Auditdienst Rijk
Ministerie van Financiën

> Retouradres Postbus 20201 2500 EE Den Haag

VERTROUWELIJK

Ministerie van Economische Zaken en Klimaat
Aan de plaatsvervangend Secretaris-generaal,
mevr. drs. G.M. Keijzer-Baldé
Bezuidenhoutseweg 73
2594 AC Den Haag

Auditdienst Rijk

Korte Voorhout 7
2511 CW Den Haag
Postbus 20201
2500 EE Den Haag
www.rijksoverheid.nl

Inlichtingen

[redacted]

T 06-[redacted]
F 070-[redacted]
[redacted]@minfin.nl

Ons kenmerk
2021-0000269220

Uw brief (kenmerk)
-

Bijlagen
1

Datum 28 december 2020
Betreft Eindrapportage ADR onderzoek

Geachte mevrouw Keijzer-Baldé,

Bij deze doen wij u als opdrachtgever het eindrapport toekomen van het onderzoek "vastleggingen binnen project PEGA" met ons kenmerk 2021-0000269209 Onderzoek vastleggingen binnen project PEGA. De conceptversie is op vrijdag 17 december jl. besproken met u.

U heeft van de mogelijkheid om een managementreactie te geven en deze op te laten nemen in de rapportage gebruik gemaakt. We hebben u daartoe gelegenheid gegeven om een reactie op te stellen. De managementreactie hebben wij onverkort in het bijgevoegde eindrapport opgenomen.

Hopende u hiermede van dienst te zijn geweest. Bedankt voor het door u in ons gestelde vertrouwen.

Als Auditdienst Rijk hechten wij veel waarde aan opdrachtgevers tevredenheid. Daarom zijn wij benieuwd hoe u de toegevoegde waarde van het onderzoek en de betrokkenheid van het projectteam heeft ervaren middels onderstaande link naar een korte vragenlijst. Het invullen duurt circa twee minuten. Uw mening stellen wij zeer op prijs en zullen die vertrouwelijk behandelen. Onze dank dat u daar de tijd voor wil nemen.

Naar de vragenlijst

Mocht het klikken niet werken dan kan u deze link kopiëren naar uw internetbrowser:
https://www.adronderzoek.nl/index.php/162795?lang=nl

Met vriendelijke groet,

Auditdienst Rijk
Ministerie van Financiën

Onderzoeksrapport
Onderzoek vastleggingen binnen project PEGA

Definitief

Colofon

Titel	Onderzoek vastleggingen binnen project PEGA
Uitgebracht aan	pSG van EZK
Datum	28 december 2021
Kenmerk	2021-0000269209

Inlichtingen
Auditdienst Rijk
070-[redacted]

Inhoud

Aanleiding opdracht

De Tweede Kamer heeft op 5 maart 2019 met algemene stemmen de motie-Van der Lee c.s. aangenomen, die uitspreekt dat een parlementaire enquête naar de aardgaswinning in Groningen wenselijk is. Binnen het ministerie van Economische Zaken en Klimaat (EZK) is in het verlengde hiervan per oktober 2020 een projectgroep PEGA (team Parlementaire Enquête Groninger Aardgaswinning) opgericht.

De parlementaire enquêtecommissie heeft haar onderzoeksvoorstel[1] op 4 februari 2021 aangeboden aan de Tweede Kamer. Op 9 februari is deze enquêtecommissie naar de aardgasgaswinning in Groningen formeel ingesteld. Het doel van de enquête is waarheidsvinding en het verkrijgen van verklarend inzicht in de besluitvorming over de aardgaswinning, de schadeafhandeling en de fysieke versterking van de meest risicovolle woningen in Groningen. Dit maakt het voor de commissie mogelijk te komen tot oordeelsvorming over de gehele periode en lessen te trekken, om daarmee bij te dragen aan toekomstperspectief voor Groningen en de ontwikkeling van toekomstig beleid. Zij doet dit onder meer door informatie op te vragen bij betrokken partijen (zogenoemde vorderingen). Elke vordering bevat meerdere vragen die leiden tot dossiers, inclusief bescheiden ter onderbouwing, waarmee beoogd wordt de vragen van de commissie te beantwoorden.

De opdrachtgever, zijnde de plaatsvervangend Secretaris Generaal (pSG) EZK, wil graag de bevindingen van de ADR bij de wijze waarop, binnen het team PEGA, de dossiers tot stand zijn gekomen. Hiermee kan de opdrachtgever achteraf verantwoording afleggen over dit proces. Dit is belangrijk omdat de opdrachtgever ervoor heeft gekozen een apart PEGA-team de werkzaamheden te laten uitvoeren, niet zijnde de beleidsmedewerkers die (direct) betrokken waren bij het dossier rond de gaswinning).

Het doel van de opzet van een apart team was tweeledig: 1. een zo objectief mogelijke samenstelling van de geselecteerde thema's en van de dossiers te krijgen, waarin de reconstructies en daaraan ten grondslag liggende documenten zijn opgenomen; en 2. de beleidskolom te ontzien, omdat de (voorbereidingen op) vorderingen van een parlementaire enquêtecommissie een zwaar tijdsbeslag leggen op de beschikbare capaciteit van beleidsmedewerkers. Dit leidde echter wel tot een dilemma bij team PEGA: namelijk het gelijktijdig kunnen waarborgen van de vakinhoudelijke kwaliteit en het kunnen waarborgen van een objectieve beoordeling.

Het door het PEGA-team ontworpen proces ter beantwoording van de vordering van de enquêtecommissie kent twee belangrijke handmatige stappen:

1. het opstellen van de zoekvragen, waarmee middels "Zoek en Vind" [2] de digitale bronnen worden doorzocht en suggesties/zoekresultaten gegenereerd worden, en;
2. het beoordelen welke van de suggesties/zoekresultaten relevant zijn en dus in het dossier worden opgenomen en welke niet.

De opdrachtgever wil deze bevindingen omdat zij graag een volledige en transparante beantwoording wil geven op de vragen van de parlementaire enquête. Mochten de bevindingen aanleiding geven tot extra relevante documenten dan kan zij deze nog naleveren aan de parlementaire enquête. Daarnaast vormen deze bevindingen input voor leerpunten met betrekking tot eventuele volgende parlementaire enquêtes.

---

[1] onderzoeksvoorstel voor een parlementaire enquete over de aardgaswinning in groningen

[2] Zoek & Vind is de (generieke) zoekfunctionaliteit welke gebruikt wordt door EZK

De ADR heeft de doelstelling van de opdrachtgever vertaald in twee onderzoeksvragen:

1. Welke bevindingen heeft de ADR bij de opgestelde zoekvragen?
   Toelichting: voor een volledig[3] dossier is het van belang dat de bronnen zelf doorzoekbaar zijn (dit valt buiten de scope van dit onderzoek) en is het van belang dat alle relevante zoektermen gebruikt worden. Anders gezegd: zijn de gekozen/gehanteerde zoektermen afdoende om alle potentieel relevante suggesties te genereren? Immers niet gevonden door "Zoek en Vind" levert een groot risico op dat documenten onterecht niet aanwezig zijn in het dossier.
2. Welke bevindingen heeft de ADR bij de, bij beantwoording van de vragen van de commissie gehanteerde, beoordeling van de suggesties door de projectgroep PEGA?
   Toelichting: Voor een volledig[3] dossier zijn de volgende aspecten van belang:
   1 het selecteren of juist niet selecteren van documenten uit de zoekresultaten van "Zoek en Vind",
   2 de beoordelingen/overwegingen die de betrokken medewerkers hebben gemaakt en die uiteindelijk hebben geleid tot de inhoud van de dossiers.

Bij de start van het onderzoek in maart 2021 bleek dat team PEGA onder tijdsdruk stond, veroorzaakt door de beantwoording (met beperkte mankracht) van de ruim 70 vorderingen van de enquêtecommissie. Vervolgens is in overleg met de opdrachtgever gekozen voor een andere aanpak (inzet data-analyse) van het onderzoek waardoor de belasting van het PEGA-team minimaal zou zijn. Het nadeel van deze aanpak is dat de ADR niet rechtstreeks naar de zoekvragen en de selectie en beoordeling heeft kunnen kijken, en alleen gekeken heeft naar de uiteindelijk gevonden documenten op basis van de ADR–zoekvragen. De ADR heeft vervolgens een aantal documenten opnieuw ter beoordeling voorgelegd aan het PEGA-team. Het voordeel van deze aanpak is dat hiermee het dossier van het PEGA-team kon worden aangevuld.

