

```
In [ ]: !jupyter nbconvert GemAmsterdam_2022_09_21.ipynb --to slides --post serve
```

## Titel en abstract

### Woo dossiers zijn niet FAIR

Maik Larooij, Jaap Kamps, Maarten Marx  
Archiefwetenschappen/Informatica  
Universiteit van Amsterdam

We laten zien dat de meeste nu gepubliceerde Woo-dossiers niet voldoen aan de FAIR data principes, die zeggen dat data vindbaar, ~~toegankelijk~~ archiveerbaar, uitwisselbaar, en herbruikbaar moet zijn. Dit heeft ons inziens twee oorzaken: er wordt veel gebruik gemaakt van een niet passend datamodel voor Woo-dossiers, en de dossiers worden op enorm veel plekken gemaakt, zonder veel sturing van bovenaf.

Het voldoen aan die FAIR data principes is niet alleen leuk voor wetenschappers die een *Woogle*, een zoekmachine speciaal voor Woo-dossiers, willen maken, het is ook verplicht volgens artikel 2.4.3 van de Woo.

Gelukkig is het helemaal niet moeilijk om aan die principes te voldoen, en het kan ook nog bijna kosteloos. Er is eigenlijk alleen software nodig die de Wob jurist helpt en stuurt. Die zijn we deze herfst aan de UvA aan het ontwikkelen. Liefst met behulp van wat directe "probleemhouders", gemeentes en provincies die de Woo's produceren en publiceren.

Om het gebruik van die software aantrekkelijk te maken voor gemeenten, laten we in de software zien wat er allemaal kan als je je data FAIR produceert: een interne *Woogle* voor de eigen Woo's; een data dashboard voor de eigen gepubliceerde Woo's; een API en automatische upload naar PLOOI.

Zie ook het stukje op [Kennisplatform Informatiehuishouding Overheden](https://kia.pleio.nl/groups/view/80563a73-206e-42c2-878f-678eef75af2b/kennisplatform-informatiehuishouding-overheden/blog/view/bfdc0597-c668-4391-aaa6-ce6ed632bcee/woogle-een-zoekmachine-speciaal-voor-wob-en-woo) (<https://kia.pleio.nl/groups/view/80563a73-206e-42c2-878f-678eef75af2b/kennisplatform-informatiehuishouding-overheden/blog/view/bfdc0597-c668-4391-aaa6-ce6ed632bcee/woogle-een-zoekmachine-speciaal-voor-wob-en-woo>) van Jeroen Jonkers.

### Woo dossiers zijn niet FAIR

Maik Larooij, Jaap Kamps, Maarten Marx  
Archiefwetenschappen/Informatica  
Universiteit van Amsterdam

Archief Gem. Amsterdam 2022-09-21

- Gebaseerd op scriptie van Maik en ESB artikel.
- Onderdeel van NWO ACCESS project naar publiceren van stukken van lagere overheden.

# Aanleiding, wie zijn wij?

- Beide achtergrond in zoektechnologie
- Veel ervaring met toegankelijk maken van overheidsdocumenten
  - Handelingen Staten Generaal 1814-2016 in XML
  - [waaroverheid.nl](http://waaroverheid.nl/) (<http://waaroverheid.nl/>)



## Aanleiding voor Wob onderzoek

- Lezing van Follow The Money, over hoe lastig het voor hen is om in de dossiers te zoeken.
- Allemaal leuke wetenschappelijke uitdagingen in die Wob dossiers.
- *Laten we een **Woogle** of **Woobipedia** gaan maken!*

## Oogst sinds winter 2022

- 3 artikelen gepubliceerd, nog 3 in de pen
- 8 scripties over Woo verdedigd, nu 4 bezig

## Inhoud: Woo dossiers zijn niet FAIR

1. Wat is het probleem?
2. Hoe is het op te lossen?
3. Is dat moeilijk? Welnee!
  - Zelf aan de slag.
4. Fair Woo software
  - jullie hulp bij nodig

