slot kan de ontwerpaanpak worden gefocust op een breder publiek, om tot een ontwerp van gebruikerspaden en karakteristieken te komen die meer specifiek zijn toegespitst op verschillende gebruikers. Daarmee is het gebruik ervan in het algemeen realistischer. Dit leidt tot de volgende onderzoeksvraag: *Naar welke inzichten worden gezocht in open data en welke data-analyses zijn hiervoor nodig?*

De mogelijkheden voor vervolgonderzoek maken het mogelijk om nieuwe onderzoeksvragen te maken voor specifieke richtingen. Dit onderzoek is gefundeerd op onderzoeksvragen waarop is gereflecteerd in de volgende paragraaf.

6.6 Reflectie op gemaakte keuzes voor onderzoek

In deze reflectie is het onderzoeksproject beschouwd vanuit een algemene kijk op het onderzoek.

Allereerst zijn er de keuzes die zijn gemaakt voor de hoofd- en subonderzoeksvragen. Deze vragen zijn gelinkt aan de content in de verschillende hoofdstukken vanaf hoofdstuk 2 en verder. Gezamenlijk hebben ze geleid tot de onderzoeksresultaten voor het beantwoorden van de hoofdvraag. Ze staan in dienst van de hoofdvraag, en met behulp van onderzoeksmethoden zijn ze beantwoord. Alle elementen hebben uiteindelijk een bijdrage geleverd aan het beantwoorden van de hoofdvraag en het komen tot de wetenschappelijke- evenals maatschappelijke bijdrage van dit onderzoeksproject. Dit betekent dat het onderzoeksproject goede samenhang heeft voor het verkrijgen van de onderzoeksresultaten.

Het project was realistisch uitvoerbaar, zowel in de tijdspanne van het onderzoeksproject als de planning van tussenliggende onderzoeksresultaten waarvoor ook diverse experts zijn geïnterviewd Er is rekening gehouden met onvoorziene omstandigheden van respondenten. Zoals een conferentiebezoek waardoor de interviews gespreid en per respondent werden afgesproken om zo de informatie te verkrijgen na een doorlooptijd van meerdere weken. Ook was begonnen met het literatuuronderzoek en werden de resultaten besproken alvorens door te gaan met het evalueren van mogelijke oplossingen.

De volgende zaken zouden met de kennis van nu anders zijn aangepakt. De analyse van het huidige Nederlandse portaal zou in combinatie met een vergelijking met andere portalen zijn uitgevoerd. Verder zou additioneel onderzoek naar de gebruikersbehoeften onder verschillende datagebruikers zijn uitgevoerd, om de interesse in open data te meten. Op deze manier kan er een duidelijker beeld worden geschetst vanuit burgers op open data. Het is namelijk de vraag of burgers interesse hebben in analyseren van open data en of zij de kennis hebben om dit te kunnen doen. Op deze manier kunnen eisen van het portaal nog specifieker gemaakt worden dan aan de hand van literatuur nu het geval is. Het vermoeden is dat regio's verschillen in interesse voor open data. Zo zou een regio als Amsterdam meer met open data kunnen dan een plattelandsgemeente waar doorgaans meer structuur is onder de bevolking, is de verwachting.

In dit onderzoeksproject heb ik meer systeemperspectief gebruikt dan actorperspectief om het portaal te verbeteren. Uiteindelijk is het ontwerp van aanpakken voor het analyseren van gegevens een systeem zelf, terwijl strategische belangen in de werkelijkheid de overhand voeren en daarmee het verschil kunnen maken tussen succes en teleurstelling van het portaal. Omdat zowel actorperspectief als systeemperspectief van toepassing zijn op het portaal, kan een geïntegreerde analyse van beide perspectieven leiden tot waardevolle inzichten. Deze inzichten gaan over hoe het ene perspectief domineert op het andere bij het maken van beslissingen over de verdere ontwikkeling van het portaal. Dit is inclusief de zoektocht naar informatie die het actorperspectief in beeld kan brengen en dit vraagt ook om onderzoek naar

de interactie van verschillende overheidspartijen. Mogelijk zijn die minder geïnteresseerd in systemen maar wel in hun doelstellingen die ze constant nastreven, waar mogelijk systemen een rol kunnen spelen. Gebruikerspaden en functies die als middel kunnen worden gebruikt voor de ontwikkeling van het gebruik van open data via het portaal is de juiste aanpak. Doordat het kansen voor toekomstig onderzoek en daarmee constante verbetering van het datagebruik niet in de weg ligt. Het leunt namelijk op wetenschappelijke ontwikkeling van data-analysetechnieken.

Tot slot is deze thesis een verlengstuk van het studieprogramma, waarbij de analyse en het ontwerp van socio-technische Multi-actor systemen centraal staat. Het portaal van de Nederlandse overheid is beschouwd als een dergelijke complexe socio-technische systeem dat kan bijdragen aan maatschappelijke vraagstukken. Dit onderzoek heeft gestreefd hieraan een bijdrage te leveren door de gebruiksvriendelijkheid van het portaal te verbeteren in het belang van alle stakeholders. Hierbij zijn naast bestaande technische oplossingen ook gebruikerspaden als nieuw concept geïntroduceerd om het gebruik van open data te structureren voor datagebruikers. Door het gevolgde studieprogramma heb ik de vaardigheden ontwikkeld om dit te doen op een gestructureerde manier. Hierdoor was de analyse ervan evenals het ontwerp van een oplossing de kern van de thesis en daarmee een contributie voor mijn studieprogramma. De contributie is dat ik perspectieven en een selectie van technieken uit mijn studie evenals algemene beschouwing van systemen heb gebruikt. Deze beschouwing kan worden herhaald voor het verbeteren van een open-dataportaal.