In hoofdstuk 1 gaan wij kort in op de gekozen aanpak van de ADR om de onderzoeksvragen te beantwoorden. In hoofdstuk 2 geven wij antwoord op onderzoeksvraag 1 waarna wij in hoofdstuk 3 onderzoeksvraag 2 behandelen. In hoofdstuk 4 gaan we in op de uitkomsten van de aanvullende werkzaamheden, waarna in hoofdstuk 5 de aanbevelingen volgen. In hoofdstuk 6 staat de verantwoording van het onderzoek en in hoofdstuk 7 de ondertekening van dit rapport.

[3] Volledig in de zin van gevuld met alle relevante documenten

Samenvatting

De Tweede Kamer heeft op 5 maart 2019 met algemene stemmen de motie-Van der Lee c.s. aangenomen, die uitspreekt dat een parlementaire enquête naar de aardgaswinning in Groningen wenselijk is. Binnen het ministerie van Economische Zaken en Klimaat (EZK) is in het verlengde hiervan per oktober 2020 een PEGA-team (team Parlementaire Enquête Groninger Aardgaswinning) opgericht, dat is belast met het verzamelen van de informatie voor de enquêtecommissie. Omdat het ministerie van EZK zich wil kunnen verantwoorden voor de totstandkoming en de inhoud van het dossier voor de enquêtecommissie, heeft het de ADR gevraagd hiernaar onderzoek te doen.

De ADR heeft, in afstemming met de opdrachtgever, besloten tot een aanpak die een minimale belasting legt op het team PEGA. Het PEGA-team stond namelijk onder tijdsdruk door de beantwoording van de 70 vorderingen van de enquêtecommissie. De ADR heeft daartoe zoekvragen gegenereerd op basis van publieke bronnen. In onderling overleg is besloten dat EZK de vorderingen van de enquêtecommissie niet aan de ADR zou verstrekken, mede gezien het vertrouwelijke karakter ervan. Ten behoeve van ons onderzoek naar de opgestelde zoekvragen van team PEGA, is het deel van de documenten dat gevonden is met de zoekvragen van de ADR en niet gevonden is met de zoekvragen van team PEGA voorgelegd aan medewerkers van PEGA om te beoordelen op relevantie voor de enquêtecommissie (zie voor details de verantwoording van het onderzoek in paragraaf 6.1.1).

De ADR heeft een aanzienlijk aantal documenten gevonden dat niet door het team PEGA was gevonden. Uit een nader onderzoek van 121 van deze documenten gaf het team PEGA aan dat 28 documenten (= 23%) relevant voor de enquêtecommissie waren (zie hoofdstuk 2 voor nadere uitwerking).

Uit ons onderzoek blijkt verder dat PEGA-medewerkers 99 documenten in eerste instantie als niet relevant beoordeelden, maar in een later stadium 29 hiervan alsnog als relevant bestempelden. De medewerkers van het PEGA-team gaven hiervoor als reden aan dat er sprake was van voortschrijdend inzicht. Een factor die volgens hen ook meespeelde is de ervaren tijdsdruk ten tijde van het opleveren van documenten (zie hoofdstuk 3 voor nadere uitwerking).

Uit aanvullend onderzoek naar de documenten die de ADR wel had gevonden, maar het team PEGA niet, bleek het volgende:

- Het team PEGA heeft aanvullend 494 documenten (uit de meest kansrijke clusters) beoordeeld die de ADR had gevonden, maar het PEGA-team niet. De uitkomst was dat het PEGA-team 119 documenten (=24%) daarvan relevant voor aanlevering aanmerkte[4].
- Het team PEGA heeft veel documenten die de ADR had gevonden als niet relevant aangemerkt voor de enquêtecommissie, omdat deze niet binnen de scope van de vordering van de enquêtecommissie lagen of niet afkomstig waren van het niveau van het Directieteam of hoger. Ook kon het zijn dat de status van deze documenten niet vast te stellen is.
- De beoordeling van documenten kost veel tijd van het team PEGA en de experimentele aanpak van de ADR heeft ervoor gezorgd dat het onderzoek uitliep. Daarbij bleek het niet mogelijk tijdig een andere manier te vinden voor de beoordeling van de stukken die de ADR wel had gevonden, maar het team PEGA niet (zie hoofdstuk 4 voor nadere uitwerking).

De bestuursraad van EZK (BR) heeft daarom besloten het onderzoek van de ADR af te laten ronden en geen nadere analyse meer te laten uitvoeren naar de resterende documenten die de ADR wel, en het team PEGA niet had gevonden. De contactpersoon heeft over het besluit van de Bestuursraad het volgende teruggekoppeld:

"De Bestuursraad van EZK zag zich, mede als gevolg van het feit dat het onderzoek niet volgens planning is verlopen, geconfronteerd met het dilemma dat de ADR heeft vastgesteld dat er zich in een totale voorraad van circa 137.000 documenten mogelijk nog niet-gevonden relevante documenten bevinden, maar dat de ADR en EZK er samen niet in geslaagd zijn om de uitkomsten van deze data-analyse op een gevalideerde en praktisch uitvoerbare wijze te verbinden aan de concrete vorderingen. Er was dus onvoldoende zicht op de daadwerkelijke hoeveelheid relevante documenten. Voortzetting van het zoeken naar die concrete documenten zou een zwaar beperkend effect hebben op de capaciteit van het PEGA-team in een periode dat alle capaciteit moet worden ingezet voor de voorbereiding van de EZK-genodigden op de besloten voorgesprekken vanaf januari 2022. In het besluit om het ADR-onderzoek af te ronden is naast het capaciteitsvraagstuk ook meegewogen dat onzeker is hoeveel tijd nodig is om de bewuste documenten alsnog te vinden en dat in deze fase van het onderzoek van de commissie – nl. na afrondingen van het documentenonderzoek – eventuele aanvullende documenten überhaupt steeds minder toegevoegde waarde hebben. Wel heeft de Bestuursraad aangegeven opnieuw met dit materiaal aan de slag te willen gaan als er aanvullende, gerichte vorderingen vanuit de commissie worden ontvangen."

[4] Omdat het geen steekproef betreft, maar een deelwaarneming van de meest kansrijke clusters, is extrapolatie niet mogelijk en kan dus geen uitspraak worden gedaan over de gehele massa, maar geeft het een indicatie.

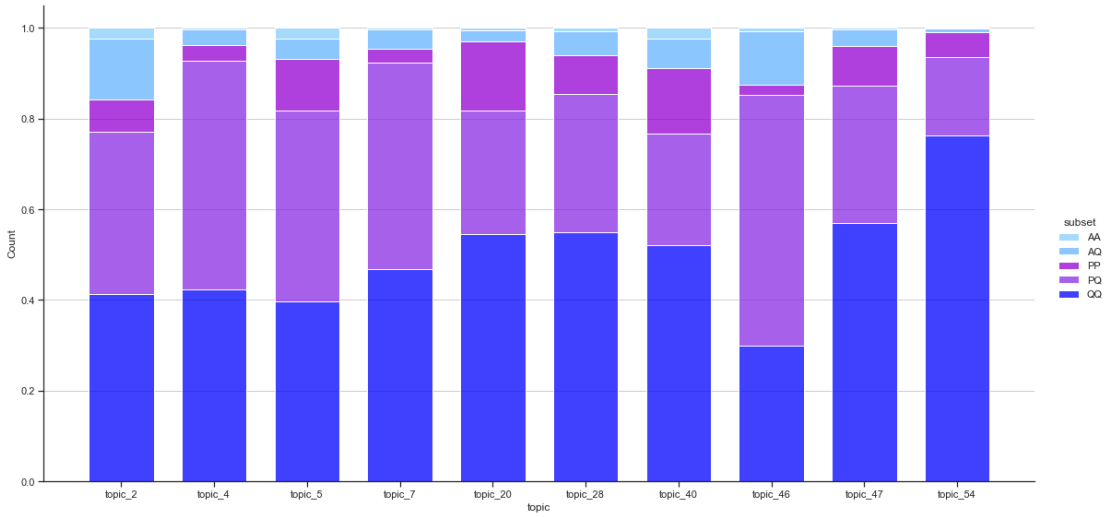
1 Gekozen (experimentele) aanpak ADR

In dit hoofdstuk belichten we kort de gevolgde aanpak, zodat de lezer de uitkomsten in de navolgende hoofdstukken kan volgen.
Bij de start van het onderzoek in maart 2021, bleek dat team PEGA onder tijdsdruk stond, veroorzaakt door de beantwoording (met beperkte mankracht) van de ruim 70 vorderingen van de enquêtecommissie. De ADR heeft, in afstemming met de opdrachtgever, besloten tot een (experimentele) aanpak die een minimale belasting legde op het team PEGA. Een nadeel van de gekozen aanpak was dat we niet rechtstreeks naar de zoekvragen (deelvraag 1) en de selectie en beoordeling (deelvraag 2) hebben gekeken, maar alleen naar de gevonden documenten op basis van de ADR zoekvragen. Vervolgens hebben we een aantal documenten opnieuw ter beoordeling voorgelegd. Een voordeel was dat deze aanpak voor het dossier relevante aanvullende documenten kan opleveren. Het onderzoek heeft echter wel een langere doorlooptijd gekend dan gepland omdat dit de eerste keer was dat de ADR deze aanpak heeft toepast in een onderzoek.