## Theoretisch kader

- We toetsen de Wob/Woo dossiers aan de FAIR data principes [Wilkinson et al, 2016]
- Data moet
  - vindbaar,
  - ~~toegankelijk~~ archiveerbaar,
  - uitwisselbaar, en
  - herbruikbaar zijn.
- En eigenlijk ook volgens artikel [2.4.3 van de Woo](https://wetten.overheid.nl/BWBR0045754/2022-08-01#Hoofdstuk2_Artikel2.4) ([https://wetten.overheid.nl/BWBR0045754/2022-08-01#Hoofdstuk2\\_Artikel2.4](https://wetten.overheid.nl/BWBR0045754/2022-08-01#Hoofdstuk2_Artikel2.4)).

# Onze test hiervoor: kunnen we een Woogle maken?

- zoekmachine speciaal voor (alle) woo-dossiers

## Heel lastig met de huidige wob dossiers

**geen woorden**

**geen stukken**

**geen gegevens**

**Dit zijn precies de bouwstenen voor**

- zo'n zoekmachine,
- de 4 FAIR principes.

## Resultaat

- Zonder woorden, stukken en gegevens
- kan je alleen met enorm veel geduld en mankracht
  - zaken *terugvinden*,
  - feiten *opzoeken*, en
  - dingen *uitzoeken*.

**dagelijkse praktijk bij clubs als FTM, Argos, P. Omtzigt,...**

# Samenvatting probleem

1. Geen woorden en geen stukken
  - Komt door opnieuw inscannen, zonder OCR, van stapels documenten.
  - Zelfs met Ctrl F kan je niet zoeken
  - Enorm groot onhandelbaar "aan elkaar geplakt" bestand
2. Geen gegevens (metadata)
  - Metadata model past niet.
  - Een Wob is geen file/document, **maar een dossier**

## Intermezzo, resultaat Wob Amsterdam

- Website gescraped herfst 2021
- 201 "besluit PDFs"
- Resultaten kennis extractie gisteren gedaan door Julian Venhuizen.

### geen woorden

- Slechts 61/202 = 30% bevatten woorden.
- Dit is het **besluit**, niet een vrijgegeven stuk!

### Metadata extractie

- kon dus maar op 30% van de besluiten
- Hoe goed ging het?
  - **Datums**
    - besluit datum: alle gevonden, 90% goed.
    - verzoek datum alle gevonden, 75% goed.
  - **Besluit**
    - overeenkomst 51%
  - **het verzoek**
    - overeenkomst 38%

# Conclusie klein experimentje

- Extractie met AI technieken is nog veel te slecht voor in een "Woobipedia archief."
- Heel vaak gaat het niet eens omdat de stukken niet eens machine leesbaar zijn.
- En dit is nog een "nette gemeente".
  - Vaak zitten verzoek, besluit, inventarislijst en alle vrijgegeven stukken aan elkaar in 1 PDF.

## Geef liever zelf als metadata wat er toch al is

- mijn studenten kunnen spannender dingen gaan doen
- jullie hebben een fijn gevoel, (en voldoen aan de Woo, artikel 2.4.3)

In [22]: `import pandas as pd`

```
adam= pd.read_csv('evaluated_amsterdam_wobs.csv', index_col='file_name')
.drop(['folder', 'Unnamed: 0', 'path_x', 'text', 'clean_text'], axis=1)
print(adam.shape)
adam.head()
```

(20, 36)

Out[22]:

|                                                                                                 | decision_date_x | request_date_x |
|-------------------------------------------------------------------------------------------------|-----------------|----------------|
| file_name                                                                                       |                 |                |
| 1_besluit_wob-verzoek_genhoes_19.pdf                                                            | 4-8-2020        | 13-7-2020      |
| 17_getekend_wob-besluit_geredigeerd.pdf                                                         | 8-10-2020       | 25-6-2020      |
| 2018-09-20_geanonimiseerd_wob-besluit_inzake_werkinstructie_beoordelen_deskundigenrapporten.pdf | 20-9-2018       | 6-8-2018       |
| 181211_deelbesluit_wob_verzoek_3_contracten_leadbuyer_ict.pdf                                   | 31-10-2018      | 26-11-2018     |
| 20201203_wob_besluit_handhavingsverzoeken_geredigeerd.pdf                                       | 7-12-2020       | 29-9-2020      |