Literatuurreferenties

Alexopoulos, C., Loukis, E. and Charalabidis, Y. (2014). A Platform for Closing the Open Data Feedback Loop based on web2.0 functionality. *JeDem eJournal Of Democracy & Open Government*, Vol. 6, Iss. 1, pp. 62-68.

Application programming interface (API). (2018). In *Wikipedia*. Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Application_programming_interface

Atom (bestandsformaat). (2018). In *Wikipedia*. Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Atom_(bestandsformaat)

Atwal, A. and Caldwell, K. (2002). Do multidisciplinary integrated care pathways improve interprofessional collaboration? *Scandinavian journal of sciences,* Vol. 16, Iss. 4, pp. 360-367. https://doi.org/10.1046/j.1471-6712.2002.00101.x

Berners-Lee, T. (2019). 5* Open Data. Geraadpleegd op 10-02-2019 van: http://5stardata.info/.

Berthold, M. & Hand, D.J. (2007). Intelligent Data Analysis, An Introduction. *2nd revised and extended edition,* Springer.

Centraal bureau voor de statistiek (CBS). (2018, 20 december). Internationale handel; in-, uit- en wederuitvoer, SITC 2002-2015 [XML-Feed & JSON-API]. Geraadpleegd van XML-Feed: https://opendata.cbs.nl/ODataFeed/OData/70905ned en JSON-API: https://opendata.cbs.nl/ODataApi/OData/70905ned

Chan, C.M.L. (2013). From Open Data to Open Innovation Strategies: Creating e-Services Using Open Government Data. *2013 46th Hawaii International Conference on System sciences*, pp. 1890-1899. DOI: 10.1109/HICSS.2013.236.

CKAN ('Comprehensive Knowledge Archive Network'). (2018). In *Wikipedia.* (2018). Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Atom_(bestandsformaat)

Colpaert, P., Joye, S., Mechant, P., Mannens, E. and Van de Walle, R. (2013). The 5 Stars of Open Data Portals. *Proceedings MeTTeG*, Vol. 7, pp. 61-67. http://pieter.pm/5stardataportals/

Dawes, S.D., Vidiasova, L., and Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly,* Vol. 33 Iss. 1 pp. 15-27. https://doi.org/10.1016/j.giq.2016.01.003.

Ding, L., Peristeras, V., and Hausenblas, M. (2012). Linked Open Government Data. *IEEE Intelligent Systems*, Vol. 27 Iss. 3 pp. 11-15. Doi: 10.1109/MIS.2012.56.

Drupal. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Drupal

Duval, E., Hodgins, W., Sutton, S., and Weibel, S.L. (2002). Metadata Principles and Practicalities. *D-lib Magazine Access Terms and Conditions.* Doi: 10.1045/april2002-weibel.

Geels, F.W. and Schot, J. (2007). Typology of sociotechnical transition pathways. *Research Policy*, Vol. 36, Iss. 3, pp. 399-417. https://doi.org/10.1016/j.respol.2007.01.003

Hossain, M.A., Dwivedi, Y.K. & Rana, N.P. (2015). State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of organizational Computing and Electronic Commerce*, 23:1-2, 14-40, Doi: 10.1080/10919392.2015.1124007.

Hypertext Markup Language (HTML). (2018). In *Wikipedia*. Geraadpleegd op 31-12-2018, van https://en.wikipedia.org/wiki/HTML

Iamamphai, P., Noymance, J., San -Um, W. and Pasupa, K. (2016). Investigations and Comparisons of Government Open Data Websites through Systematic Functional Analysis and Efficient Promotion Approach. *The 2016 Management and Innovation Technology International Conference (MITiCON-2016)*, pp. 142-147.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning, with applications in R*. Springer. Doi: 10.1007/978-1-4614-7138-7. http://www-bcf.usc.edu/~gareth/ISL/

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, Vol. 29 pp. 258-268. https://doi.org/10.1080/10580530.2012.716740.

JavaScript. (2018). In *Wikipedia*. Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/JavaScript

Jeffery, K. (z.d.). Metadata: an overview and some issues. Geraadpleegd op 10-02-2019 van: https://www.ercim.eu/publication/ws-proceedings/11th-EDRG/jefferey.pdf

JSON. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/JSON

Krishnamurthy, L., Nadeau, J., Ozsoyoglu, G., Ozsoyoglu, M., Schaeffer, G., Tasan, M. and Xu, W. (2003). Pathways database system: an integrated system for biological pathways. *BioInformatics*, Vol. 19, No. 8, pp. 930-937. Doi: 10.1093/bioinformatics/btg113.