De ADR heeft zoekvragen gegenereerd op basis van publieke bronnen. (Zie voor details de verantwoording van het onderzoek in paragraaf 6.1.1.) In onderling overleg is besloten dat EZK de vorderingen van de enquêtecommissie niet aan de ADR zou verstrekken, mede gezien het vertrouwelijke karakter ervan. Om toch te kunnen reflecteren op de opgestelde zoekvragen van het team PEGA, zijn de documenten die gevonden zijn met de zoekvragen van de ADR en niet met de zoekvragen van team PEGA voorgelegd aan de PEGA-medewerkers ter beoordeling op relevantie.

In totaal heeft de ADR 1.596.851 documenten gevonden met de zoekvragen gebaseerd op zoektermen uit publieke bronnen (voornamelijk krantenartikelen). De ADR kan geen 1.6 mln. documenten voorleggen aan het team PEGA om te beoordelen op relevantie. Daarom hebben we opnieuw de techniek “topic modeling” (zie paragraaf 6.1.2 voor uitleg over topic modeling) toegepast op de gehele massa om de documenten in onderwerpen te verdelen op basis van de woorden die in de documenten voorkomen en hebben we samen met het team PEGA relevante “topics” bepaald. Hieruit kwamen 10 relevante topics naar voren.

In de onderstaand figuur zijn deze topics met de genormaliseerde aantallen documenten weergegeven. De betekenis van de subsets is als volgt:
AQ: Aanlevering en Query (aangeleverd en gevonden door de ADR zoekvraag)
AA: Alleen in Aanlevering (dus wel aangeleverd, niet gevonden door de ADR zoekvraag)
PP: Prullenbak en niet in de ADR zoekvraag
PQ: Prullenbak en wel in de ADR zoekvraag
QQ: Alleen in Query (niet in prullenbak/aanlevering alleen gevonden door de ADR zoekvraag)



Figuur 1: overzicht genormaliseerd aantal gevonden subsets per topic. A staat voor aanlevering PEGA, P staat voor prullenbak PEGA, en Q staat voor de zoekvraag ADR.

Figuur 1 is per relevant topic uitgesplitst naar drie categorieën, zijnde documenten die door PEGA reeds zijn aangeleverd (**lichtblauw, AA+AQ**), documenten die door PEGA zijn gevonden maar als niet relevant zijn geacht (**magenta, PP+PQ**) en documenten die niet door PEGA zijn gevonden (**donkerblauw, QQ**). De y-as geeft niet het aantal documenten weer dat in een topic zit, maar de som van de gewogen waarschijnlijkheden. Deze laatste maat is zuiverder dan het rapporteren van aantallen documenten. Dat heeft er mee te maken dat het topic modeling algoritme een document aan elk van de zestig topics toekent met een bepaalde waarschijnlijkheid. Zo kan een document met 50% waarschijnlijkheid tot topic 1 behoren, 40% tot topic 2 en de overige 10% verdeeld over de overige topics. In dat geval tellen we dit document in bovenstaande figuur voor topic 2 niet als één document, maar als 0.40 document. De waarde op de y-as geeft daarmee dus de som van alle waarschijnlijkheden van de documenten die tot dat topic behoren. Zie paragraaf 6.1.2 voor een verdere uitwerking hiervan.

Uit elk topic zijn vervolgens, indien mogelijk, 30 documenten getrokken (15 uit de subset QQ en 15 uit de overige massa) en deze zijn voorgelegd aan team PEGA om te beoordelen op relevantie. Omdat het geen steekproef betreft is extrapolatie niet mogelijk en kunnen we dus geen uitspraak doen over de gehele massa, maar geeft het een indicatie.

In hoofdstuk 2 gaan wij in op de uitkomsten uit subset QQ waarmee we een antwoord geven op onderzoeksvraag 1. In hoofdstuk 3 gaan wij nader in op de uitkomsten van de documenten uit de overige massa waarmee we een antwoord geven op onderzoeksvraag 2.

De opdrachtgever wil zo volledig mogelijk aanleveren aan de enquêtecommissie en heeft naar aanleiding van de eerste onderzoeksresultaten besloten om een aanvulling te doen op de oorspronkelijke onderzoeksvraag. De vraag luidt: "In hoeverre zitten er, voor het beantwoorden van de vordering van de parlementaire enquêtecommissie nog mogelijk relevante documenten in de door de ADR gevonden documenten welke niet door team PEGA zijn gevonden?" In hoofdstuk 4 staan de uitkomsten voor de beantwoording van de aanvullende onderzoeksvraag.

2 28 relevante documenten gevonden door nieuwe databronnen en bijlagen in e-mails

In dit hoofdstuk gaan we in op de eerste onderzoeksvraag:

1. Welke bevindingen heeft de ADR bij de opgestelde zoekvragen?

Toelichting: Voor een volledig[3] dossier is het van belang dat de bronnen doorzoekbaar zijn (dit valt nu buiten scope) en is het van belang dat alle relevante zoektermen gebruikt worden. Anders gezegd: zijn de gekozen/gehanteerde zoektermen afdoende om alle potentieel relevante suggesties te genereren? Immers, niet gevonden door de zoekmachine "Zoek en Vind" levert een groot risico op dat documenten onterecht niet aanwezig zijn in het dossier.

2.1 Van 121 niet door PEGA gevonden documenten zijn er 28 als relevant voor aanlevering beoordeeld door PEGA

In interviews met de medewerkers van PEGA hebben we in totaal 144 documenten besproken, die wel door onze zoekvragen zijn gevonden, maar niet door de zoekvragen van PEGA. Hiervan zijn 21 documenten reeds in een andere vorm (bijvoorbeeld een identieke mail uit een andere mailbox) aangeleverd, en resteren 121 die niet aangeleverd zijn. Van deze 121 documenten, dus de documenten die alleen door de ADR en niet door PEGA zijn gevonden en die ook niet in een andere vorm zijn aangeleverd, hebben de medewerkers van PEGA in totaal 28 documenten als relevant voor aanlevering beoordeeld (= 23%).

Uitgesplitst per topic (zie aanpak in paragraaf 6.1) hebben wij met name voor topics 4 , 5, 40 en 46 relatief veel documenten gevonden die PEGA niet heeft gevonden en waarvan de medewerkers bij navraag hebben aangegeven dat deze wel relevant zijn voor levering.

	Totaal besproken documenten	Totaal niet aangeleverd	Relevant bevonden	Percentage (t.o.v. niet-aanlevering)
Topic 2	15	11	2	18%
Topic 4	15	14	6	43%
Topic 5	10	6	3	50%
Topic 7	15	13	1	8%
Topic 20	15	13	2	15%
Topic 28	15	3	0	0%
Topic 40	15	14	5	36%
Topic 46	15	10	7	70%
Topic 47	15	14	1	7%
Topic 54	14	14	1	7%
Totaal	144	121	28	23%

Tabel 1 – Besproken documenten niet-gevonden door PEGA per topic

2.2 Medewerkers PEGA geven twee mogelijke oorzaken voor niet vinden van documenten

In de gesprekken met de medewerkers van PEGA werd een tweetal redenen genoemd voor het mogelijk niet vinden van documenten:

1. Aansluiten van nieuwe databronnen nadat de zoekacties waren gestart (bijvoorbeeld: nieuwe mailbox beschikbaar in Zoek & Vind die niet opnieuw doorzocht is voor oudere vorderingen.)
2. Er is wel in mails gezocht, maar als deze geen "hit" gaven op de ingegeven zoekwoorden, zijn de bijlagen bij die mails niet doorzocht. Er is dus alleen gescand op woorden in de mail en in de naam van de bijlage, maar niet in de inhoud van de bijlage.

3 Voortschrijdend inzicht bij de beoordeling op relevantie

In dit hoofdstuk gaan we in op de tweede onderzoeksvraag:

2. Welke bevindingen heeft de ADR bij de, bij beantwoording van de vragen van de commissie gehanteerde, beoordeling van de suggesties door de projectgroep PEGA?

Toelichting: Voor een volledig[3] dossier zijn de volgende aspecten van belang:
1 het selecteren of juist niet selecteren van documenten uit de zoekresultaten van "Zoek en Vind",
2 de beoordelingen/overwegingen die de betrokken medewerkers hebben gemaakt en die uiteindelijk hebben geleid tot de inhoud van de dossiers.