5 rows × 36 columns

# Metadata voor een Wob dossier

## Voor het dossier

1. Datums
2. De "vraag"
3. Verantwoordelijke voor antwoord
4. etc

## En per (vrijgegeven) document

- De kolommen in de inventarislijst
- De tekst in de documenten
- Typische 'overheid.nl' metadata
  - onderwerp
  - datum
  - van wie komt het?
  - type document
  - etc

# Voorbeeld metadata dossier en per document

```
{
  "title": "Zonneweide Jaagweg (sept, okt, nov, dec 2021)",
  "identifier": "Woo.ZW8",
  "type": "Actieve openbaarmaking",
  "topic": "Zonneweide Jaagweg",
  "sourceUrl": "https://www.noord-holland.nl/Loket/Pilot_actieve_openbaarm
ak
    ing/Documenten/Zonneweide_Jaagweg_sept_okt_nov_dec_2021",
  "decisionDate": "2021-12-23",
  "handledBy": "Provincie Noord-Holland",
  "valuation": "Deels openbaar",
  "documents": [
    {
      "title": "Besluit actieve openbaarmaking Zonneweide Jaagweg
        sept okt nov dec 2021",
      "documentType": "Besluit",
      "date": "2021-12-23",
      "language": "nld",
      "valuation": "Openbaar",
      "mimeType": "application/pdf",
      ....
    },
    {
      "title": "Memo BO verslag gesprek 20210907",
      "documentType": "Vrijgegeven document",
      "date": "2021-07-20",
      "recipient": "Gedeputeerde Stigter, wethouder Van Dolder",
      "annexType": "Memo",
      "valuation": "Deels openbaar",
      "groundsOfRefusal": ["10.2.e"],
      .....
    }
    ...
  ],
}
```

# Wat levert al dat extra metadata gedoe nou op?

Voorbeeld vraag: Hoe lang duurt de afhandeling van een wob verzoek nou gemiddeld?

- Onderwerp van *Ondraagelijk Traag* [IMM, Openstate, 2021]
  - 12 HBO studenten, half jaar bezig
  - zo'n 1000 wobs
  - data van indiening en besluit staan "gewoon" in het besluit.
- simpele Excel vraag als je de **metadata** maar hebt.

## Verdeling van de afhandelingsduur (N=1088)

```
df['Antwoordtermijn']=  
    (df['Datum van antwoord'] - df['Datum van binnenkomst']).dt.days  
sns.displot(df,x='Antwoordtermijn', kde=True);  
df.Antwoordtermijn.describe().round(0)
```



## Uitsplitsen naar ministerie

```
sns.boxplot(data=df[df.Antwoordtermijn<250], # malle uitschieters eruit  
            y='Antwoordtermijn',  
            x='Ministerie' , order=box.index);
```

# Twée simpele beleidsvragen: voorbeeld 2

## Vraag 2: Waarover vragen mensen informatie op via de Wob?

- Zou je uit de *beschrijving* of *trefwoorden* kunnen halen.
- Voorbeeld voor Ministerie LNV
- steeds het stukje tekst na iets als *Wob-verzoek over* in de besluitbrief
- Nog best lastig om automatisch naar onderwerpen te groeperen....

```
'VION-groep',  
'Covid-19 en nertsen',  
'de coronacrisis',  
'Uroplatus alluaudi',  
'Trioceros Deremensis',  
'Onderzoeken visteelt',  
'Wegenbouwbitumenzaak uit 2006',  
'Rechtmatig vissen',  
'PAS-meldingen',  
'Register zelfzuivelaars',  
'Gesloten pootringen toekans',  
'Dierenkadavers',  
'POP3 project Mierbeek',  
'welzijn paarden',  
'Lijst erkende kopers van melk',  
'Varkenshouderij in Oijen',  
'Geregistreeerde varkens',  
'Klachten diervoederwebsite',  
'tegen hondenfokker',  
'Eendenhouderijen',  
'Verkoper van honden',  
'Diertellingen Eindhovensebaan',
```

## Conclusie: wat is het probleem?

- Wob dossiers zijn vrijwel niet computer leesbaar,
  - geen woorden
  - geen stukken
  - geen gegevens
- en daardoor
  - niet terug te vinden, dus niet herbruikbaar
  - niet doorzoekbaar
  - niet geschikt om iets uit te zoeken.

