

Dataset hackathon open overheid

Beschrijving

De dataset bevat een deel van de onder de Wet open overheid op open.overheid.nl actief openbaar gemaakte documenten. Het gaat om ruim 77.000 documenten, van verschillende bronnen:

- Documenten die door departementen actief openbaar zijn gemaakt via het Woo-aanleverloket van KOOP
- Stukken die door verschillende departementen zijn gepubliceerd op [Rijksoverheid.nl](https://rijksoverheid.nl)
- Parlementaire stukken

Gebruik van de data

De data die is te gebruiken voor de hackathon is op verschillende manieren toegankelijk:

- Via de index file (.csv formaat) die te downloaden is. Elke regel hierop bevat de identifier, titel en de urls van de bestanden die gekoppeld zijn aan een document
- Door te zoeken in een [Apache Solr](#) collectie die beschikbaar is gesteld. Voor gebruik van Apache Solr, zie de [documentatie](#)

Beide manieren van het gebruik van de data worden hieronder verder toegelicht, nadat de data zelf beschreven is.

Toelichting op de data

Algemeen

Voor ieder document is een aantal bestanden te vinden in een Amazon S3 bucket:

- `metadata.json` - de Woo-metadata van het document (zie toelichting hieronder)
- `relaties.json` - relaties tussen documenten
- `text.txt` - De ruwe tekst van het document
- `.pdf file` - Indien beschikbaar, het oorspronkelijke pdf-bestand zoals gepubliceerd door het bestuursorgaan.

Verwijzen naar deze bestanden zijn te vinden in [de index file](#) (.csv) en in de Solr collectie.

Metadata

Een voorbeeld van een metadatabestand is hieronder weergegeven. De namen van de metadatavelden spreken voor zich. Voor enkele aspecten van de metadata een korte toelichting:

- Een aantal velden (bijvoorbeeld verantwoordelijke en documentsoorten) is voorzien van een `id` en een `label`. Deze zijn afkomstig uit TOOI-waardelijsten die zijn te vinden op <https://standaarden.overheid.nl/tooi/waardelijsten/>. De labels die zijn opgenomen in de data die beschikbaar is gesteld kunnen afwijken van de meest actuele labels uit deze waardelijsten. Het veld `bronwaarde` dat soms is opgenomen is het label zoals dat door de publicerende organisatie is toegekend. Dit is gestandaardiseerd naar de TOOI waardelijst
- Gebruik de weblocatie als je in een applicatie het document wil presenteren.
- Het aggregatiekenmerk is een manier waarop documenten gegroepeerd worden in dossiers. Dit wordt vooral gebruikt voor kamerstukken

```
{
  "pid": "https://open.overheid.nl/documenten/oep-205c59b13753488947513a4a4bc9810db2a5e197",
  "weblocatie": "https://zoek.officielebekendmakingen.nl/kst-29544-1276.html",
  "creatiedatum": "2025-03-31",
  "geldigheid": {
    "begindatum": "2025-03-31T22:00:00Z"
  },
  "identifiers": [
    "kst-29544-1276"
  ],
  "verantwoordelijke": {
    "type": "overheid:AndereOrganisatie",
    "id": "https://identifier.overheid.nl/tooi/id/oorg/oorg10002",
    "label": "Tweede Kamer",
    "bronwaarde": "Tweede Kamer der Staten-Generaal"
  },
  "publisher": {
    "id": "https://identifier.overheid.nl/tooi/id/organisatie_onbekend",
    "label": "Onbekend"
  },
}
```

```

"opsteller": {
  "type": "overheid:AndereOrganisatie",
  "id": "https://identifier.overheid.nl/tooi/id/oorg/oorg10002",
  "label": "Tweede Kamer",
  "bronwaarde": "Tweede Kamer der Staten-Generaal"
},
"language": {
  "type": "eu:Taal",
  "id": "http://publications.europa.eu/resource/authority/language/NLD",
  "label": "Nederlands",
  "bronwaarde": "nl"
},
"omschrijvingen": [
  "Arbeidsmarktbeleid; Brief regering; Reactie op verzoek van het lid Patijn, gedaan tijdens de Regeling van
  Werkzaamheden van 11 maart 2025, over het bericht dat werknemers van Transavia met vaste contracten zes weken
  per jaar een WW-uitkering krijgen"
],
"titelcollectie": {
  "officiëleTitel": "29544;26448, nr. 1276 - Arbeidsmarktbeleid"
},
"classificatiecollectie": {
  "documentsoorten": [
    {
      "type": "overheid:Informatietype",
      "id": "https://identifier.overheid.nl/tooi/def/thes/kern/c_056a75e1",
      "label": "Kamerstuk"
    }
  ],
  "themas": [
    {
      "type": "plooi:Topthema",
      "id": "https://identifier.overheid.nl/tooi/def/thes/top/c_fc530104",
      "label": "arbeidsongeschiktheid en werkloosheid",
      "bronwaarde": "Werkgelegenheid"
    }
  ],
  "informatiecategorieen": [
    {
      "id": "https://identifier.overheid.nl/tooi/def/thes/kern/c_c76862ab",
      "label": "vergaderstukken Staten-Generaal"
    }
  ]
},
"aggregatiekenmerk": "29544;26448",
"extraMetadata": [
  {
    "prefix": "plooi.displayfield",
    "velden": [
      {
        "key": "vergaderjaar",
        "values": [
          "2024-2025"
        ]
      },
      {
        "key": "indiener",
        "values": [
          "Y.J. van Hijum"
        ]
      },
      {
        "key": "documentsubsoort",
        "values": [
          "Brief"
        ]
      }
    ]
  }
]
}

```

Relaties

Sommige documenten hebben relaties met andere documenten. In dat geval is er een bestand `relaties.json` aanwezig. Hieronder is een voorbeeld gegeven van een dergelijk bestand.

- `role` geeft het type relatie aan, zoals gedefinieerd in de betreffende TOOI waardelijst
- `relation` geeft de pid van het document aan waarmee deze relatie bestaat

```
[
  {
    "relation": "https://open.overheid.nl/documenten/513a511f-5a90-4cc3-93b4-c2c6cc6806e0",
    "role": "https://identifier.overheid.nl/tooi/def/thes/kern/c_4d1ea9ba"
  },
  {
    "relation": "https://open.overheid.nl/documenten/oep-a6ee0a2f271dc15261c9c030f9cfdbb3154b724d",
    "role": "https://identifier.overheid.nl/plooi/def/thes/documentrelatie/identiteitsgroep"
  },
  {
    "relation": "https://open.overheid.nl/documenten/oep-d022c4bf82945e6d3a5bb3f03b69cf12ebe19623",
    "role": "https://identifier.overheid.nl/plooi/def/thes/documentrelatie/identiteitsgroep"
  }
]
```

Index file

De [index file](#) (.csv) bevat voor ieder document dat beschikbaar is een regel met de volgende velden:

- `pid` - unieke identifier
- `titel` - titel van het document
- `metadata_file` - locatie van het bestand met metadata
- `txt_file` - locatie van het bestand met de ruwe tekst (indien beschikbaar)
- `pdf_file` - locatie van het pdf-bestand (indien aanwezig)
- `relaties_file` - locatie van het bestand waarin de relaties van het document staan (indien aanwezig)

Solr collectie

De dataset is geïndexeerd in een [Apache Solr](#) collectie die te vinden is op <https://innovatielab.koopoverheid.nl/hackathon-open-overheid/>. Voor gebruik van Apache Solr, zie de [documentatie](#). De inrichting van de collectie is basaal, er zijn geen geavanceerde zoekfuncties ingericht anders dan wat Solr standaard te bieden heeft. Dat geldt ook voor de behandeling van de Nederlandse taal, hier zijn slechts basisvoorzieningen (eenvoudige lemmatization bijvoorbeeld) ingericht.

Enkele voorbeelden van queries op deze collectie om data te bevragen in deze collectie.

Functie	Solr query	Opmerkingen
Opvragen van de identifier van alle documenten	<code>select?q=*&fl=identifier,titel</code>	
Opvragen van alle documenten met een bepaalde tekst	<code>select?q=tekst:toeslag&fl=identifier,titel</code>	
Alle aanwezige verantwoordelijke_id's opvragen	<code>select?q=*&facet=true&facet.field=verantwoordelijke_id&facet.mincount=1&facet.limit=-1&rows=0</code>	Zie documentatie over facets
Filteren op documenten van een specifieke organisatie	<code>select?q=tekst:toeslag&q=verantwoordelijke_id:"https://identifier.overheid.nl/tooi/id/oorg/oorg10002"</code>	Let op escaping van special character

Het default response type is json. XML is ook mogelijk, door de parameter `wt=xml` toe te voegen aan de URL. Als het nodig is om delen van de dataset op basis van een selectie te verwerken, kan dat door met de [parameters](#) `rows` en `start` steeds pagina's zoekresultaten op te vragen. Gebruik hierbij eventueel [cursors](#), dit komt de performance ten goede. Verzoek: don't hammer the server.. deze collectie is ingericht met beperkte capaciteit. Vraag geen heel grote hoeveelheden documenten op met een enkel request (door de `rows` parameter heel hoog te zetten)

Een voorbeeld van een document uit deze Solr collectie is hieronder weergegeven. Een subset van de hierboven beschreven metadata is beschikbaar.

```

{
  "pid": "https://open.overheid.nl/documenten/oep-41a916c33b325fcc63c3a793f0aa9f375b40f464",
  "identificer": "oep-41a916c33b325fcc63c3a793f0aa9f375b40f464",
  "url": "https://zoek.officielebekendmakingen.nl/kst-22112-3961.html",
  "titel": "22112, nr. 3961 - Nieuwe Commissievoorstellen en initiatieven van de lidstaten van de Europese Unie",
  "aggregatiekenmerk": "22112",
  "omschrijving": [
    "Nieuwe Commissievoorstellen en initiatieven van de lidstaten van de Europese Unie; Brief regering; Kwartaalrapportage EU-wetgevingsonderhandelingen Q2 2024"
  ],
  "verantwoordelijke": "Tweede Kamer",
  "verantwoordelijke_type": "overheid:AndereOrganisatie",
  "verantwoordelijke_id": "https://identificer.overheid.nl/tooi/id/oorg/oorg10002",
  "publisher": "Onbekend",
  "publisher_id": "https://identificer.overheid.nl/tooi/id/organisatie_onbekend",
  "documentsoort": [
    "Kamerstuk"
  ],
  "documentsoort_id": [
    "https://identificer.overheid.nl/tooi/def/thes/kern/c_056a75e1"
  ],
  "informatiecategorie": [
    "vergaderstukken Staten-Generaal"
  ],
  "informatiecategorie_id": [
    "https://identificer.overheid.nl/tooi/def/thes/kern/c_c76862ab"
  ],
  "thema": [
    "Europese zaken"
  ],
  "thema_id": [
    "https://identificer.overheid.nl/tooi/def/thes/top/c_68609e19"
  ],
  "metadata_file": "https://hackathon-open-overheid.s3.eu-central-1.amazonaws.com/2024/08/oep-41a916c33b325fcc63c3a793f0aa9f375b40f464/metadata.json",
  "txt_file": "https://hackathon-open-overheid.s3.eu-central-1.amazonaws.com/2024/08/oep-41a916c33b325fcc63c3a793f0aa9f375b40f464/text.txt",
  "pdf_file": "https://hackathon-open-overheid.s3.eu-central-1.amazonaws.com/2024/08/oep-41a916c33b325fcc63c3a793f0aa9f375b40f464/kst-22112-3961.pdf",
  "timestamp": "2025-05-06T12:28:53.418Z",
}

```