3.1 Van 99 documenten die in de "prullenbak" zaten zijn 29 als relevant beoordeeld door PEGA medewerkers

We hebben in totaal 144 documenten besproken die door PEGA zijn gevonden, maar die niet zijn aangeleverd aan de commissie (deze zijn dus in de prullenbak gezet). Van deze 144 documenten zijn er 45 documenten reeds in een andere vorm aangeleverd, dus in onze waarneming betreft de totale massa van niet-aangeleverde documenten 99 stukken. Van deze 99 niet-aangeleverde (dus "pure prullenbak") documenten zijn er in totaal 29 alsnog relevant bevonden (= 29%). Zoals uit de percentages in de onderstaande tabel blijkt, kwam dit met name voor in topics 4, 5, 40, 46 en 47.

	Totaal besproken documenten	Totaal niet aangeleverd	Relevant bevonden	Percentage (t.o.v. niet-aanlevering)
Topic 2	9	6	1	17%
Topic 4	14	14	4	29%
Topic 5	19	3	3	100%
Topic 7	15	12	3	25%
Topic 20	15	15	2	13%
Topic 28	13	7	1	14%
Topic 40	13	12	6	50%
Topic 46	15	6	3	50%
Topic 47	15	11	5	45%
Topic 54	16	13	1	8%
Totaal	144	99	29	29%

Tabel 2 – Besproken documenten uit de prullenbak per topic

3.2 PEGA medewerkers geven twee mogelijke oorzaken voor het anders beoordelen van de relevantie van een document

In de gesprekken met de experts werden als mogelijke redenen genoemd voor het anders beoordelen van de relevantie van een document:

1. Voortschrijdend inzicht over wat wel en niet relevant is. Door het langer met de materie bezig zijn is de kennis bij de PEGA medewerker gegroeid.
2. Tijdsdruk waaronder de oplevering van de vorderingen heeft plaatsgevonden.

Het anders beoordelen van documenten door voortschrijdend inzicht maakt het achteraf verantwoording afleggen over gemaakte keuzen niet eenduidig.

4 Onvoldoende zicht op aantal relevante documenten in de door de ADR gevonden documenten die niet door team PEGA zijn gevonden

In dit hoofdstuk gaan we in op de aanvullende onderzoeksvraag:
In hoeverre zitten er, voor het beantwoorden van de vordering van de parlementaire enquêtecommissie nog mogelijk relevante documenten in de door de ADR gevonden documenten welke niet door team PEGA zijn gevonden?
Toelichting: De ADR heeft door middel van data-analyse een aanzienlijk aantal documenten gevonden dat niet door team PEGA is gevonden. De vraag is echter of dit inhoudelijk relevante documenten voor de vordering van de enquêtecommissie zijn. De opdrachtgever wil zo volledig mogelijk aanleveren aan de enquêtecommissie en heeft daarom besloten een aanvulling te doen op de oorspronkelijke onderzoeksvraag.

Voor het beantwoorden van de aanvullende onderzoeksvraag heeft de ADR twee aanvullende stappen genomen. De eerste stap is het terugbrengen van het aantal documenten in de massa. Na deze stap resteerde een set van 137.000 documenten. In de tweede stap heeft de ADR deze set geclusterd om zo gelijksoortige documenten (met een gelijksoortige distributie over de topics) te bundelen en daarmee de beoordeling door het team PEGA te vergemakkelijken. Voor details zie paragraaf 6.1.4. In de eerste paragraaf gaan we nader in op de beoordeling van de clusters door het team PEGA, waarna we in de tweede paragraaf ingaan op de achtergrond voor het afronden van de opdracht.

4.1 Van 494 niet door PEGA gevonden documenten zijn 119 als relevant voor aanlevering beoordeeld door PEGA

In de eerste aanlevering zijn aan team PEGA 50 clusters ter beoordeling aangeleverd om de instellingen van de clustering te verifiëren. Deze aanlevering bevatte 167 documenten waarvan er 14 door team PEGA als relevant voor de vordering werden aangemerkt. Op basis van de ervaringen van de eerste 50 clusters die aangeleverd zijn heeft de ADR een nieuwe indicator voor relevantie gehanteerd voor de sortering van de meest kansrijke clusters en heeft op basis hiervan vervolgens 100 clusters met in totaal 327 documenten ter beoordeling aan team PEGA aangeleverd. De reeds beoordeelde documenten uit de eerste ronde zijn niet opnieuw aangeleverd.

Van deze 327 documenten zijn er 105 als relevant voor aanlevering aan de enquêtecommissie beoordeeld door het team PEGA. Het blijkt dat de nieuwe verdeling op basis van de aangepaste metrieken tot een beter resultaat leidt (van 8% naar 32% relevant).

Uit de beoordeling van de aangeleverde clusters door team PEGA blijkt dat 119 (14+105) van 494 documenten door team PEGA als relevant zijn beoordeeld en dat er 375 (494-119) van de 494 documenten niet relevant zijn voor de enquêtecommissie.

De aangeleverde clusters waren geen steekproef, maar een deelwaarneming van de meest kansrijke clusters, daarom is extrapolatie niet mogelijk en kan dus geen uitspraak worden gedaan over de gehele massa.

4.2 Onvoldoende zicht op aantal relevante documenten omdat onderzoek gestopt is vanwege tijdsdruk

Uit de werkzaamheden van de ADR blijkt dat er heel veel documenten door de ADR zijn gevonden die wellicht inhoudelijk over het onderwerp gaan maar toch zogenoemde "fout positieven" betroffen. Dit zijn documenten die hetzij:

- niet binnen de scope van de vordering van de enquêtecommissie liggen;
- van een medewerker lager dan Directieteam niveau afkomstig zijn;
- een niet vast te stellen status hebben,
- danwel een combinatie van deze eigenschappen hebben

In het geval van een niet vast te stellen status betreft het "kladstukken" van beleidsmedewerkers die input vormen voor het opstellen van nota's of het beantwoorden van Kamervragen. Dat de ADR deze laatste gevonden heeft komt doordat de ADR de bron "netwerkschijf" integraal meegenomen heeft in de data-analyse.

Een andere belangrijke oorzaak voor "fout positieven" is dat, zoals hierboven al aangegeven, in onderling overleg is besloten dat EZK de vorderingen van de enquêtecommissie niet aan de ADR zou verstrekken, mede gezien het vertrouwelijke karakter ervan. (zie aanpak 6.1.1) Bij het opstellen van de initiële zoekvraag op basis van openbare bronnen is daarom gekozen voor een zoekvraag die relevant is voor het "Groningendossier". Daarmee is er bij de beoordeling van relevantie van een door de ADR gevonden document nog een aanvullende vertaalslag naar de specifieke vorderingsvragen nodig.

Uiteindelijk heeft de ADR echter onvoldoende zicht gekregen op het totale aantal relevante documenten omdat de totale doorlooptijd van het onderzoek vanwege het experimentele karakter flink uitgelopen is in de tijd. In de oorspronkelijke planning (in maart 2021) was juni 2021 voorzien als opleverdatum van het rapport. De uitloop heeft er mede voor gezorgd dat het aanvullende onderzoek over de extra vraag in zeer kort tijdsbestek moest worden uitgevoerd, omdat in januari 2022 de enquêtecommissie start met de verhoren. Ondanks intensieve werksessies met het team PEGA en de mensen van Zoek en Vind is het niet gelukt om deze extra onderzoeksvraag te beantwoorden. De beoordeling van clusters vormt voor het team PEGA naar verluidt een te grote belasting. Daarnaast was de nog beschikbare tijd beperkt en waren er heel veel "fout positieve" uitkomsten in de data. Om deze redenen heeft de Bestuursraad van het ministerie van EZK besloten om het onderzoek af te laten ronden. De contactpersoon heeft over het besluit van de bestuursraad het volgende teruggekoppeld:

"De Bestuursraad van EZK zag zich, mede als gevolg van het feit dat het onderzoek niet volgens planning is verlopen, geconfronteerd met het dilemma dat de ADR heeft vastgesteld dat er zich in een totale voorraad van circa 137.000 documenten mogelijk nog niet-gevonden relevante documenten bevinden, maar dat de ADR en EZK er samen niet in geslaagd zijn om de uitkomsten van deze data-analyse op een gevalideerde en praktisch uitvoerbare wijze te verbinden aan de concrete vorderingen. Er was dus onvoldoende zicht op de daadwerkelijke hoeveelheid relevante documenten. Voortzetting van het zoeken naar die concrete documenten zou een zwaar beperkend effect hebben op de capaciteit van het PEGA-team in een periode dat alle capaciteit moet worden ingezet voor de voorbereiding van de EZK-genodigden op de besloten voorgesprekken vanaf januari 2022. In het besluit om het ADR-onderzoek af te ronden is naast het capaciteitsvraagstuk ook meegewogen dat onzeker is hoeveel tijd nodig is om de bewuste documenten alsnog te vinden en dat in deze fase van het onderzoek van de commissie – nl. na afrondingen van het documentenonderzoek – eventuele aanvullende documenten überhaupt steeds minder toegevoegde waarde hebben. Wel heeft de Bestuursraad aangegeven opnieuw met dit materiaal aan de slag te willen gaan als er aanvullende, gerichte vorderingen vanuit de commissie worden ontvangen."

5 Aanbevelingen en/of vervolgstappen

In bespreking met team PEGA is de data-analyse techniek “topic modeling”, die de ADR toepaste in haar onderzoek, als waardevol voor toekomstige enquêtecommissies of WOB verzoeken aangemerkt. De techniek zou gebruikt kunnen worden om op voorhand in te zetten voor een eerste grove filtering, waardoor de inzet van medewerkers efficiënter kan plaats vinden. Inmiddels zijn er ook contacten gelegd, middels dit onderzoek, bij Zoek en Vind waar men kijkt of men de techniek kan implementeren in het pakket.

We onderschrijven de intentie van de Bestuursraad welke heeft aangegeven opnieuw met het materiaal van de ADR aan de slag te willen gaan als er aanvullende, gerichte vorderingen vanuit de commissie worden ontvangen.

6 Verantwoording onderzoek

6.1 Werkzaamheden en afbakening

De opdrachtgever wilde graag de bevindingen van de ADR bij de wijze waarop, binnen het team PEGA, de dossiers tot stand zijn gekomen. Elke vordering bevat meerdere vragen welke leiden tot dossiers, inclusief bescheiden ter onderbouwing, waarmee de vragen van de commissie beantwoord worden. De opdrachtgever wil deze bevindingen omdat zij graag een volledige en transparante beantwoording wil geven op de vragen van de parlementaire enquête. Mochten de bevindingen aanleiding geven tot extra relevante documenten dan kan zij deze nog naleveren aan de parlementaire enquête. Daarnaast vormen deze bevindingen input voor leerpunten met betrekking tot eventuele volgende parlementaire enquêtes.

Deze doelstelling van het onderzoek is gerealiseerd door het beantwoorden van de volgende onderzoeksvragen:

1. Welke bevindingen heeft de ADR bij de opgestelde zoekvragen?
   Toelichting: voor een volledig[3] dossier is het van belang dat de bronnen doorzoekbaar zijn (dit valt nu buiten scope) en is het van belang dat alle relevante zoektermen gebruikt worden. Anders gezegd: zijn de gekozen/gehanteerde zoektermen afdoende om alle potentieel relevante suggesties te genereren? Immers niet gevonden door "Zoek en Vind" levert een groot risico op dat documenten onterecht niet aanwezig zijn in het dossier.
2. Welke bevindingen heeft de ADR bij de, bij beantwoording van de vragen van de commissie gehanteerde, beoordeling van de suggesties door de projectgroep PEGA?
   Toelichting: Voor een volledig[3] dossier zijn de volgende aspecten van belang:
   1 het selecteren of juist niet selecteren van documenten uit de zoekresultaten van "Zoek en Vind",
   2 de beoordelingen/overwegingen die de betrokken medewerkers hebben gemaakt en welke uiteindelijk hebben geleid tot de inhoud van de dossiers.

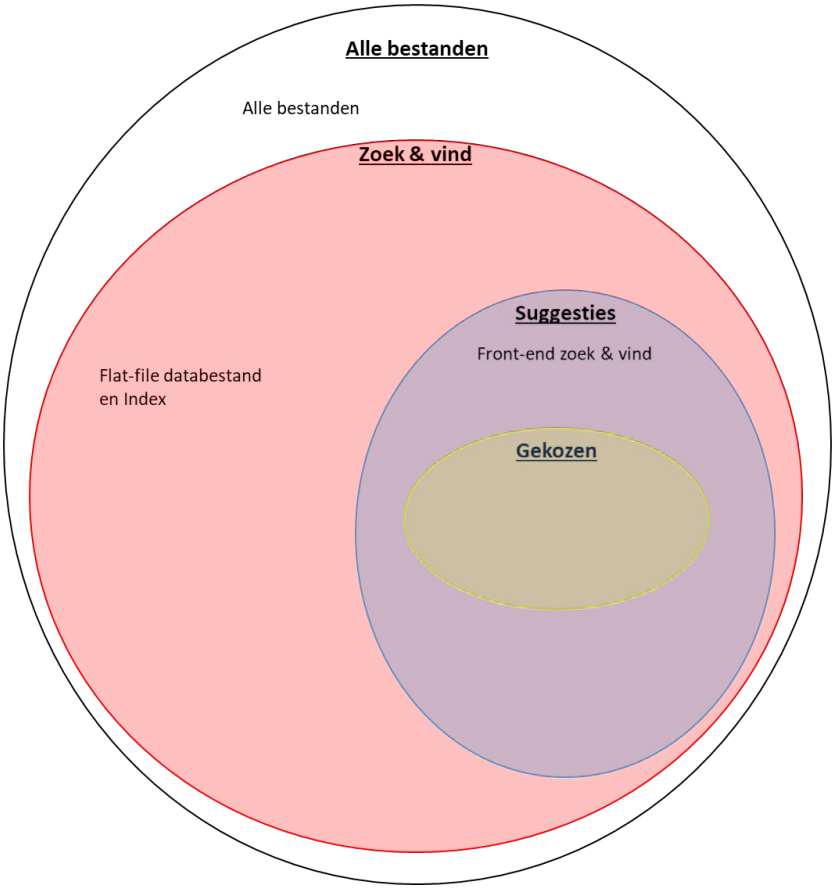
Naar aanleiding van de bevindingen uit de eerste twee onderzoeksvragen is er een aanvullende onderzoeksvraag geformuleerd:

"*In hoeverre zitten er, voor het beantwoorden van de vordering van de parlementaire enquêtecommissie nog mogelijk relevante documenten in de door de ADR gevonden documenten welke niet door team PEGA zijn gevonden?*"

Op basis van publieke bronnen, machine-learning modellen, en interviews met domeinexperts is de aard van de aanlevering in kaart gebracht. De werkzaamheden zijn uitgevoerd tussen 01-03-2021 en 01-12-2021. In paragraaf 6.1.1 wordt de (experimentele) aanpak in zijn algemeenheid toegelicht. In paragraaf 6.1.2 wordt de gehanteerde werkwijze van onderzoeksvraag 1 toegelicht. In paragraaf 6.1.3 wordt de gehanteerde werkwijze van onderzoeksvraag 2 toegelicht. In paragraaf 6.1.4 gaan we nader in op de gehanteerde werkwijze rond de aanvullende onderzoeksvraag.

De totale doorlooptijd van het onderzoek is vanwege het experimentele karakter flink uitgelopen in de tijd. In de oorspronkelijke planning in maart 2021 was juni 2021 voorzien als opleverdatum van het rapport. De uitloop heeft er mede voor gezorgd dat de aanvullende onderzoeksvraag in zeer kort tijdsbestek moest worden uitgevoerd. Ondanks intensieve werksessies met team PEGA en de mensen van Zoek en Vind is het niet gelukt om de aanvullende onderzoeksvraag te beantwoorden.

6.1.1 *Experimentele aanpak op hoofdlijnen*



Figuur 2 Documenten binnen bronnen.

Legenda Figuur 2:

Wit	Alle bestanden: alle bestanden uit alle bronnen
Rood	Zoek en Vind: alle geïndexeerde bestanden door Zoek en Vind voor heel EZK en LNV
Blauw	Suggesties: alle suggesties op basis van de zoekvragen team PEGA
Geel	Gekozen: alle als relevant beoordeelde suggesties op basis van de vordering van de commissie.

Het onderzoek richt zich voornamelijk op de rode, blauwe, en gele bollen in Figuur 2. In vraag 1 wordt gekeken naar de totstandkoming van de suggesties in "Zoek & Vind" (Z&V) op basis van de opgestelde zoekvragen (van rood naar blauw). In vraag 2 wordt gekeken naar de totstandkoming van de aanlevering die gedaan is op basis van de suggesties (van blauw naar geel). In de huidige opzet is buiten beschouwing gelaten in hoeverre Z&V alle bronnen en documenten kan benaderen en doorzoeken (van wit naar rood). Dit is gedaan omdat hiervoor toegang nodig is tot de productie-omgeving van Z&V en de onderliggende bronnen. Deze toegang is niet binnen de doorlooptijd van het onderzoek te realiseren.

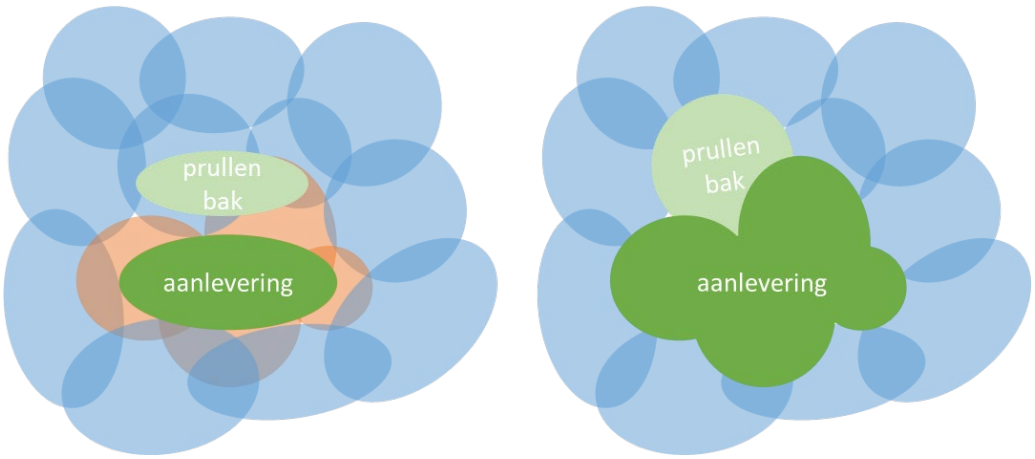
Een handmatige her-evaluatie van de aanpak door het team PEGA is niet alleen enorm tijdrovend, maar voegt bovendien niet direct waarde toe aan (het inzicht in) de aanlevering. Zo zijn er in Z&V grofweg 30,8 miljoen bestanden te vinden, waarvan ongeveer 19,3 miljoen van EZK. Bovendien hebben wij niet meer expertise in huis op het gebied van aardgaswinning dan de door PEGA geraadpleegde domeinexperts die betrokken zijn geweest. Daarom hebben we voor een experimentele opzet gekozen, waarbij we ons richten op *objectiviteit* en *volledigheid* door openbare bronnen en machine-learning modellen te gebruiken om de 19,3 miljoen documenten in Z&V te beoordelen op relevantie voor het

"Groningendossier". Daarmee is er bij de beoordeling van relevantie van een door de ADR gevonden document nog een aanvullende vertaalslag naar de specifieke vorderingsvragen nodig. Dit omdat in onderling overleg is besloten dat EZK de vorderingen van de enquêtecommissie niet aan de ADR zou verstrekken, mede gezien het vertrouwelijke karakter ervan.

Om inzicht te krijgen in de aanlevering door team PEGA, is besloten om het proces om te komen tot relevante documenten te herhalen, maar dan op basis van tekst-mining en machine-learning in plaats van domeinexpertise. Op basis van tekst-mining op openbare bronnen worden zoekvragen opgesteld die gebruikt worden om in Z&V naar documenten te zoeken, waarvan de resultaten vergeleken worden met die van team PEGA welke gezocht op basis van de vordering en kennis van domeinexperts. Op hoofdlijnen kan de experimentele opzet in vier stappen ingedeeld worden:

a. Het opstellen van zoekvragen om alle (mogelijk) relevante documenten in Zoek en Vind op te zoeken;
b. (mogelijk) relevante documenten in onderwerpen indelen en relevante onderwerpen definiëren;
c. in kaart brengen welke onderwerpen in welke mate door team PEGA gevonden zijn (vraag 1), en daarnaast welke onderwerpen in welke mate aangeleverd zijn (vraag 2);
d. voor elk relevant onderwerp representatieve documenten voorleggen om de onderwerpen op relevantie voor de vordering te beoordelen.

Op deze manier willen wij een inschatting maken in hoeverre team PEGA het rechter venn-diagram in Figuur 3 heeft weten te benaderen. Links is de situatie geschetst dat de aanleveringen niet alle documenten bevat met relevante onderwerpen. Rechts is de ideale situatie geschetst dat alle documenten met relevante onderwerpen aangeleverd zijn en de prullenbak alleen documenten met niet relevante onderwerpen bevat.



Figuur 3: venn-diagrammen van de documenten die in Z&V te vinden zijn, met links de daadwerkelijke situatie en rechts de ideale situatie. De blauwe bolletjes representeren de documenten met onderwerpen die voor de vorderingen niet relevant zijn, oranje zijn de relevante documenten, en groen representeert de door het PEGA team gevonden documenten (zoeksuggesties) met de onderverdeling in aangeleverde documenten (donker groen).

6.1.2 *Aanpak onderzoeksvraag 1*

Voor het beantwoorden van onderzoeksvraag 1 hebben we een indicatie nodig in hoeverre de blauwe bol alle relevante documenten in de rode bol bevat (Figuur 2). Daarvoor hebben we gebruik gemaakt van publieke bronnen – namelijk kamerstukken en krantenartikelen - om een set van zoekvragen op te stellen waarmee we in Z&V naar relevante documenten kunnen zoeken.[5] Vervolgens

[5] Voor een overzicht van de kranten waarvan artikelen verzameld zijn, zie appendix 1.

vergelijken we de resultaten die uit onze zoekvragen voortkomen met de resultaten die voortgekomen zijn uit de zoekvragen van team PEGA: welke documenten zitten in onze resultaten die niet in die van team PEGA zitten? De discrepantie tussen de twee sets aan resultaten geeft vervolgens inzicht in vraag 1 door de onderwerpen te belichten die mogelijk bij de totstandkoming van de zoekvragen van team PEGA over het hoofd gezien zijn.

Om krantenartikelen te verzamelen hebben we allereerst initiële zoekvragen op moeten stellen die we in een krantendatabase konden. Een deel van deze initiële zoekvragen zijn opgesteld aan de hand van de onderzoeksdoelstellingen zoals die geformuleerd zijn door de enquêtecommissie in haar rapportage, aangevuld met zoektermen die naar voren zijn gekomen door machine-learning technieken toe te passen op alle documenten in de tweede-kamer dossiers die te maken hebben met gaswinning in Groningen.[6] Vervolgens hebben wij met deze initiële zoekvragen naar krantenartikelen gezocht, en op de inhoud van deze artikelen *topic-modeling* toegepast. Omdat het onderzoek tot 1959 terugkijkt, hebben we ook een historische database aangesproken.[7] Door topic-modeling toe te passen op de krantenartikelen zijn wij tot zoekvragen gekomen die we in Z&V gebruiken om de documenten van EZK te doorzoeken. Onze zoekvragen leverden grofweg 2 miljoen documenten op.

Vervolgens zijn de resultaten die uit onze zoekvragen voortgekomen vergeleken met de zoekresultaten van team PEGA. Omdat het 2 miljoen documenten betrof, hebben we op deze zoekresultaten topic modeling toegepast om de documenten in onderwerpen te verdelen op basis van de woorden die in de documenten voorkomen. De relevante onderwerpen (‘topics’) zijn ieder apart voorgelegd en besproken met team PEGA door middel van een deelwaarneming en interviews.

In samenspraak met team PEGA zijn op vrijdag 10 september 2021 tien van de 60 besproken topics als relevant geclassificeerd (zie appendix 3 voor een overzicht) aan de hand van een lijst met de meest kenmerkende woorden van een topic. Voor elk van deze tien topics hebben we een algoritme dertig documenten gestratificeerd laten selecteren. In deze selectie is sprake geweest van een dubbele stratificatie:

- Gevonden door PEGA vs. niet gevonden door PEGA. Het algoritme heeft (waar mogelijk) per topic 15 documenten geselecteerd die door PEGA zijn gevonden en 15 documenten die niet door PEGA zijn gevonden, maar alleen in onze zoekvragen naar voren zijn gekomen.
- Waarschijnlijkheid. Het topic-modeling algoritme kent een document met een bepaalde waarschijnlijkheidsscore toe aan elk van de zestig topics. Voor alle documenten ontstaat er dan per topic een waarschijnlijkheidsbereik. We hebben per topic dit bereik in drie gelijke delen opgedeeld, en uit elke deel vijf documenten aselect getrokken.[8]

Op basis van de massa ‘niet gevonden door PEGA’ kunnen we iets zeggen over de zoekresultaten van PEGA, ofwel onderzoeksvraag 1. Op basis van de massa ‘wel gevonden door PEGA’ kunnen we iets zeggen over onderzoeksvraag 2 door een aantal door PEGA beoordeelde documenten ter herbeoordeling voor te leggen. Bij het aanleveren van de documentselectie aan team PEGA hebben we er bewust voor gekozen alleen ruwe tekst op te sturen, en metadata achterwege te laten. Onder

[6] Dit is gebeurd op basis van de tekst opgesteld in kamerstuk 35561-2. Voor een overzicht van de zoekvragen die gebruikt zijn om krantenartikelen te zoeken, de bronnen uit welke deze zoekvragen voortgekomen zijn, en de relevantie Tweede-Kamerdossiers, zie appendix 2.
[7] Voor de actuele periodes hebben we gebruik gemaakt van LexisNexis. Voor historische artikelen hebben we gebruik gemaakt van Delpher. Voor meer informatie over de verzamelde artikelen, zie appendix 3.
[8] Ter verduidelijking: we hebben eerst berekend met welke waarschijnlijkheid een leeg document tot een bepaald topic behoort. Vervolgens hebben we de allerhoogste waarschijnlijkheid binnen een topic bepaald. Stel nu dat de eerste 10% is, en de tweede waarschijnlijkheid 70%. De drie stratificaties lopen dan van 10%-30%, 30%-50% en 50%-70%. Uit elke bak trekken we aselect vijf documenten, en dat doen we twee keer, namelijk één keer voor de massa met door PEGA gevonden documenten en één keer voor de niet door PEGA gevonden documenten.

deze metadata valt onder andere of een document door PEGA gevonden is of niet, of ze het document reeds hebben aangeleverd of dat het bij een eerdere beoordeling in de prullenbak is gezet.

De documenten zijn op donderdag 16 september voorgelegd aan het team PEGA en in drie interviews besproken op maandag 20 september.

6.1.3 Aanpak onderzoeksvraag 2

Voor het beantwoorden van onderzoeksvraag 2 is gekeken of er qua informatie vergelijkbare documenten in de blauwe massa van suggesties zitten die niet in het dossier geel (gekozen) zitten (Figuur 2). De ADR heeft ook gekeken of er nog vergelijkbare documenten in de rode massa van "Zoek en Vind" zaten die niet in het dossier geel (gekozen) waren opgenomen. Eventuele opvallende zaken zijn ieder apart voorgelegd en besproken met team PEGA door middel van een deelwaarneming en interviews.

In de huidige opzet is buiten beschouwing gelaten in hoeverre "Zoek en Vind" alle bronnen en documenten kan benaderen en doorzoeken. Dit is gedaan omdat hiervoor toegang nodig is tot de productieomgeving van "Zoek en Vind" en de bronnen. Deze toegang was niet op korte termijn (binnen de doorlooptijd van het onderzoek) te regelen.

6.1.4 Aanpak aanvullende onderzoeksvraag

Voor het beantwoorden van de aanvullende onderzoeksvraag heeft de ADR twee aanvullende stappen genomen. De eerste stap is het terugbrengen van het aantal documenten in de massa. Dit is in de eerste plaats gedaan door de set te ontdubbelen[9]. De tweede stap is binnen "Zoek en Vind" het aantal documenten terug te brengen door een aantal directories (bv. een directory voor communicatie) uit te sluiten die door het team PEGA als niet relevant werden beschouwen en het toepassen van NOT query's opgesteld door het team PEGA. Een NOT-query bevat een aantal termen waarvan het voorkomen van de term in een document ervoor zorgt dat het document juist níet als resultaat naar voren komt. Na voornoemde stappen resteerde een set van 137.000 documenten. De ADR heeft deze set geclusterd om zo gelijksoortige documenten (met een gelijksoortige distributie over de topics) te bundelen en daarmee de beoordeling door het team PEGA te vergemakkelijken.

Elk cluster bevat soortgelijke documenten en is daarmee in meer of mindere mate relevant voor het onderwerp. Door elk cluster een indicator van relevantie mee te geven, kan het team PEGA iteratief de meest interessante clusters bekijken. Voor de aanlevering van de eerste vijftig clusters (Deze bevatten alleen niet-mails omdat de mails toen nog niet verwerkt waren) is de indicator voor relevantie gedefinieerd als de som van de probabilities voor alle 8 inhoudelijke topics (dus de topics 20 en 40 uitgesloten omdat deze enkel woorden over de vorm zijnde brief of mail bevatten).

Op basis van de beoordeling van de eerste vijftig clusters, bleek de indicator voor relevantie aangescherpt te kunnen worden. De ADR kon de relevante topics onderverdelen in inhoud- en vormtopics: vormtopics zijn topic 20 en topic 40 en de rest zijn inhoudelijke topics. Sommige inhoudelijke topics (28, 47, 54) bleken na evaluatie inhoudelijk minder relevante woorden te bevatten dan de overige 5 topics. De nieuwe indicator voor relevantie is daarom gedefinieerd als de som van de probabilities voor topics 2,4,5,7 en 46. De tweede aanlevering met clusters zijn de top 100 clusters – die niet al in de eerdere 50 hadden gezeten, en dit keer met zowel mails als niet-mails – volgens deze nieuwe indicator, gesorteerd van hoog naar laag.

[9] Een aantal documenten komt meermaals voor in de door de ADR gevonden documenten. Denk bijvoorbeeld aan één mail die in meerdere mailboxen terug te vinden is. Om het aantal documenten terug te brengen, is een ontdubbeling uitgevoerd om het aantal duplicaten zo klein mogelijk te maken.

6.2 Gehanteerde Standaard

Deze opdracht is uitgevoerd in overeenstemming met de Internationale Standaarden voor de Beroepsuitoefening van Internal Auditing. Dit onderzoek verschaft geen zekerheid in de vorm van een oordeel of conclusie, omdat het een onderzoeksopdracht betreft en geen controle-, beoordelings- of andere assurance-opdracht. Als hier wel sprake van was geweest, dan zouden we wellicht andere zaken hebben geconstateerd en gerapporteerd.

De opdracht is uitgevoerd conform de algemene uitgangspunten voor de uitoefening van de interne auditfunctie bij de rijksdienst. Daarbij hoort ook een stelsel van kwaliteitsborging. Een onderdeel daarvan is dat er een onafhankelijke kwaliteitstoetsing heeft plaatsgevonden op deze onderzoeksopdracht.

6.3 Verspreiding rapport

De opdrachtgever, pSG EZK, is eigenaar van dit rapport. Dit rapport is primair bestemd voor de opdrachtgever met wie wij deze opdracht zijn overeengekomen. Hoewel het rapport de context van het onderzoek zo goed mogelijk probeert te beschrijven, is het mogelijk dat iemand die de context niet (volledig) kent, de uitkomsten anders interpreteert dan bedoeld.

In de ministerraad is besloten dat het opdrachtgevende ministerie waarvoor de Auditdienst Rijk (ADR) een rapport heeft geschreven, het rapport binnen zes weken op de website van de rijksoverheid plaatst, tenzij daarvoor een uitzondering geldt. De minister van Financiën stuurt elk halfjaar een overzicht naar de Tweede Kamer met de titels van door de ADR uitgebrachte rapporten en plaatst dit overzicht op www.rijksoverheid.nl.

7 Ondertekening

Oosterhesselen, 28 december 2021

Bijlage(n)

Appendix 1

Naam	Bron	Locatie
Parlementaire stukken	Officiële Bekendmakingen	https://zoek.officielebekendmakingen.nl/
Kranten	LexisNexis	http://signin.nexis.com
Documenten EZK	Zoek & Vind	Ministerie van Economische Zaken

Gebruikt dossiernummer officiële bekendmakingen:

- 33529

Gebruikte kranten LexisNexis:

- Dagblad van het Noorden'
- De Stentor'
- Leeuwarder Courant'
- Friesch Dagblad'
- Trouw',
- NRC Handelsblad'
- De Volkskrant'
- De Telegraaf'

Appendix 2

Nr	Zoekvraag o.b.v. tweede-kamer stukken gebruikt bij zoeken krantenartikelen
Topic 1	'groningen OR schade OR aardbeving AND allcaps(ncg) OR risicogebied OR referentiegebied'
Topic 2	'groningen OR allcaps(nam) OR gas AND seismisch OR sodm OR winningsplan'
Topic 3	'groningen OR schade OR gaswinning AND allcaps(tcmg) OR allcaps(img) OR mijnbouwschade'
Topic 5	'norg OR allcaps(nam) OR groningenveld AND akkoord OR vergoeding OR overeenkomst'
Rapport	'groningen AND *gaswinning OR schadeafhandeling OR versterking'

Appendix 3

topic	gemiddeld ADR	PEGA	gemiddeld	voorlopig 10-09-2021	definitief		
4	3	3	3	3	3	1	niet relevant
7	3	3	3	3	3	2	onzeker
46	3	3	3	3	3	3	relevant
2	3	2	2,5	3	3		
28	2	3	2,5	3	3		
5	2	2	2	3	3		
47	2	2	2	2	3		
54	2	2	2	3	3		
20	1	2	1,5	2	3		
40	1	2	1,5	3	3		

Bijlage Managementreactie EZK

Deze bijlage bevat de managementreactie van EZK op het conceptrapport 0.96 voor managementreactie. De reactie is op 27 december/januari 2021 jl. ontvangen. De reactie is geen onderdeel van het uitgevoerde onderzoek en de inhoud valt buiten onze verantwoordelijkheid.