## Hoe is dit op te lossen?

## Nou.....

- wel woorden
- wel losse documenten
- metadata voor het dossier als geheel en voor elk los document

## Is dat nou duur?

- *Ondragelijk Traag* schatte de kosten van een Wob dossier op **150 Euro per pagina**.

## Kosten van ons voorstel: verwaarloosbaar

### Losse leesbare documenten

- staan hoogstwaarschijnlijk al zo bij de Wob-jurist op haar PC;
- alleen een kwestie van even wennen aan een cultuuromslag
  - niet meer die hele stapel uitprinten, en inscannen.

### Metadata kaartje

- Eigenlijk staat alles ook al in de lopende tekst in het besluit
- en per document in de inventarislijst

# Kosten baten

- Gemiddeld Wob dossier is XX paginas.
  - Nb: ook zo'n leuke onuitvoerbare Excel vraag ;-)

## Nu

- Kosten:  $XX \text{ paginas} \times 150 \text{ Euro} = \dots$
- Baten: eenmalig wob-aanvrager "tevreden gesteld"

## Straks met echte *open overheid*

- Kosten:  $XX \times 150 + 1 \text{ extra uur voor de metadata} = \dots$
- Baten:
  - wob-aanvrager stuk beter geholpen
  - door herbruikbaarheid en terugvindbaarheid
    - allerlei nieuwe vragen worden voorkomen
    - als ze toch komen, kan Wob-jurist eerdere stukken veel makkelijker hergebruiken

# Is zo'n metadata kaartje maken moeilijk? Welnee!

- Zelf aan de slag met [WooFairify prototype van Maik \(http://woofairify.herokuapp.com/\)](http://woofairify.herokuapp.com/).

## Opdracht

- **Gegeven:** een Wob besluit op papier
- **Te doen:**
  - maak een metadatakaartje
    - met markeer stiften op papier
    - invoeren in Maiks software.

# We willen jullie overtuigen

## Onze wens en frustratie

- Wij willen een mooie Woogle maken.
- Het **verzamelen** van de Woo dossiers gaat prima.
- Maar het "terugtoeveren" van de woorden en documenten en metadata kan niet foutloos
  - ook niet met de beste AI technieken (daar gaan al die scripties over)

## Onze "wortel"

- Pak dit probleem bij de bron aan: gemeentes, provincies, etc.
- Krijg gemeentes mee door de **voordelen** tastbaar te maken.

# Fair Woo CMS

- lokaal draaiend CMS voor Woo dossiers

## Onderdelen

### 1. Invoer

- van Woo dossier in Word/Excel/mapje met PDFs naar Woo dossier in CMS

### 2. **Publicatie** volgens artikel 2.4.3 Woo

- direct op eigen website
- naar PLOOI
- via API

### 3. **Eigen Woogle**

- interne zoekmachine voor de eigen Woo dossiers
- handig voor de Woo jurist

### 4. **Woo dashboard**

- Plotjes, grafiekjes, tellingen voor overzicht en jaarverslag

# Fair Woo CMS: doe mee!

## Hoe ver zijn we?

- Versie 0.1 draait
  - met de 1000 Openstate Wobs erin
  - versie 1 dashboard
  - *full text search* (via OCR)
  - invoer per stuk en in bulk
  - publicatie als HTML lijst

## Gezocht: mee-doeners en -denkers

- extra wensen, wat kan wel weg
- hoe dit handig in organisatie in te passen
- veiligheid (lokaal, cloud, via VNG, ....)