Machine Learning (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://en.wikipedia.org/wiki/Machine_learning

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2015). Open overheid in actie, actieplan 2016-2017. Geraadpleegd op 21 December 2018, van: https://open-overheid.nl/wp-content/uploads/2016/02/actieplan-open-overheid-2016-2017-2.pdf

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2018, 20 december). Internationale handel; in-, uit- en wederuitvoer, SITC 2002-2015. [Link]. Geraadpleegd van: https://data.overheid.nl/dataset/e2b41b8d-da40-4846-a515-7a39f5896ca2

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (z.d.). *Data* [screenshot]. Geraadpleegd op 30 december 2018, van https://data.overheid.nl/dataset

MySQL. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/MySQL

Open data van de Overheid. (2018, 30 december). Dataregister van de Nederlandse overheid [website]. Geraadpleegd van https://data.overheid.nl

OpenDataMonitor. (2018). Open Data Terminology. http://knowhow.opendatamonitor.eu/odresearch/open-data-terminology/

Opensourcesoftware. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Opensourcesoftware

Overheid.nl.(2018). Open data van de Overheid, over deze website. Geraadpleegd op 20 December, 2018 van: https://data.overheid.nl/over-deze-website.

Oxford. (2018). Pathway. English Oxford Living Dictionaries. Geraadpleegd op 18 September 2018, van: https://en.oxforddictionaries.com/definition/pathway

Peffers, K., Tuunanen, T., Rothenberger, M.A., and Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, Vol. 24 Iss. 3 pp. 45-78.

Portable document format (pdf). (2018). In *Wikipedia*. Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Portable_document_format

Publicatieblad van de Europese Unie (PbEU) (2016, 4 mei). Geraadpleegd op 7 November 2018,van:https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/richtlijn_2016_-_680_definitief.pdf

Publicatieblad van de Europese Unie (PbEU). (2013, 27 juni). Geraadpleegd op 7 November 2018, van: https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32013L0037&from=FR

Publicatieblad van de Europese Unie (PbEU). (2014, 24 juli). Geraadpleegd op 7 November 2018, van: https://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:52014XC0724(01)&from=EN

Ragsdale, C.T. (2008). Spreadsheet Modeling & Decision Analysis. *Fifth edition.*

RDF/XML. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/RDF/XML

Resource Description Framework (RDF). (2018). In wikipedia. Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Resource_Description_Framework

Rozanski, N. and Woods, E. (2005). *Software Systems Architecture, working with Stakeholders Using Viewpoints and Perspectives.* Addison-Wesley.

SOAP. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://en.wikipedia.org/wiki/SOAP

Thorsby, J., Stowers, G.N.L., Wolslegel, K. and Tumbuan, E. (2016). Understanding the content and features of open data portals in American cities. *Government Information Quarterly*. http://dx.doi.org/10.1016/j.giq.2016.07.001.

Ubaldi, B. (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance,* No. 22, OECD publishing, Paris. http://dx.doi.org/10.1787/sk46bj4r03s7-en.

Uniform resource identifier (URI). (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Uniform_resource_identifier

Uniform Resource Locator (URL). (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Uniform_Resource_Locator

Usecasediagram. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Usecasediagram

Webservice. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://nl.wikipedia.org/wiki/Webservice

XML. (2018). In *Wikipedia.* Geraadpleegd op 31-12-2018, van https://en.wikipedia.org/wiki/XML

Zuiderwijk, A. (2015). Open Data Infrastructures, The design of an infrastructure to enhance the coordination of open data use. *PhD thesis*.

Zuiderwijk, A., and Janssen, M. (2015). Towards decision support for disclosing data: Closed or open data? *Information Polity*, vol. 20 pp. 103-117. IOS Press. Doi: 10.3233/IP-150358.

Zuiderwijk, A., Janssen, M. & Parnia, A. (2013). The complementarity of open data infrastructures: An analysis of functionalities. *Paper presented at the 14th Annual International Conference on Digital Government Research*, Quebec, Canada.

Zuiderwijk, A., Janssen, M. and Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity*, Vol. 19, IOS Press, pp. 17-33. Doi: 10.3233/IP-140329.

Bijlage 1: interview met opendatadeskundige

Naar aanleiding van een interview, zijn er notulen gemaakt en deze zijn gevalideerd met de geïnterviewde deskundige. Dit is de eerste aanzet voor het literatuuronderzoek dat is uitgevoerd voor dit onderzoek. De resultaten van dit interview zijn gebruikt om relevante literatuur te selecteren.

Interview met opendatadeskundige Mevr. Zuiderwijk

Datum: 14-06-2018
Tijd: 12.30-13.30uur

Antwoord naar aanleiding van eerste vraag:
'Welke functionaliteiten zijn er voor het hergebruik van Public Sector Information (PSI) via open data portalen die Open Government Data (OGD) publiceren? Welke relevante titels zou ik hiervoor kunnen raadplegen?'

Tip 1: PhD thesis over open data, hierin staan in het bijzonder in hoofdstuk 6 functionaliteiten voor een open data OGD-infrastructuur. In hoofdstuk 5 staan requirements. (Open Data Infrastructures, The design of an infrastructure to enhance the coordination of open data use (by Zuiderwijk, 2015)

Daarnaast is er een conferentie artikel, met een vergelijking van 3 verschillende portalen. (Titel: "The complementarity of open data infrastructures: An analysis of functionalities" door Zuiderwijk et al., 2013)

Open data maturity rapport (titel: "Open data for a European data community" door European Data Portal, 2017). Hier wordt gekeken naar wat belangrijk is met betrekking tot data op portalen. (Machine -readable, trendsetting etc.)

Open data benchmarking, waaronder barometer, index, Scoreboard om Open Data adoption te kunnen beoordelen inclusief geleerde lessen. (Title: "Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned" door Susha et al., 2015).

Artikelen over open data waaronder:

"Innovation with open data: Essential elements of open data ecosystems" door Zuiderwijk et al. (2014)

"Open Data for Competitive Advantage: Insights from Open Data Use by Companies" door Zuiderwijk et al. (2015)

"The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments" door Zuiderwijk et al. (2016)

"Open data for democracy: Developing a theoretical framework for open data use" by Ruijer et al. (2017)

"Lifting Open Data Portals to the Data Web" door van der Waal et al. (2014)

"Towards decision support for disclosing data: Closed or open data?" door Zuiderwijk and Janssen (2015)

"Special Issue on Innovation through Open Data - A Review of the State-of-the-Art and an Emerging Research Agenda: Guest Editors' Introduction" door Zuiderwijk et al. (2014)

"Open Government Portals Assessment: A Transparency for Accountability Perspective" door Lourenco. (2013)

"An open data and open services repository for supporting citizen-driven application development for governance" door Charalabidis et al. (2016)

"Open Data Platforms and Their Usability: Proposing a Framework for Evaluating Citizen Intentions" door Kapoor et al. (2015)

"An analysis of open government portals: A perspective of transparency for accountability" door Lourenco (2015)

"Investigations and Comparisons of Government Open Data Websites through Systemic Functional Analysis and Efficient Promotion Approach" door Iamamphai et al. (2016)

Ieder artikel moet gescand worden of het bruikbaar is voor het onderzoek.

Tussenvraag: '*Welke actoren ziet u als belanghebbende bij het open data portaal en functie voor gebruik ervan'.*

- Ambtenaren zelf, werkzaam bij overheidsinstelling
- Journalisten, om nieuwswaardige conclusies te trekken uit ruwe open data
- Onderzoekers, gebruik van ruw data voor wetenschappelijk onderzoek
- (Eind)gebruikers, burgers
- Ontwikkelaars (Intermediair of infomediar), bijv. app ontwikkelaars
- Bibliothecarissen en archivisten (voor duurzaamheid, opslag en om toekomstig gebruik mogelijk te maken)

Tip 2: Bekijk de video's over open data gebruik en over open data infrastructuur. Referentiesheet met links naar relevante videos: Open Data Governance and Use, An overview of all the videos used in the course. (OD01xP - 2016)

Antwoord naar aanleiding van tweede vraag: *'Welke theorieën kunnen het belang van actoren en hun rol voor hergebruik van publieke sector informatie? En welke Keyboards kan ik gebruiken om relevante publicaties te vinden?'*

Resulterend in de volgende zoekresultaten na het invoeren van de relevante zoektermen in GoogleScholar. Ze kunnen in het literatuuronderzoek als een tabel worden weergegeven:

(AND en OR kunnen afwisselend worden gebruikt bij dezelfde zoektermen in de zoekbalk.)

"Open data" AND/OR functionalit*

Gekozen resultaat:

(p. 2) A platform for closing the open data feedback loop based on Web 2.0 functionality (by Alexopoulos et al., 2014)

(p. 4) Barriers and development directions for the publication and usage of open data: A socio-technical view (by Zuiderwijk & Janssen, 2014)

"Open data" AND/OR portal*

(Socio-technial Impediments of Open Data by Zuiderwijk et al., 2012)

Antwoord naar aanleiding van tweede vraag: *"Welke criteria zijn er bekend, waar een multi-actor systeem, open data portaal met verschillende functionaliteiten die voorzien in de rollen van stakeholders?"*

"Open data" AND/OR platform*

(p. 10) The 5 stars of open data portals (by Colpaert et al., 2013)

"Open data AND/OR ecosystem*

(p. 2) From open data to open innovation strategies: Creating e-services using open government data (by Chan, 2013)

"Open data" AND/OR requirement*

(Is al aanwezig, Zie Chapter 5 PhD thesis by Zuiderwijk, 2015)

"Open data" AND/OR infrastructure*

(Is al aanwezig, Zie publicatie *A Platform for Closing the Open Data Feedback Loop based on Web2.0 functionality (by Alexopoulos)*

Tussenvraag: *Welke auteur zou u mij aanraden voor informatie over open data portalen?*

Gebruikte zoekterm: "Dawes open data ecosystem"

Gekozen resultaat: Planning and designing open government data programs: An ecosystem approach (by Dawes et al., 2016).

-(opmerking) Over het algemeen genomen is er weinig theorie over open data, waardoor buiten het domein moet worden gezocht.

-(opmerking) Beleidsmakers kunnen leren van conclusies uit open data portalen.

-(opmerking) Publisher kan worden geassocieerd met intermediairs, dat zijn mediators tussen technische ontwikkelaars en beleidsmakers.

Tussenvraag: *hoe zou u binnen dit onderzoek kunnen worden gerefereerd als expert?*

Expert Mevr. Zuiderwijk behoort tot de meest invloedrijke personen op het gebied van open data onderzoek (Referentie: "State-of-the-art in open data research: Insights from existing literature and a research agenda" By Hossain et al., 2015).

Bijlage 2: overzicht na analyse van portaalfunctionaliteiten

In tabel 13 staan de resultaten na een enquête te hebben gehouden onder vijf deskundigen. Deze deskundigen zijn als volgt geselecteerd. Ze waren allereerst beschikbaar en hadden op korte termijn gelegenheid voor het geven van een interview. Ten tweede, ze deden allen onderzoek binnen de TU Delft op het gebied van open data en vielen binnen een onderzoekgroep die focust op open overheidsgegevens. Voor dit onderzoek had het ook gekund om deskundigen van andere universiteiten te interviewen, of standaard eindgebruikers waarvan de gebruikersbehoeften kunnen worden onderzocht. De functies zijn geëvalueerd met behulp van deskundigen doordat het optimaal gebruik van open datamogelijkheden ook afhankelijk is van de mogelijkheden die de wetenschap biedt om het gebruik van open data te ondersteunen. Het belang van de functies is onder de deskundigen gemeten op een schaal van 1 tot 5.

Tabel 133: Functiebelang, Schaal: 1= niet belangrijk- 5= erg belangrijk

Functie	Deskundige 1	Deskundige 2	Deskundige 3	Deskundige 4	Deskundige 5	Gemiddelde
F1	5	5	5	5	5	5
F2	2	4	5	5	5	4,2
F3	1	1	2	2	3	1,8
F4	4	5	4	5	5	4,6
F5	3	4	1	5	4	3,4
F6	3	5	5	5	4	4,4
F7	5	4	5	5	5	4,8
F8	3	5	4	5	2	3,8
F9	2	5	4	5	5	4,2
F10	3	4	3	5	4	3,8
F11	4	3	2	2	2	2,6
F12	5	5	5	5	5	5
F13	4	4	5	5	5	4,6
F14	4	5	5	5	4	4,6
F15	4	4	5	1	4	3,6

De opgestelde scores zijn gedeeld door 5 om het gemiddelde voor elk onafhankelijke functie uit te rekenen. De gemiddelde score van alle functiescores is 4.0. Op tabel 12 vallen wederom F3 & F11 op door de lager dan gemiddelde scores. Met uitzondering van deze twee zwak

scorende functies, onder de vijf deskundigen, scoren alle andere functies bovengemiddeld op belangrijkheid. Met uitschieter van F1 en F12, wat respectievelijk zoekfunctie naar datasets (F1) en Versie management voor datasets (F12) is. De laatstgenoemde twee functies worden dus essentieel gevonden voor het portaal.

Bijlage 3: feedback op gebruikerspaden per respondent

De gebruikerspaden werd de volgende open vraag gesteld: 'Welke open data gebruikersdoelen zijn momenteel voor u van toepassing? Beschrijf deze.'
Daarnaast is feedback gegeven de vragen: ‘Zijn er ontbrekende stappen? En klopt de volgorde van de stappen?

De eerste respondent is van mening dat het gebruik van open data te maken heeft met het ontwikkelen van modellen, om ze vervolgens te kunnen testen op beschikbare open data. Hierbij is de datakwaliteit het meest uitdagende onderdeel voor het gebruik van open data. De respondent stelt hierbij voor om feedbackloops te stimuleren tussen verschillende datagebruikers en dataprovider van de open data om de kwaliteit te kunnen verbeteren. Daarnaast om de koppeling met andere datasets te herkennen.
De tweede respondent gebruikt open data voor wetenschappelijke analyses. Bijvoorbeeld om trends in verkeer te ontdekken voor logistieke inzichten. Het meest uitdagende bij het gebruik van open data is hierbij: onvolledige datasets, structuur van de dataset, en niet-machineleesbaarheid. Een essentiële laatste stap voor een gebruikerspad is de beantwoording van de onderzoeksvraag en/of hypothese. Hierbij is bij de uitvoering van de stappen uitleg en tussenresultaten benodigd.

Voor de derde respondent is open data een bron voor vertrouwen en transparantie van de overheid. Daarnaast is open data een bron voor zakelijke kansen voor nieuwe economische locaties waarvan de samenleving kan profiteren. Ook kan overheidsbeleid worden verbeterd doordat de burgers in staat zijn om betere eisen te stellen met behulp van open data. Hierbij is het gedrag van overheidsdiensten voor de verstrekking van open data belangrijk. Het is ook van belang dat datagebruikers open data opvragen en gebruiken. Er is een afhankelijkheid van open data. In het geval van grote datasets, kan het meer tijd kosten als alles moet worden gescheiden in subsets. Of bij een te kleine gegevensreeks, die vaker moet worden bijgewerkt. Het ligt aan praktische termen wat er uit de gegevens moet worden gehaald. Hierbij is een terugkoppeling met de dataprovider nodig om de datakwaliteit te kunnen verbeteren. De huidige gebruikerspaden zijn te weinig voor wetenschappers die er meer nodig hebben. Misschien zijn de huidige gebruikerspaden voor journalisten wel relevant. Het ligt eraan of mensen het daadwerkelijk in de praktijk gaan gebruiken.

Volgens de vierde respondent kan open data worden gebruikt voor het creëren van meer voordelen als gebruiksdoel in het algemeen. Voor alle functies moet de technische veiligheid van gegevens en het waarborgen van privacy op actorniveau worden afgewogen. Om informatie voldoende te beveiligen tegen negatieve bijwerkingen zal daarom gegevensbeheer moeten worden toegepast. De afzonderlijke stappen in een gebruikerspad zouden meer kunnen worden gespecificeerd op basis van het domein waar het individuele onderzoek onder valt. Er is namelijk sprake van verschillende regio’s en domeinen van onderzoek, waar het opzetten van experimenteeronderzoek bij hoort. Hiervoor moeten wel de doelstellingen worden uiteengezet.

Voor de vijfde respondent is het gebruiksdoel het makkelijker maken van big data-analyse, door het bijeenvoegen van verschillende databronnen in de vorm van een raamwerk. De overheid zou bij alle gebruikerspaden moeten beoordelen in welke voordelen de ontsluiting

van open data kunnen resulteren. Het onderzoeksdoel moet vanaf het begin duidelijk zijn, bijvoorbeeld de aanduiding in de vorm van een hypothese. Dit hangt bijvoorbeeld af van de klant voor wie de open data wordt geanalyseerd. Het is niet nodig om datasets op te slaan voordat er wordt geanalyseerd aangezien gegevens worden geaggregeerd en daarmee meerdere databronnen worden gebruikt bij het analyseren van datasets. Het is wel nodig om de kwaliteit van open data te beoordelen alvorens de geproduceerde modellen te fitten op datasets. De evaluatietechnieken voor de modellen zijn afhankelijk van het soort data-analysetechnieken die zijn gebruikt. Bijvoorbeeld voor de evaluatie van een toegepaste clusteranalyse, is gemiddelde kwadratische fout (MSE) niet geschikt, maar voor regressiemodellen is MSE wel geschikt om modelprestaties uit te drukken. In het algemeen moet de datakwaliteit worden beoordeeld voordat data-analysemodellen kunnen worden gebouwd. Hiermee kan worden voorkomen dat de gemiddeldes in datasets te sterk worden beïnvloed door uitschieters zonder valide reden, maar door de slechte datakwaliteit. Doordat alle data-analysemethoden afhankelijk zijn van de onderzoeksdoelen die aan het begin van een gebruikerspad zijn geformuleerd. Moet de datagebruiker aan het eind ook een hypothese en/of onderzoeksvraag beantwoorden.

Bijlage 4: vernieuwde gebruikerspaden

In deze paragraaf zal worden beschreven welke aanpassingen zijn toegepast op de gebruikerspaden en op welke manier dit is gedaan ter verbetering ervan. De nieuwe versies van gebruikerspaden, na het verwerken van de feedback door vijf deskundigen, hebben tot een aantal additionele stappen geleid welke zijn aangegeven met 'extra stap':

1 Ondernemerpad

Stap 1: definieer de onderzoeksvragen of definieer hypothesen die u wilt bevestigen of ontkrachten door middel van statistische gevolgtrekking.
Stap 2: verkrijg gegevens van het portaal. Dit kan handmatig worden gedaan door de dataset te downloaden. Of automatisch, als u wilt dat het een herhaalbaar proces is op een URL (Uniform Resource Locator).
Extra stap A: de verkregen gegevens hebben invloed op de manier waarop de volgende stappen: 3, 4 en 5 worden uitgevoerd en/of gedefinieerd. Als er een grote dataset is, moet deze worden gescheiden in subsets, anders worden gegevens inefficiënt geanalyseerd, wat een verspilling van tijd is. Als er slechts een kleine dataset voorhanden is, kan deze na feedback aan de dataprovider worden bijgewerkt. Of gelinkt met andere mogelijk relevante datasets naar aanleiding van constructieve feedback.
Stap 3: ontleedt de gegevens voor kwaliteitsbeoordeling, met betrekking tot het doel ervan. Als dit acceptabel is, gaat u verder, anders herhaalt u de vorige stap met andere gegevens op het portaal bijvoorbeeld.
Extra stap B: geef feedback aan de dataprovider voor het onderzoeken van de datakwaliteit. Indien de kwaliteit van dataset niet naar wens is, kan dit kenbaar worden gemaakt aan de dataprovider in de vorm van constructieve feedback, ter verbetering ervan van de dataset.

Extra stap C: problemen met datakwaliteit delen met anderen vergelijkbare datagebruikers om interactie over de datakwaliteit mogelijk te maken met constructieve feedback tot gevolg. De datakwaliteit is een probleem wanneer er foute conclusies worden getrokken uit foutieve resultaten.
Stap 4: filter relevante subsets uit datasets en link naar andere (externe) gegevens, bereiken bepalen, categorieën definiëren enzovoorts.
Stap 5: de gegevens analyseren, te beginnen met beschrijvende statistieken. Als dit acceptabel is en er voldoende informatie kan worden verkregen met de dataset. Bijvoorbeeld met behulp van lineaire regressielijnen om relaties tussen geselecteerde variabelen te beschrijven. Dit is parametrisch indien u aannames maakt over de functie van het gegevensmodel dat u representatief voor de dataset wilt maken door middel van parametrische benadering (James et al., 2013: 61). Anders gebruikt u statistische inferentie, in de vorm van een niet-parametrisch model. Hierbij gebruikt u een techniek om de structuur van de gegevens te achterhalen. Bijvoorbeeld met Support Vector Machines (SVM) (Berthold & Hand, 2007: 181). Uiteraard moeten beide methoden bij het model representatief zijn voor dataset. Hiervoor kunnen verschillende testen worden gebruikt, bijvoorbeeld gemiddelde kwadratische fout (James et al., 2013: 29). Dit is om te kwantificeren in hoeverre de voorspelde responswaarde voor een gegeven responswaarde dichtbij de werkelijke respons in de dataset ligt.
Stap 6: presenteer de resultaten in de vorm van visualisaties die de resultaten bevatten in een onderzoeksrapport.

Extra stap D: geef feedback aan dataproviders voor het onderzoeken van de datakwaliteit. Bij het uitvoeren van de analyses kan aan de dataprovider fouten kenbaar worden gemaakt, in de vorm van constructieve feedback ter verbetering van datakwaliteit.
Stap 7: verfijn op basis van visualisaties het relevantie niveau van gegevensanalyses voor inzichten om onderzoeksvragen te beantwoorden. Als dit niet aanvaardbare antwoorden geeft op de onderzoeksvragen en/of hypothesen, herhaal dan de vorige stap.
Stap 8: deel visualisaties met andere gebruikers op het portaal en/of sociale mediakanalen.
Extra stap E: beantwoord de onderzoeksvragen en hypothesen volgens afgeleide resultaten en publiceer de resultaten in een peer-reviewed journaal. Met peer-reviewed bedoelen we dat verschillende deskundigen binnen het vakgebied waar de data-analyse onder valt, een kritische blik werpen op het geschreven werk in de vorm van feedback.

2 Wetenschapperpad

Extra stap A: begin met het definiëren van onderzoeksvragen en zet de onderzoeksdoelen uiteen.
Stap 1: onderzoeken en zoeken naar datasets op thema's, filters en gegevensattributen.
Extra stap B: geef feedback aan de dataproviders als er iets ontbreekt of fout is met de open data, of doe een datakwaliteitsbeoordeling. Er is hier geen standaard voor aangezien dit afhankelijk is van de data zelf en dient met een kritische blik te worden geëvalueerd. Zijn de metingen geschikt voor het gebruiksdoel is hierbij de hoofdvraag die verder kan worden uiteengezet in subvragen voor verschillende perspectieven op de geschiktheid van de dataset voor specifieke gebruiksdoelen.
Stap 2: hypotheses definiëren op basis van de gegevensverkenning die moet worden bevestigd of ontkracht.
Stap 3: data-analysemethoden selecteren. Bijvoorbeeld een voorspellingsmodel, in de vorm van randomforest waarmee via een verzameling van beslissingsbomen de variantie kan worden uitgemiddeld. Een techniek die beslissingsstructuren decoreert voor de classificatie van gegevens (James et al., 2013).
Extra stap C: uitleggen wat voor soort informatie is gewenst, en voor welke verwerkingstaken, om te begrijpen welke aanvullende gegevens nodig zijn.
Stap 4: open data verkrijgen van portaal, handmatig of automatisch herhaalbaar vanaf een URL.
Stap 5: ontleed de gegevens voor kwaliteitsbeoordeling voor het gebruikersdoel of onderzoeksdoel. Indien aanvaardbaar, ga door, anders herhaal voorgaande stap.
Extra uitleg voor stap 5: de criteria voor datakwaliteit moeten worden gedefinieerd om te beslissen of de gegevens nuttig zijn voor het beantwoorden van de onderzoeksvragen en/of het bevestigen of ontkrachten van de vooraf gedefinieerde hypothesen.
Stap 6: filter relevante attributen en koppel de gegevens eventueel aan andere (externe) datasets.
Stap 7: het verzamelen van de gegevens, voor het maken van het model en het testen van de prestaties daarvan. Bijvoorbeeld met behulp van een neuraal netwerk als modelleertechniek (Berthold & Hand, 2007: 269). Vervolgens het valideren van het model door middel van K-voudige kruisvalidatie om de beste configuratie te kiezen voor de model parameters (Berthold & Hand, 2007: 58). Dus, het scheiden van de dataset in meerdere subsets en het testen van de prestaties van modelconfiguraties op elke verschillende subsets van de dataset.
Stap 8: visualiseer de resultaten en presenteer de dataklassen die relevant zijn om de hypothesen te beantwoorden.

Stap 9: verfijn op basis van inzichten uit de visualisaties de conclusies.
Extra stap D: geef feedback aan dataproviders over de kwaliteit van gebruikte gegevens.
Stap 10: deel de resultaten met anderen voor feedback.

3 Ontwikkelaarpad

Extra stap A: begin met het definiëren van de onderzoeksvragen.
Stap 1: zoek naar datasets of API/URL's op data.overheid.nl portaal en daarbuiten eventueel
Stap 2: verwerven of downloaden met een geautomatiseerd proces of het opslaan van de gegevens in een database. Dit moet een database zijn met een vooraf gedefinieerde structuur voor tijdreeksgegevens.
Stap 3: met behulp van machine-learning, in termen van clusteranalyses, om subgroepen met gedeelde kenmerken binnen de data te ontdekken (James et al., 2013: 386).
Extra stap B: interpreteer de resultaten in het licht van verder relevant onderzoek.
Stap 4: onderzoeksvragen en/of hypothesen definiëren op de ontdekte clusters.
Extra stap C: vertaal de onderzoeksvragen en hypothesen in een gegevensmodel dat alle vragen kan beantwoorden. Dit model representeert de waarden van de dataset die gebruikt is voor produceren van het model. Bijvoorbeeld om voorspellingen te doen over toekomstige ontwikkelingen door hypotheses te toetsen met behulp van geproduceerde voorspellingsmodellen.
Stap 5: vooraf verwerken van de gegevens, koppelen aan andere databronnen voor het verrijken van de gegevens, en in staat zijn om relevant onderzoek te doen.
Stap 6: gebruik de resultaten voor het voorspellen van trends. Er zijn meerdere staten en technieken die kunnen worden gebruikt om het gedrag te modelleren. De toestanden zijn als volgt (Ragsdale, 2008: 486):

- Stationair, zonder trends in de tijd
- Niet-stationair, inclusief opwaartse en neerwaartse trends in de gegevens
- Seizoensgebonden, inclusief patronen in zowel stationaire als niet-stationaire tijdreeksgegevens

Stap 7: ontleedt de gegevens voor kwaliteitsbeoordeling om onderzoeksvragen te beantwoorden en/of hypothesen te bevestigen of te ontkrachten. Als dit niet acceptabel is, herhaalt u de vorige stap.
Extra stap D: definieer de criteria die u wilt gebruiken voor de beoordeling van de gegevenskwaliteit, dit hangt af van wat voor soort onderzoeksvragen u wilt beantwoorden en de hypothesen die u wilt bevestigen of juist ontkrachten. Ga vervolgens door naar de kwaliteitsbeoordeling om deze te verbeteren door feedback te geven aan dataproviders.
Stap 8: om een goed voorspellingsmodel te maken, kunnen de volgende technieken worden gebruikt (Ragsdale, 2008):

- Voortschrijdend gemiddelde, voor het markeren van langetermijntrend over kortetermijn willekeurige fluctuaties,
- Gewogen voortschrijdend gemiddelde, met gewichten
- Exponentiële afvlakking voor exponentieel afnemende gewichten in de loop van de tijd met behulp van exponentiële functies, die het gegevensgedrag in de loop van de tijd representeren
- Holt's-Winters' Methode voor additieve seizoensgebonden effecten, dit is een prognosemodel dat bestaat uit een voorspellingsvergelijking, samen met drie additieve smoothing-vergelijkingen. Eén voor het niveau, één voor de trend en een seizoen component.

Stap 9: ontleedt de gegevens voor kwaliteitsbeoordeling, indien aanvaardbaar ga door. Anders herhaling van vorige stap.
Stap 10: pas de fitheidstest toe voor het beoordelen van prestaties van modelvoorspelling. Bijvoorbeeld met behulp van de volgende technieken (Ragsdale, 2008: 487)

- Mediaan absolute afwijking, om de variabiliteit van univariate steekproef van kwantitatieve dataset te meten
- Gemiddeld absolute percentuele fout, om de nauwkeurigheid van modellen in percentages te meten
- Gemiddelde kwadratische fout, om het gemiddelde gekwadrateerde verschil tussen geschatte waarden en waarnemingen in de gegevens te meten
- Standaardafwijking, om de voorspellingsfouten voor verschillende modellen te vergelijken op dezelfde dataset.

Extra stap E: kies het juiste model evaluatiemethode die past bij de modelleer techniek.
Stap 11: de resultaten weergeven in de vorm van visualisaties.
Stap 12: de vorige stappen verfijnen, volgens inzichten uit de visualisaties. Als dit acceptabel is, bevestig dan de onderzoeksresultaten door de onderzoeksvragen te beantwoorden en/of hypothesen te bevestigen/ontkrachten.
Stap 13: deel de onderzoeksresultaten in de vorm van een onderzoeksrapport.

4 Journalistpad

Extra stap A: doe een kwaliteitsbeoordeling van mogelijke gegevens.
Stap 1: kies een dataset op basis van een interessant thema dat de voorkeur heeft.
Stap 2: verkrijg de dataset handmatig door deze naar de harde schijf van een lokale computer te downloaden.
Extra stap B: geef feedback aan gegevensproviders over de gekozen gegevens.
Extra stap C: deel data kwaliteitsbeoordeling met andere gebruikers, voor feedback om het te combineren met andere gegevens om die gegevenskwaliteit te verbeteren. Hierbij is het belangrijk dat koppelingen tussen gegevens waar mogelijk worden herkend en toegevoegd. De identificatienamen van de kolommen zijn handige informatie bij het maken van keuzes voor de koppeling van verschillende datasets.
Stap 3: filter relevante subsets en link naar andere gegevens van externe bronnen. Selecteer waar nodig reeksen van waarden voor verschillende gegevensattributen.
Stap 4: clusteranalyse toepassen op de dataset, voor het ontdekken van groepen met gedeelde kenmerken (James et al., 2013: 386).
Stap 5: definieer hypothesen volgens gedefinieerde clusters.
Stap 6: analyseer de dataset, mogelijk met statistische gevolgtrekking of beschrijvende statistieken, afhankelijk van de hypotheses.
Stap 7: test de prestaties van het model door middel van fitheidstest (Ragsdale, 2008). Deze meting kan worden gebruikt voor het testen van statistische hypotheses, bijvoorbeeld of twee monsters worden getrokken uit dezelfde verdeling die de dataset vertegenwoordigt.
Stap 8: representatie van de resultaten in vorm van visualisaties voor inzichten.
Stap 9: de analyse volgens afgeleide inzichten uit visualisaties verfijnen totdat het acceptabel is.
Stap 10: deel inzichten in de vorm van een online beschikbaar onderzoeksrapport.