Ministerie van Economische Zaken en Klimaat

> Retouradres Postbus 20401 2500 EK Den Haag

ADR
T.a.v. de [redacted]
T.a.v. de [redacted]

Bureau Bestuursraad

Bezoekadres
Bezuidenhoutseweg 73
2594 AC Den Haag

Postadres
Postbus 20401
2500 EK Den Haag

Overheidsidentificatienr
00000001003214369000

T 070 379 8911 (algemeen)
F 070 378 6100 (algemeen)
www.rijksoverheid.nl/ezk

Behandeld door
[redacted]
T 070 [redacted]
[redacted]@minezk.nl

Ons kenmerk
BBR / 21326664

Uw kenmerk

Bijlage(n)

Datum 27 december 2021
Betreft Definitief ADR rapport PEGA

Geachte [redacted] en [redacted]

Managementreactie ADR-rapport PEGA

Hierbij doe ik u mijn reactie toekomen op uw onderzoeksrapport vastleggingen binnen project PEGA.

Bewuste keuze voor een experimentele werkwijze
Onderhavig onderzoek had voor EZK als doel om door middel van een externe validatie door de ADR een kwaliteitsslag te kunnen maken op het proces van het vinden en leveren van documenten aan de commissie. Beperkende factor was hierbij dat het onderzoek parallel liep aan de vorderingenfase, waardoor het PEGA-team zoveel mogelijk moest worden ontzien, gelet op de grote hoeveelheden informatie die in zeer korte tijd moesten worden verwerkt. De ADR heeft er daarom, uiteraard in overleg met de opdrachtgever, voor gekozen om op een experimentele wijze, met behulp van big data analyse, door het beschikbare bronmateriaal te gaan en te kijken of die aanpak nog aanvullende, door het PEGA-team niet gevonden relevante documenten zou opleveren.

Werkwijze EZK vs. ADR
EZK heeft zich bij het proces van zoeken, beoordelen en leveren van relevante documenten gericht op de onderwerpen en de bewoordingen van de vorderingen van de enquêtecommissie. De (series van) zoekvragen, die zijn samengesteld in overleg met inhoudelijke experts, bevatten daarom gerichte, vaak technische zoektermen. Er is voldoende reden om aan te nemen dat het PEGA-team daarbij zeer grondig te werk is gegaan en heel veel relevante informatie heeft kunnen vinden en heeft geleverd. Uiteindelijk heeft EZK, verdeeld over zeven leveringen, in totaal ruim 100.000 documenten aan de enquêtecommissie geleverd. Daarbij was vanaf het begin duidelijk dat 100% volledigheid onmogelijk zou zijn, gelet op de looptijd van dit dossier (ruim 60 jaar archief in diverse bronnen, waaronder veel fysieke archieven die zijn gedigitaliseerd door middel van scans).

Pagina 1 van 4

De ADR-werkwijze richtte zich bewust niet op de vorderingen, maar leverde op basis van een big data analyse een andere keuze in bepalende zoektermen en zoekvragen op, waarmee in de totale hoeveelheid beschikbare documenten werd gezocht. Dit leverde een geheel andere dwarsdoorsnede van het materiaal op. Dit heeft als voordeel dat de kans groot is dat daarbij documenten naar boven komen die in de meer gerichte werkwijze van het PEGA-team mogelijk toch zijn gemist. Het nadeel daarvan is echter dat de hoeveelheid mogelijk relevant materieel groot is en blijft, omdat er onherroepelijk ook veel niet-relevant materiaal in de verzameling documenten zit.

Het vereist alsnog diverse analyses en het nodige handwerk om deze hoeveelheid beheersbaar te maken en uiteindelijk alleen daadwerkelijk relevante documenten aan de enquêtecommissie te kunnen leveren. Er is daarom, zowel door de ADR-medewerkers als het PEGA-team, in oktober en november van 2021 nog veel tijd en energie gestoken in een poging de verbinding met de vorderingen te leggen.

Reactie op de bevindingen
De constatering van de ADR dat er niet gevonden relevante documenten zijn, komt voor EZK niet als een verrassing. Dat geldt ook voor de bevinding dat sprake is van voortschrijdend inzicht waardoor documenten door het PEGA-team op een later moment alsnog als relevant werden beoordeeld. Beide constateringen vloeien in hoge mate voort uit de bewuste keuze die de opdrachtgever, op advies van de Bestuursraad van EZK, in de zomer van 2020 heeft gemaakt om het PEGA-team op te bouwen met relatief veel nieuwe medewerkers om zo de inzet van inhoudelijk betrokken beleidsmedewerkers zoveel mogelijk te beperken, zodat de lopende zaken binnen het Groningen-dossier geen vertraging zouden ondervinden als gevolg van de werkzaamheden voor de parlementaire enquête. Voordeel hiervan was dat het PEGA-team onafhankelijk van de betrokken beleidsdirecties kon worden gepositioneerd. Een neveneffect van deze keuze is dat de kennis van het onderwerp binnen het PEGA-team gaandeweg sterk is gegroeid, waardoor een beoordeling van een document een half jaar later anders kan zijn.

Dilemma
Als gevolg van de complexiteit van de experimentele werkwijze van de ADR en de mede daardoor ontstane vertragingen, bleek de geplande afronding van het onderzoek vóór de zomer van 2021 niet mogelijk. Hierdoor ontstond eind november een dilemma: de ADR had weliswaar vastgesteld dat er zich in een totale voorraad van 137.000 documenten mogelijk nog niet gevonden relevante documenten bevinden, maar was er op dat moment nog niet in geslaagd om de uitkomsten van de data-analyse op een gevalideerde en praktisch uitvoerbare wijze te verbinden aan de concrete vorderingen van de enquêtecommissie. Het was dus niet helder welke documenten daadwerkelijk relevant zijn in het licht van de vorderingen en welke documenten niet toch al in een andere (meer/minder definitieve) vorm zijn geleverd. Het was onduidelijk om hoeveel unieke relevante documenten het zou gaan en hoeveel tijd het zou kosten om deze te vinden en alsnog te leveren aan de enquêtecommissie, mede gelet op het handwerk dat dit vergt aan de kant van het PEGA-team.

Besluit tot afronding van het onderzoek

Het leveren van meer documenten in deze fase van het onderzoek van de enquêtecommissie – kort voor de start van de besloten voorgesprekken per 6 januari 2022 – leek weinig opportuun, aangezien de commissie het onderzoek van documenten grotendeels had afgerond om met de uitkomsten daarvan de fase van de besloten voorgesprekken in te gaan. Dat betekent dat de toegevoegde waarde van nieuwe leveringen van documenten voor het onderzoek van de enquêtecommissie steeds verder afneemt.

Voortzetting en verdieping van de analyse zou een zwaar beperkend effect hebben op de capaciteit van het PEGA-team dat inmiddels bezig was om tientallen genodigden voor de besloten voorgesprekken te begeleiden door middel van het maken van persoonlijke dossiers met voor deze personen relevante documenten. De stappen om de ontbrekende verbinding met de vorderingen van de enquêtecommissie toch te kunnen leggen, zijn tijdrovend. Bovendien is deze nadere analyse in meerdere opzichten een onzeker proces waarvan niet zeker is of het concreet voldoende relevant materiaal zou opleveren.

Derhalve heeft de Bestuursraad op 29 november 2021 besloten om de ADR te verzoeken dit vraaggestuurde onderzoek af te ronden. De Bestuursraad heeft daarbij wel aangetekend dat de ADR-verzameling met mogelijk relevante documenten alsnog wordt geanalyseerd als de enquêtecommissie op enig moment een gerichte vordering naar bepaalde onderwerpen doet.

Aanbeveling ADR
Ondanks de vroegtijdige afronding, zijn de leereffecten van het onderzoek aanzienlijk. De ADR doet in dit rapport dan ook een belangrijke aanbeveling die EZK van harte ondersteunt, nl. dat een big data analyse bij uitstek waardevol is in de voorbereidende fase van een parlementaire enquête. Hiermee kan immers het 'zoekgebied' met relevante documenten aanzienlijk kan worden verkleind, waarna gerichter kan worden gezocht zodra de vorderingen zijn ontvangen. Echter, als de noodzakelijke stappen voor die analyse achteraf moeten worden doorlopen – zoals in de casus bij EZK het geval was – wordt in feite de gehele vorderingenfase over gedaan. Dat bleek in dit stadium redelijkerwijs niet meer uitvoerbaar.

Ik dank u hartelijk voor uw onderzoek.

[redacted]

drs. G.M. Keijzer – Baldé
Plv. Secretaris – Generaal EZK

Auditdienst Rijk
Postbus 20201
2500 EE Den Haag
(070) [redacted]