## Hoe?

Check <https://wooverheid.nl> (<https://wooverheid.nl>)

# Fair Woo CMS samengevat

A relational DB based CMS (MySQL) with strong text search facility serving 4 main goals:

1. Woo-file quality control
  - A. through constraint checking
  - B. help with repairing errors
    - a. auto-completion
    - b. value suggestions
    - c. ...
2. FAIR publishing of Woo-files,
  - A. via an API
    - a. for outside developers (like openstate, or UvA)
    - b. to connect to PLOOI schema
  - B. via a page on the municipalities website listing all of them, with metadata
3. Advanced internal search system
  - A. Mainly for woo-lawyers to search for documents and files that they may re-use in new cases.
4. Reporting and aggregation
  - A. data dashboard
  - B. Templates for doing data analysis, a data dashboard, and making reports like an annual report.
  - C. Possibly via an Excel export and a pandas notebook.

All these activities are carried out using SQL queries (with additional `MATCH` queries on the indexed text fields.)

## The problem that we solve

Small municipalities must, by law, comply with the FIA regulations and publish their Woo-files. But they most often do not have IT expertise in house to do this. So we see in the Netherlands (with around 500 governmental bodies publishing Woo-files) a complete lack of uniformity in the way things are published plus, from the perspective of FAIR data, very poor data quality of the published files.

This is the problem that we seek to solve.

## Hartelijk dank!

## Contact

- [maartenmarx@uva.nl](mailto:maartenmarx@uva.nl)

# FAIR WOO CMS FAQ

Hier verzamelen we alle vragen van potentiële gebruikers rond ons CMS .

1. Hoe ziet een samenwerkingstraject er uit?

- Wij (UvA) leveren de software, de Woo-aanbieder gebruikt die om haar Woo dossiers in op te slaan, te verbeteren, desgewenst te publiceren, om in de eigen Woo dossiers te kunnen zoeken, en om overzichten te maken. Wij verzamelen de gepubliceerde Woo dossiers. In de pilot fase passen we de software aan gedeelde wensen van aanbieders aan.

2. Hoe lang is de planning?

- Afhankelijk van de Woo-aanbieder. Zo lang de software naar wens functioneert kan die gebruikt worden.
- Wij hopen eind Oktober te kunnen starten.
- Wij houden het systeem in ieder geval tot eind 2024 in de lucht.

3. Zijn er eventueel kosten aan verbonden?

- De software is open source en gratis en gebouwd op open source en gratis (zonder licentie) software.
- Installatie (via Docker) lokaal is voor rekening Woo aanbieder.
- Installatie via de UvA server is voorlopig nog gratis, en tijdens de pilot zeker. We hopen dat ook zo te houden.

4. Wat is technisch noodzakelijk?

- Of alleen een browser, voor invoer van Woo-dossiers, en uitvoer van overzichten, en publicatiepagina. Dan draait de software voor een specifieke Woo aanbieder "in the cloud" (= op een server van de Universiteit van Amsterdam).
- Of een PC/server die werkt met [Docker \(https://nl.wikipedia.org/wiki/Docker\\_\(software\)\)](https://nl.wikipedia.org/wiki/Docker_(software)) als men het liever lokaal draait.

5. Hoe is de beveiliging en de toegankelijkheid van de digitale documenten geborgd?

- **Beveiliging** Het Woo-CMS is alleen toegankelijk voor de Woo-aanbieder "van wie het is". Als het lokaal draait via Docker is dat de eigen verantwoordelijkheid. Voor de server op de UvA werken we met twee-weg authenticatie.
- **Toegankelijkheid** Dit is waar het project om gaat! Zorgen dat de dossiers toegankelijk worden, zowel bij de Woo-aanbieder lokaal, als (wanneer dat speelt, waarschijnlijk begin 2024) via PLOOI en <https://www.open-overheid.nl/> (<https://www.open-overheid.nl/>), als via de *Woogle* van de Universiteit van Amsterdam, de beoogde zoekmachine voor alle Woo-dossiers.

In [ ]: